Can SGD Handle Heavy-Tailed Noise?

Ilyas Fatkhullin ILYAS.FATKHULLIN@AI.ETHZ.CH
ETH AI Center, ETH Zurich

Florian Hübler FLORIAN.HUEBLER@INF.ETHZ.CH
ETH Zurich

Guanghui Lan GEORGE.LAN@ISYE.GATECH.EDU
Georgia Institute of Technology

Abstract

Stochastic Gradient Descent (SGD) is a cornerstone of large-scale optimization, yet its theoretical behavior under heavy-tailed noise—common in modern machine learning and reinforcement learning—remains poorly understood. In this work, we rigorously investigate whether vanilla SGD, devoid of any adaptive modifications, can provably succeed under such adverse stochastic conditions. Assuming only that stochastic gradients have bounded p-th moments for some $p \in (1, 2]$, we establish sharp convergence guarantees for (projected) SGD across convex, strongly convex, and non-convex problem classes. In particular, we show that SGD achieves minimax optimal sample complexity under minimal assumptions in the convex and strongly convex regimes: $\mathcal{O}(\varepsilon^{-\frac{p}{p-1}})$ and $\mathcal{O}(\varepsilon^{-\frac{p}{2(p-1)}})$, respectively. For non-convex objectives under Hölder smoothness, we prove convergence to a stationary point with rate $\mathcal{O}(\varepsilon^{-\frac{2p}{p-1}})$, and complement this with a matching lower bound specific to SGD.

1. Introduction

Consider a stochastic optimization problem

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E}\left[f(x, \xi)\right], \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is closed and convex, and ξ is a random variable distributed according to some unknown distribution $\mathcal{D}$. Access to F is only available through unbiased stochastic gradients $\nabla f(x, \xi)$, which may exhibit heavy-tailed behavior. Recent empirical studies have highlighted the prevalence of heavy-tailed phenomena in modern machine learning datasets and environments, especially in deep learning [1, 4, 43] and reinforcement learning [18]. To capture heavy-tailedness formally, we adopt the following moment assumptions.

Assumption 1 *Let $F(\cdot)$ be differentiable on $\mathcal{X}$.We have access to stochastic gradients with $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla F(x)$ and there exists $p \in (1, 2]$ such that the p-th moment is bounded, i.e.,*

$$p\text{-}\mathbf{BM} \qquad \mathbb{E}\left[\|\nabla f(x, \xi)\|^p\right] \leq G^p \qquad \textit{for all } x \in \mathcal{X}.$$

This assumption is standard in recent theoretical works on heavy-tailed optimization. When $p < 2$, gradient estimates can exhibit unbounded variance, precluding the use of conventional analysis techniques. Such heavy-tailed behavior of data is often used to explain the empirical success of adaptive algorithms over vanilla SGD. The examples of such adaptive schemes include methods based on gradient clipping [36, 42, 50], Normalized-SGD [22, 25, 33], their combinations [7, 8], and even more general non-linear schemes [2, 26, 38]. We provide a more detailed comparison of our results to related work in Appendix C. In this

	Convex	**Strongly Convex**	**Non-convex**
Convergence Criterion	$\mathbb{E}[F(\widetilde{x}_T) - F^*] \leq \varepsilon$	$\mathbb{E}[(F(\bar{x}_T) - F^*)^{p/2}] \leq \varepsilon^{p/2}$ or $\mathbb{E}[\|x_T - x^*\|^p] \leq \varepsilon^{p/2}$	$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] \leq \varepsilon^2$
Complexity in $\Theta(\cdot)$	$\left(\frac{G D_{\mathcal{X}}}{\varepsilon}\right)^{\frac{p}{p-1}}$ Thm. 1, [34]	$\left(\frac{G^2}{\mu\,\varepsilon}\right)^{\frac{p}{2(p-1)}}$ Thm. 2, [50]	$\Delta_1\left(\frac{L_p^{1/p} G}{\varepsilon^2}\right)^{\frac{p}{p-1}}$ Thms. 4, 9

Table 1: Summary of sample complexity bounds of SGD for solving (1) under (p-BM) Assumption 1. The "Converg. Criterion" row specifies the convergence criterion/measure for each setting with $\widetilde{x}_T$, $\bar{x}_T$, x_T denoting *average*, *random* and *last* iterate convergence respectively. The symbol $\Theta(\cdot)$ means that our rates are unimprovable for SGD under our assumptions. The constants $D_{\mathcal{X}}$, μ, Δ_1 and L_p denote the diameter of $\mathcal{X}$, strong convexity modulus, initial function value gap and the Hölder smoothness constant respectively; refer to corresponding section for formal definitions.

work, we revisit the analysis of vanilla Stochastic Gradient Descent with (optional) projection under heavy-tailed noise:

SGD: step-size sequence, $\{\eta_t\}_{t\geq 1}$, $\quad x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t \nabla f(x_t, \xi_t)),$

where $\Pi_{\mathcal{X}}(\cdot)$ is the Euclidean projection onto $\mathcal{X}$.[1] While this is one of the simplest and most popular stochastic optimization algorithms, its convergence in the heavy-tailed settings remains elusive. While many adaptive methods mentioned above can achieve optimal convergence, the majority of these works hardly question if SGD may have similar properties as these more sophisticated adaptive algorithms. The main argument discussed in the literature concerning the failure of SGD in such settings is centered around the following simple example. Consider the 1-dimensional quadratic function $F(x) = \frac{1}{2}x^2$ on $\mathcal{X} = \mathbb{R}$ and let $\xi \sim \mathcal{D}$ be any zero-mean noise with infinite variance and bounded p-th moment. In that case we have after one step of SGD starting from $x_1 = 0$

$$\mathbb{E}\left[F(x_2)\right] = \tfrac{1}{2}\mathbb{E}\left[\|\nabla F(x_2)\|^2\right] = \tfrac{1}{2}\mathbb{E}\left[\eta_1^2\,\|\xi_1\|^2\right] = +\infty. \tag{2}$$

The last equality holds due to infinite variance whenever the step-size η_1 is non-adaptive, i.e., predefined/deterministic. Therefore, vanilla SGD does not converge in the usual sense neither in expectation of the function value nor gradient norm squared. One might make an erroneous conclusion out of this example that SGD is useless under this heavy-tailed model of noise if $p < 2$. However, this example and convergence measure is fairly artificial and in this work we aim to investigate under what conditions and in what sense SGD may still converge under infinite variance.

Contributions We provide a comprehensive study of SGD under the (p-BM) assumption. Our analysis spans the convex, strongly convex, and non-convex settings and the main results are summarized in Table 1. The key contributions are as follows:

Convex. We establish that a weighted average function value of SGD converges for a wide range of step-size sequences. Moreover, if the diameter of the set $\mathcal{X}$ is bounded, even the simple average iterate of

[1] The projection is optional for most results but it is useful to discuss the implications when $\mathcal{X}$ is bounded.

SGD converges in expectation. In the latter case, when the step-size sequence η_t is tuned properly, SGD achieves the optimal sample complexity $\mathcal{O}(\varepsilon^{-\frac{p}{p-1}})$ to find $\widetilde{x}_T$ with $\mathbb{E}[F(\widetilde{x}_T) - F^*] \le \varepsilon$.[2]

Strongly convex. In this case we show that SGD with step-size $\eta_t = 2/\mu\, t, t \ge 1$ converges in function value, $\mathbb{E}[(F(\bar{x}_T) - F^*)^{p/2}] \le \varepsilon^{p/2}$ for a point $\bar{x}_T$ sampled uniformly from the iterates $\{x_t\}_{t \le T}$, and in terms of the distance to the optimum, $\mathbb{E}[\|x_T - x^*\|^p] \le \varepsilon^{p/2}$, with optimal sample complexity $\mathcal{O}(\varepsilon^{-\frac{p}{2(p-1)}})$. This implies that in strongly convex case, SGD only requires knowledge of strong convexity modulus μ and is the first optimal algorithm that does not require knowledge of the tail index p.

Non-convex. If function $F(\cdot)$ is Hölder smooth of order $\nu = p$, we show that SGD converges to a stationary point in expectation. In particular, in the unconstrained case this implies that we can find a point $\bar{x}_T$ with $\mathbb{E}\|\nabla F(\bar{x}_T)\|^2 \le \varepsilon^2$ after $\mathcal{O}\left(L_p^{\frac{1}{p-1}} \varepsilon^{-\frac{2p}{p-1}}\right)$ iterations of SGD, where L_p is the Hölder constant.

Overall, our results unify and extend the theoretical understanding of SGD, demonstrating that optimal convergence rates in heavy-tailed regimes can be achieved without any adaptive schemes across a wide range of problem classes. The interested reader can find additional complementary results, including (i) lower-bounds in-probability and in-expectation that imply tightness of our results, (ii) upper-bounds under the bounded *central* p-th moment assumption (i.e., $\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right] \le \sigma^p$) and (iii) empirical verifications in Appendix E.

2. Convex Setting

Our main running assumption in this section is standard.

Assumption 2 *The objective function $F(\cdot)$ is convex on $\mathcal{X} \subseteq \mathbb{R}^d$ and there exists an optimizer $x^* \in \arg\min_{y \in \mathcal{X}} F(y) \subseteq \mathcal{X}$.*

We are now ready to provide our main convergence guarantee for this setting.

Theorem 1 *Let Assumptions 1 and 2 hold ((p-BM) and convexity) with $p \in (1,2]$, and SGD is run with non-negative step-sizes $\{\eta_t\}_{t \ge 1}$. For any $t \ge 1$, define the weights $w_t := \eta_t \|x_t - x^*\|^{p-2}$. Then for any $T \ge 1$*

$$\frac{\sum_{t=1}^T \mathbb{E}[w_t(F(x_t) - F(x^*))]}{\sum_{t=1}^T \mathbb{E}[w_t]} \le \frac{\|x_1 - x^*\|^2 + 4G^2\left(\sum_{t=1}^T \eta_t^p\right)^{2/p}}{\sum_{t=1}^T \eta_t}.$$

If, additionally, the set $\mathcal{X}$ is bounded with diameter $D_{\mathcal{X}}$, then for any $T \ge 1$

$$\mathbb{E}\left[F(\widetilde{x}_T) - F(x^*)\right] \le \frac{D_{\mathcal{X}}^{2-p}\left(\|x_1 - x^*\|^p + 4G^p \sum_{t=1}^T \eta_t^p\right)}{\sum_{t=1}^T \eta_t}, \quad \widetilde{x}_T := \frac{\sum_{t=1}^T \eta_t\, x_t}{\sum_{t=1}^T \eta_t}.$$

To prove Theorem 1, we consider the Lyapunov (potential) function $\mathbb{E}\|x_t - x^*\|^p$, generalizing the classical analysis with $\mathbb{E}\|x_t - x^*\|^2$. We choose this new potential since the standard one with the power 2 does not have to be bounded even after making the first step of SGD, while the p-th power is bounded and is hence a suitable choice. The key technical challenge of the analysis is to build up an appropriate recursion for the p-th power of the norm. This is not straightforward since the direct argument of unrolling the square of the Euclidean norm by using inner product does not apply anymore. In order to overcome this challenge, our main observation is that the Euclidean norm raised to the power p, $p > 1$ (i.e., $\|x\|^p$) is $(p-1)$-Hölder smooth [41, Theorem 6.3]. The complete proof can be found in Appendix B.1.

[2] Throughout the work, we use the standard $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot)$ complexity notations [23], $\widetilde{\mathcal{O}}(\cdot)$ additionally hides poly-logarithmic factors. In some cases, we slightly abuse this notation to highlight the dependence of the complexities on certain variables of our focus, e.g., ε, p, which will hopefully be clear from the context.

Discussion. Notice that convergence is shown for an average suboptimality gap weighted with the dynamic random weights $w_t = \eta_t \|x_t - x^*\|^{p-2}$. Unfortunately, due to the correlation between the suboptimality and the weights, it is challenging to characterize the convergence for a specific point x_t in the sequence $\{x_t\}_{t\geq 1}$ even if we knew the weights $\{w_t\}_{t\geq 1}$. Our lower-bound in Appendix E.1.1 additionally clarifies that, for any predefined (non-stochastic) weighted output, SGD cannot converge in expected sub-optimality in the general setting when the set $\mathcal{X}$ is allowed to be unbounded. However, when the diameter of $\mathcal{X}$ is bounded we can establish convergence for implementable quantities: the running average iterate $\widetilde{x}_T = \sum_{t=1}^{T} \eta_t x_t / \sum_{t=1}^{T} \eta_t$ or the simple average $\widetilde{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$.

Now we discuss several choices for the step-size sequence. First, if we know all problem parameters, we can set the step-sizes diminishing as $\eta_t = \frac{\|x_1 - x^*\|}{G\, t^{1/p}}$ for all $t \geq 1$, or constantly $\eta_t \equiv \frac{\|x_1 - x^*\|}{G\, T^{1/p}}$ to obtain $\mathbb{E}\left[F(\widetilde{x}_T) - F(x^*)\right] = \widetilde{\mathcal{O}}\left(\frac{G D_{\mathcal{X}}^{2-p} \|x_1 - x^*\|^{p-1}}{T^{\frac{p-1}{p}}}\right)$ for all $T \geq 1$. This implies the $\mathcal{O}(\varepsilon^{-\frac{p}{p-1}})$ sample complexity, which is known to be optimal under Assumptions 1 and 2 for all $p \in (1,2]$, see e.g., [34, Chapter V, Section 3.1], [40] or Theorem 8. Second, we explore a universal step-size strategy which does not require knowledge of any parameters, $\eta_t = 1/\sqrt{t}$ for all $t \geq 1$. In this case we have for all $T \geq 1$ the bound $\mathbb{E}\left[F(\widetilde{x}_T) - F(x^*)\right] \leq \widetilde{\mathcal{O}}\left(\frac{D_{\mathcal{X}}^{2-p}\|x_1 - x^*\|^2}{\sqrt{T}} + \frac{D_{\mathcal{X}}^{2-p} G^p}{T^{\frac{p-1}{2}}}\right)$. This result shows that even untuned projected SGD converges under heavy-tailed noise.

3. Strongly Convex Setting

In this section, we revisit the convergence of (projected) SGD for strongly convex functions. We first recall the definition.

Assumption 3 *The objective function $F(\cdot)$ is μ-strongly convex on $\mathcal{X} \subseteq \mathbb{R}^d$, i.e., $F(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex on $\mathcal{X}$ with $\mu > 0$.*

Now we are ready to state the main convergence result of strong convexity SGD.

Theorem 2 *Let Assumptions 1 and 3 ((p-BM) and μ-strong convexity) hold with $p \in (1,2)$. Then the iterates generated by SGD with step-size $\eta_t = \frac{2}{\mu\, t}$ satisfy for any $T \geq 1$*

$$\mathbb{E}\left[\left(F(\bar{x}_T) - F(x^*)\right)^{p/2} + \left(\tfrac{\mu}{2}\right)^{p/2}\|x_{T+1} - x^*\|^p\right] \leq \frac{8G^p}{(2-p)\,\mu^{p/2}\,T^{p-1}}.$$

where $\bar{x}_T$ is sampled uniformly from the iterates $\{x_1, \ldots, x_T\}$.

The proof uses similar techniques as the convex case, however strong convexity allows us to relate w_t to our convergence measure and hence the removal of w_t from the guarantee. The complete proof can be found in Appendix B.2.

Discussion. The above theorem shows that instead of the classical convergence in expectation of the function sub-optimality and the distance to the optimum squared, these quantities should be raised to the power $p/2$ for any $p \in (1,2)$. This modification of convergence measure is meaningful in heavy-tailed setting and helps to circumvent the non-convergence example shown in Section 1. Indeed, for that quadratic example we have $\mathbb{E}\left[(F(x_2) - F^*)^{p/2}\right] = 2^{p/2}\eta_1^p \mathbb{E}\left[\xi_1^p\right] < \infty$ due to Assumption 1. We remark here that while Theorem 2 requires the *non-central* (p-BM) assumption, it can be extended to the central p-BCM version when smoothness is available allowing for unbounded domains and the variance reduction effect; we omit this extension for brevity due to the page limit. The sample complexity $\mathcal{O}\left(\varepsilon^{-\frac{p}{2(p-1)}}\right)$ — after accounting for the above mentioned nuances related to the convergence measure — implied by Theorem 1 matches the lower-bound for first-order methods in this setting [50]. We remark that the above theorem prescribes to use the classical step-size $\eta_t = 2/\mu t$, which is independent of tail index p. This implies that knowledge of p is not required for achieving the optimal sample complexity.

4. Non-Convex Setting

In this section, we will establish convergence rates for non-convex SGD under (p-BM). Before we proceed with our main result, we need to introduce some notions and additional assumptions for the non-convex case. First, we introduce the weakly Hölder convex and Hölder smooth functions [13, 14, 48].

Assumption 4 (Hölder smoothness) *Let $\nu \in [1,2]$ and $\ell_\nu, L_\nu > 0$. The objective function $F : \mathcal{X} \to \mathbb{R}$ is (ℓ_ν, L_ν, ν)-Hölder smooth with curvature exponent ν on $\mathcal{X}$, i.e.,*

$$-\tfrac{\ell_\nu}{\nu} \|x-y\|^\nu \le F(x) - F(y) - \langle \nabla F(y), x-y \rangle \le \tfrac{L_\nu}{\nu} \|x-y\|^\nu \qquad \text{for all } x, y \in \mathcal{X}.$$

In the case $\nu = 2$, we say $F(\cdot)$ is (ℓ, L)-smooth or simply smooth.

Note that the Hölder smoothness with $\nu < 2$ is weaker than standard smoothness (corresponding to $\nu = 2$) whenever the set $\mathcal{X}$ is bounded. Specifically, if smoothness holds with $L = L_2$, then Hölder smoothness holds with $L_\nu = \nu L_2 D_{\mathcal{X}}^{2-\nu}/2$. Thus, using Assumption 4 with $\nu = p \in (1,2]$ is not limiting in the constrained case if our main focus is the dependence on ε and p, but it may not guarantee the tightest possible bound in smoothness constant $L = L_2$ if the function is smooth.

Convergence criterion. We measure progress in the non-convex setting using a generalized stationarity measure based on *Forward-Backward Envelope (FBE)* [15], which is tailored to possibly constrained and non-smooth settings.

Definition 3 (FBE and Generalized Stationarity Measure) *Let $F : \mathcal{X} \to \mathbb{R}$ be differentiable and satisfy Assumption 4 with a curvature exponent $\nu > 1$. Let $\rho > 0$ be a regularization parameter. The Forward-Backward Envelope (FBE) of order ν at a point $x \in \mathcal{X}$ is defined as*

$$\mathcal{D}_\rho^\nu(x) := -\tfrac{\nu \rho^{\frac{1}{\nu-1}}}{\nu-1} \min_{y \in \mathcal{X}} Q_\rho^\nu(x,y), \qquad Q_\rho^\nu(x,y) := \langle \nabla F(x), y-x \rangle + \tfrac{\rho}{\nu} \|y-x\|^\nu . \tag{3}$$

We define the associated stationarity measure $S_\rho^\nu(x) := \left(\mathcal{D}_\rho^\nu(x)\right)^{\frac{2(\nu-1)}{\nu}}$.

In the unconstrained Euclidean case (i.e., $\mathcal{X} = \mathbb{R}^d$ or when projection is inactive), the FBE reduces to the gradient norm to the appropriate power: $\mathcal{D}_\rho^\nu(x) = \|\nabla F(x)\|^{\frac{\nu}{\nu-1}}$. Thus our stationarity measure $S_\rho^\nu(x)$ reduces to $S_\rho^\nu(x) = \|\nabla F(x)\|^2$ for any $\nu > 1$ when $\mathcal{X} = \mathbb{R}^d$.

Theorem 4 *Let $F(\cdot)$ be lower bounded by F^* and Assumptions 1, 4 ((p-BM), Hölder smooth with $\nu = p$) hold, denote $\Delta_1 := F(x_1) - F^*$. Then for any $T \ge 1$ SGD with arbitrary step-size sequence $\{\eta_t\}_{t \ge 1}$ satisfies*

$$\tfrac{1}{\sum_{t=1}^T \eta_t} \sum_{t=1}^T \eta_t \mathbb{E}\left[S_{\rho+L_p}^p(x_t)\right] \le 16 \left(\tfrac{p}{p-1}\right)^\gamma \tfrac{\Delta_1 + 2\rho G^p \sum_{t=1}^T \eta_t^p}{\sum_{t=1}^T \eta_t}, \quad \text{where} \quad \gamma := \tfrac{2(p-1)}{p}, \rho := \tfrac{2(L_p + 2\ell_p)}{p-1}.$$

Discussion If $\mathcal{X} = \mathbb{R}^d$, we recall that $S_\rho^p(x) = (\mathcal{D}_\rho^p(x))^{\frac{2(p-1)}{p}} = \|\nabla F(x)\|^2$ for any $\rho > 0$ and any $p \in (1, 2]$. Therefore, the above theorem implies in the unconstrained case

$$\mathbb{E}\,\|\nabla F(\bar{x}_T)\|^2 \le 16\left(\frac{p}{p-1}\right)^{\gamma} \frac{\Delta_1+2\,\rho\, G^p \sum_{t=1}^T \eta_t^p}{\sum_{t=1}^T \eta_t} = \mathcal{O}\left(\frac{(\ell_p+L_p)^{1/p}\Delta_1^{\frac{p-1}{p}} G}{T^{\frac{p-1}{p}}}\right),$$

where the last equality holds setting $\eta_t = \sqrt[p]{\Delta_1}/G\sqrt[p]{\rho T}$ and $\bar{x}_T$ is sampled from the iterates $\{x_t\}_{t\le T}$ either uniformly or with probabilities proportional to η_t. In any of these cases, perhaps surprisingly, SGD converges in the standard measure for unconstrained non-convex optimization: expectation of the gradient squared. We remark that while our Hölder smoothness Assumption 4 is weaker that standard smoothness (with $\nu = 2$) on any compact set $\mathcal{X}$, there is still no contradiction with the folklore quadratic example discussed in Section 1. The crux is that to include the quadratic example $F(x) = \frac{1}{2}x^2$ in our function class satisfying $(0, L_p, p)$-Hölder smoothness, we need to assume a bounded domain, which gives $L_p \le pD_{\mathcal{X}}^{2-p}/2$. This means the derivation in (2) is not valid anymore due to the projection on the bounded set $\mathcal{X}$, and the projected SGD actually converges in terms of expectation of gradient norm squared. In Appendix E.1.2 we prove that the above mentioned convergence rate is not improvable for SGD with arbitrary polynomial step-sizes under our exact assumptions up to a numerical constant.

Acknowledgments

This work is supported by Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343, and the Office of Naval Research grant N00014-24-1-2654. **IF** is supported by the ETH AI Center Doctoral Fellowship.

References

[1] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.

[2] Aleksandar Armacki, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability convergence bounds for nonlinear stochastic gradient descent under heavy-tailed noise. *arXiv preprint arXiv:2310.18784*, 2023.

[3] Site Bai and Brian Bullins. Tight lower bounds under asymmetric high-order hölder smoothness and uniform convexity. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fMTPkDEhLQ.

[4] Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788, 2024.

[5] Witold M. Bednorz, Rafał Martynek, and Rafal M. Lochowski. On tails of symmetric and totally asymmetric alpha-stable distributions. *Probability and Mathematical Statistics*, 41(2), 2020.

[6] Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. In *Conference on Learning Theory*, 2024.

[7] Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

[8] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

[9] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(1):2237–2274, 2021.

[10] Nikita Doikov. Lower complexity bounds for minimizing regularized functions. *arXiv preprint arXiv:2202.04545*, 2022.

[11] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667, 2020.

[12] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

[13] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

[14] Pavel Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization. *arXiv preprint arXiv:1703.09180*, 2017.

[15] Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general bregman divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3501, 2024.

[16] Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, volume 99, 2019.

[17] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

[18] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, 2021.

[19] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[20] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

[21] Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *J. Complex.*, 31(1):1–14, 2015.

[22] Chuan He, Zhaosong Lu, Defeng Sun, and Zhanwang Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

[23] Rodney Howell. On Asymptotic Notation with Multiple Variables. *Tech. Rep.*, 2008.

[24] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869, 2024.

[25] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[26] Dusan Jakovetic, Dragana Bajovic, Anit Kumar Sahu, Soummya Kar, Nemanja Milosevic, and Dusan Stamenkovic. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023.

[27] Sajad Khodadadian and Martin Zubeldia. A general-purpose theorem for high-probability bounds of stochastic approximation with polyak averaging. *arXiv preprint arXiv:2505.21796*, 2025.

[28] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[29] Guanghui Lan, Tianjiao Li, and Yangyang Xu. Projected gradient methods for nonconvex and stochastic optimization: New complexities and auto-conditioned stepsizes. *arXiv preprint arXiv:2412.14291*, 2024.

[30] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: From theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

[31] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[32] Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *International Conference on Learning Representations*, 2024.

[33] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

[34] Arkadij Nemirovskij and David Yudin. Efficient methods of solving convex programming problems of high dimensionality. *Ekonomika i matem. methody (in Russian)*, XV(1), 1979.

[35] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2̂). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[36] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-[]Paper-[]Conference.pdf.

[37] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[38] Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Adaptive estimation algorithms: Convergence, optimality, stability. *Avtomatika i telemekhanika*, pages 71–84, 1979.

[39] Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864, 2024.

[40] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 803–510, 2009.

[41] Anton Rodomanov and Yurii Nesterov. Smoothness parameter of power of euclidean norm. *Journal of Optimization Theory and Applications*, 185(2):303–326, 2020. ISSN 1573-2878. doi: 10.1007/s10957-[]020-[]01653-[]6.

[42] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *International conference on machine learning*, 2023.

[43] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837, 2019.

[44] Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

[45] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102, 2022.

[46] Yijun Wan, Melih Barsbey, Abdellatif Zaidi, and Umut Simsekli. Implicit compressibility of overparametrized neural networks trained with heavy-tailed sgd. *arXiv preprint arXiv:2306.08125*, 2023.

[47] Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *Advances in Neural Information Processing Systems*, 34:18866–18877, 2021.

[48] Maryam Yashtini. On the global convergence rate of the gradient descent method for functions with hölder continuous gradients. *Optimization letters*, 10:1361–1370, 2016.

[49] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

[50] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[51] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012, 2022.

Appendix A. Useful Lemmata

First, we recall the definition of FBE of order ν that was given in (3):

$$\mathcal{D}^{\nu}_{\rho}(x) := -\tfrac{\nu\,\rho^{\frac{1}{\nu-1}}}{\nu-1}\min_{y\in\mathcal{X}} Q^{\nu}_{\rho}(x,y), \qquad Q^{\nu}_{\rho}(x,y) := \langle \nabla F(x), y-x\rangle + \tfrac{\rho}{\nu}\left\|y-x\right\|^{\nu}.$$

The following lemma is a modified version of Step I in the proof of the main theorem in [15]. This descent type lemma is useful in the proof of Theorem 4.

Lemma 5 *For any $x \in \mathcal{X}$ and any $\rho_1 \geq \rho + L$ we have $F^{\nu}_{1/\rho}(x) \leq F(x) - \frac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}}\mathcal{D}^{\nu}_{\rho_1}(x)$.*

Proof Notice that for any $x \in \mathcal{X}$, we have for any $x^+ \in \mathcal{X}$

$$F^{\nu}_{1/\rho}(x) = F(\hat{x}) + \tfrac{\rho}{\nu}\left\|\hat{x}-x\right\|^{\nu} \leq F(x^+) + \tfrac{\rho}{\nu}\left\|x^+ - x\right\|^{\nu},$$

where $\hat{x} \in \arg\min_{y\in\mathcal{X}} F(y) + \frac{\rho}{\nu}\left\|y-x\right\|^{\nu}$ as before. We use a mirror descent step starting from a point x as $x^+ := \arg\min_{y\in\mathcal{X}} \langle \nabla F(x), y\rangle + \frac{\rho_1}{\nu}\left\|y-x\right\|^{\nu}$ with $\rho_1 \geq \rho + L$. Then by Hölder smoothness of $F(\cdot)$ (upper bound in Assumption 4)

$$\begin{aligned} F^{\nu}_{1/\rho}(x) &\leq F\left(x^+\right) + \tfrac{\rho}{\nu}\left\|x^+ - x\right\|^{\nu} \\ &\leq F\left(x\right) + \left\langle \nabla F\left(x\right), x^+ - x\right\rangle + \tfrac{L}{\nu}\left\|x^+ - x\right\|^{\nu} + \tfrac{\rho}{\nu}\left\|x^+ - x\right\|^{\nu} \\ &= F\left(x\right) - \tfrac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}}\mathcal{D}^{\nu}_{\rho_1}(x) + \tfrac{\rho+L-\rho_1}{\nu}\left\|x^+ - x\right\|^{\nu} \leq F\left(x\right) - \tfrac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}}\mathcal{D}^{\nu}_{\rho_1}(x). \end{aligned} \tag{4}$$

where the last equality holds by definitions of x^+, $\mathcal{D}^{\nu}_{\rho}(x)$ and the last step is due to condition $\rho_1 \geq \rho + L$. ■

Recall the definition of the proximal point of order $\nu \in [1,2]$

$$\hat{x}^{\nu} := \arg\min_{y\in\mathcal{X}}\left[F(y) + \tfrac{\rho}{\nu}\left\|y-x\right\|^{\nu}\right] \quad \text{for any } x \in \mathcal{X}.$$

Lemma 6 *Let $\nu \in (1,2]$. For any $x \in \mathcal{X}$ and any $\rho > \frac{\rho_1 + 2\,\ell}{\nu}$ we have*

$$\left\|\hat{x}^{\nu} - x\right\|^{\nu} \leq \tfrac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}} \cdot \tfrac{\mathcal{D}^{\nu}_{\rho_1}(x)}{\rho - \frac{\rho_1+2\,\ell}{\nu}}.$$

Proof By the definition of $\hat{x}^{\nu}$ and the optimality condition, we have for any $u \in \mathcal{X}$

$$\left\langle \nabla F(\hat{x}^{\nu}) + \rho(\hat{x}^{\nu} - x)\left\|\hat{x}^{\nu} - x\right\|^{\nu-2}, u - \hat{x}^{\nu}\right\rangle \geq 0.$$

Setting $u = x$, we obtain

$$\rho\left\|\hat{x}^{\nu} - x\right\|^{\nu} \leq \langle \nabla F(\hat{x}^{\nu}), x - \hat{x}^{\nu}\rangle = \underbrace{\langle \nabla F(x), x - \hat{x}^{\nu}\rangle}_{\text{(I)}} + \underbrace{\langle \nabla F(\hat{x}^{\nu}) - \nabla F(x), x - \hat{x}^{\nu}\rangle}_{\text{(II)}}. \tag{5}$$

We will bound terms (I) and (II) separately. First, by the definition of FBE:

$$\begin{aligned} \text{(I)} &= \langle \nabla F(x), x - \hat{x}^{\nu}\rangle - \tfrac{\rho_1}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu} + \tfrac{\rho_1}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu} \\ &\leq \max_{y\in\mathcal{X}}\left\{\langle \nabla F(x), x - y\rangle - \tfrac{\rho_1}{p}\left\|y - x\right\|^{\nu}\right\} + \tfrac{\rho_1}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu} \\ &= -\min_{y\in\mathcal{X}} Q^{\nu}_{\rho_1}(x,y) + \tfrac{\rho_1}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu} = \tfrac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}}\mathcal{D}^{\nu}_{\rho_1}(x) + \tfrac{\rho_1}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu}. \end{aligned}$$

Second, by Hölder smoothness Assumption 4 we have for any $x, y \in \mathcal{X}$

$$-\tfrac{\ell}{\nu}\left\|x-y\right\|^{\nu} \le F(x) - F(y) - \langle \nabla F(y), x-y\rangle,$$

$$-\tfrac{\ell}{\nu}\left\|x-y\right\|^{\nu} \le F(y) - F(x) + \langle \nabla F(x), x-y\rangle.$$

Summing up the above two inequalities for $x = x$ and $y = \hat{x}^{\nu}$, we can bound the term:

$$\text{(II)} = -\langle \nabla F(\hat{x}^{\nu}) - \nabla F(x), \hat{x}^{\nu} - x\rangle \le \tfrac{2\ell}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu}.$$

Using the two upper bounds in (5), we derive

$$\rho\left\|\hat{x}^{\nu} - x\right\|^{\nu} \le \tfrac{\nu-1}{\nu\rho_1^{\frac{1}{\nu-1}}}\mathcal{D}^{\nu}_{\rho_1}(x) + \tfrac{\rho_1+2\,\ell}{\nu}\left\|\hat{x}^{\nu} - x\right\|^{\nu}.$$

It remains to rearrange and use the restriction for ρ. ■

Lemma 7 *[Lemma 10 in [25]] Let $p \in [1,2]$, and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}\left[[]\, X_{j-1}, \ldots, X_1\right]X_j = 0$ a.s. for all $j = 1, \ldots, n$ satisfying: $\mathbb{E}\left[\|X_j\|^p\right] < \infty \quad$ for all $j = 1, \ldots, n$. Define $S_n := \sum_{j=1}^{n} X_j$, then*

$$\mathbb{E}\left[\|S_n\|^p\right] \le 2\sum_{j=1}^{n}\mathbb{E}\left[\|X_j\|^p\right].$$

Appendix B. Missing Proofs

In this section we present the missing proofs of the main paper.

B.1. Convex Setting

Proof of Theorem 1. We start with the Hölder smoothness of $\|\cdot\|^p$ [41, Theorem 6.3], i.e.,

$$\|v+w\|^p \le \|v\|^p + p\tfrac{\langle v,w\rangle}{\|v\|^{2-p}} + 2^{2-p}\left\|w\right\|^p \qquad \text{for any } v, w \in \mathcal{X}, v \ne 0. \tag{6}$$

By non-expansiveness of the projection, using the update rule of SGD and the above inequality with $v := x_t - x^*$, $w := \eta_t\, \nabla f(x_t, \xi_t)$, and assuming $x_t \ne x^*$,[3] we have

$$\begin{aligned}\|x_{t+1} - x^*\|^p &= \|\Pi_{\mathcal{X}}(x_t - \eta_t\nabla f(x_t,\xi_t)) - \Pi_{\mathcal{X}}(x^*)\|^p \le \|x_t - x^* - \eta_t\,\nabla f(x_t,\xi_t)\|^p \\ &\overset{(6)}{\le} \|x_t - x^*\|^p - \eta_t p\tfrac{\langle \nabla f(x_t,\xi_t), x_t - x^*\rangle}{\|x_t - x^*\|^{2-p}} + 2^{2-p}\eta_t^p\left\|\nabla f(x_t,\xi_t)\right\|^p.\end{aligned}$$

Next we take conditional expectation and use convexity Assumption 2 along with (p-BM) to derive

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^p \,|\, x_t\right] \le \|x_t - x^*\|^p - \eta_t p\tfrac{F(x_t) - F(x^*)}{\|x_t - x^*\|^{2-p}} + 2^{2-p}\eta_t^p G^p. \tag{7}$$

Define $r_t := \|x_t - x^*\|$, $\Delta_t := F(x_t) - F(x^*)$, $w_t := \eta_t\, r_t^{p-2}$. Taking the total expectation of (7), we obtain:

$$\mathbb{E}[r_{t+1}^p] \le \mathbb{E}[r_t^p] - p\,\mathbb{E}\left[w_t\Delta_t\right] + 2^{2-p}\eta_t^p G^p. \tag{8}$$

[3] We implicitly assume throughout that $x_t \ne x^*$ for any $t \le T$, otherwise the problem is solved.

Ignoring the negative term on the RHS and unrolling, we establish for any $t \geq 1$

$$\mathbb{E}\left[r_t^p\right] \leq r_1^p + 2^{2-p} G^p \sum_{\tau=1}^{t-1} \eta_\tau^p \leq r_1^p + 2^{2-p} G^p \sum_{\tau=1}^{t} \eta_\tau^p =: C_t. \tag{9}$$

Coming back to (8), we derive the bound

$$\frac{\sum_{t=1}^T \mathbb{E}[w_t \Delta_t]}{\sum_{t=1}^T \mathbb{E}[w_t]} \quad \leq \quad \frac{C_T}{\sum_{t=1}^T \mathbb{E}[w_t]} = \frac{C_T}{\sum_{t=1}^T \eta_t \mathbb{E}[(r_t^p)^{\frac{p-2}{p}}]} \leq \frac{C_T}{\sum_{t=1}^T \eta_t \mathbb{E}[r_t^p]^{\frac{p-2}{p}}} \leq \frac{C_T}{\sum_{t=1}^T \eta_t C_T^{\frac{p-2}{p}}} = \frac{C_T^{2/p}}{\sum_{t=1}^T \eta_t},$$

where the second inequality follows by convexity of $x \mapsto x^{\frac{p-2}{p}}$ along with Jensen's inequality, and the last inequality follows from (9) and the fact that $\frac{p-2}{p} \leq 0$. The final bound follows by recalling the definition of C_T given in (9) and simplifying the rate using the fact that $p \in (1, 2]$.

If, additionally, the diameter is bounded, then we can upper bound the denominator in (7) and repeat similar steps to derive the second inequality in the theorem statement. ■

B.2. Strongly Convex Setting

Proof of Theorem 2. We start similarly to our analysis in convex case using Hölder smoothness of $\|x\|^p$ to get

$$\|x_{t+1} - x^*\|^p \leq \|x_t - x^*\|^p - \eta_t p \tfrac{\langle \nabla f(x_t, \xi_t), x_t - x^* \rangle}{\|x_t - x^*\|^{2-p}} + 2^{2-p} \eta_t^p \left\|\nabla f(x_t, \xi_t)\right\|^p.$$

Next we take the conditional expectation and use strong convexity to derive

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|^p \mid x_t\right] &\leq \|x_t - x^*\|^p - \eta_t p \tfrac{\frac{\mu}{2}\|x_t - x^*\|^2 + F(x_t) - F(x^*)}{\|x_t - x^*\|^{2-p}} \\
&\quad + 2^{2-p} \eta_t^p \mathbb{E}\left[\|\nabla f(x_t, \xi_t)\|^p \mid x_t\right] \\
&= \left(1 - \tfrac{p\,\eta_t \mu}{2}\right) \|x_t - x^*\|^p - \eta_t p \tfrac{F(x_t) - F(x^*)}{\|x_t - x^*\|^{2-p}} + 2^{2-p} \eta_t^p \mathbb{E}\left[\|\nabla f(x_t, \xi_t)\|^p \mid x_t\right] \\
&\leq \left(1 - \tfrac{\eta_t \mu}{2}\right) \|x_t - x^*\|^p - \eta_t \left(\tfrac{\mu}{2}\right)^{\frac{2-p}{2}} \left(F(x_t) - F(x^*)\right)^{p/2} \\
&\quad + 2^{2-p} \eta_t^p \mathbb{E}\left[\|\nabla f(x_t, \xi_t)\|^p \mid x_t\right],
\end{aligned} \tag{10}$$

where in the last step we used the quadratic growth condition $F(x) - F(x^*) \geq \frac{\mu}{2} \|x - x^*\|^2$ to bound the term in the denominator and the fact that $p \in (1, 2] > 1$. Define the distance to the optimum by $r_t := \|x_t - x^*\|$ and the suboptimality $\Delta_t := F(x_t) - F(x^*)$. Then taking the total expectation of (10), using (p-BM), and setting the step-sizes as $\eta_t = {}^{2}/_{\mu\, t}$, we have

$$\mathbb{E}[r_{t+1}^p] \leq \tfrac{t-1}{t} \mathbb{E}[r_t^p] - \eta_t \left(\tfrac{\mu}{2}\right)^{\frac{2-p}{2}} \mathbb{E}\left[\Delta_t^{p/2}\right] + \tfrac{2^2 G^p}{\mu^p\, t^p}.$$

Multiplying both sides by $t \geq 1$, summing up over $t = 1, \ldots, T$, and rearranging,

$$\left(\tfrac{\mu}{2}\right)^{\frac{2-p}{2}} \sum_{t=1}^T \mathbb{E}[t\, \eta_t \cdot \Delta_t^{p/2}] + T\, \mathbb{E}\left[r_{T+1}^p\right] \leq \sum_{t=1}^T \tfrac{4 G^p}{\mu^p\, t^{p-1}} \leq \tfrac{8 G^p}{(2-p)\, \mu^p\, T^{p-2}}.$$

Pluggining in $\eta_t = {}^{2}/_{\mu\, t}$ and dividing both sides by $({}^{\mu}/_{2})^{p/2}\, T$ we obtain

$$\tfrac{1}{T} \sum_{t=1}^T \mathbb{E}[\Delta_t^{p/2}] + \left(\tfrac{\mu}{2}\right)^{p/2} \mathbb{E}\left[r_{T+1}^p\right] \leq \tfrac{8 G^p \left(\frac{\mu}{2}\right)^{p/2}}{(2-p)\, \mu^p\, T^{p-1}}.$$

■

B.3. Non Convex Setting

First, we assume the function is lower bounded, which is a standard assumption in this setting.

Assumption 5 *The function $F : \mathcal{X} \to \mathbb{R}$ is lower bounded by $F^* > -\infty$ on $\mathcal{X}$.*

Next, we introduce some useful definitions needed for our analysis. For any $F : \mathcal{X} \to \mathbb{R}$ and a real $\rho > 0$, the Moreau envelope and its associated proximal operator (at a point $x \in \mathcal{X}$) are defined respectively by

$$F^{\nu}_{1/\rho}(x) := \min_{y \in \mathcal{X}} \left[F(y) + \tfrac{\rho}{\nu} \|y - x\|^{\nu}\right], \qquad \hat{x}^{\nu} := \arg\min_{y \in \mathcal{X}} \left[F(y) + \tfrac{\rho}{\nu} \|y - x\|^{\nu}\right].$$

In our analysis of SGD we will use Hölder smoothness Assumption 4 and Moreau envelope with $\nu = p$. Later, in the analysis of Mini-batch SGD we will use the above concepts with $\nu = 2$. In case $\nu = 2$, we will omit the superscript and denote $F_{1/\rho}(\cdot) := F^2_{1/\rho}(\cdot)$, $\hat{x} := \hat{x}^2$.

Proof of Theorem 4. We start with definition of Moreau envelope and the optimality of $\hat{x}^p_{t+1}$:

$$\begin{aligned} F^p_{1/\rho}(x_{t+1}) &= F\left(\hat{x}^p_{t+1}\right) + \tfrac{\rho}{2} \left\|\hat{x}^p_{t+1} - x_{t+1}\right\|^p \le F\left(\hat{x}^p_t\right) + \tfrac{\rho}{2} \left\|\hat{x}^p_t - x_{t+1}\right\|^p \\ &= F\left(\hat{x}^p_t\right) + \tfrac{\rho}{2} \left\|\Pi_{\mathcal{X}}(\hat{x}^p_t) - \Pi_{\mathcal{X}}(x_t - \eta_t \nabla f(x_t, \xi_t))\right\|^p \\ &\le F\left(\hat{x}^p_t\right) + \tfrac{\rho}{2} \left\|(\hat{x}^p_t - x_t) + \eta_t \nabla f(x_t, \xi_t)\right\|^p, \end{aligned}$$

where in the last inequality above we used non-expansiveness of projection. Next, similarly to convex case we use the Hölder smoothness of $\|\cdot\|^p$:

$$\begin{aligned} F^p_{1/\rho}(x_{t+1}) &\le F\left(\hat{x}^p_t\right) + \tfrac{\rho}{2} \left\|x_t - \hat{x}^p_t\right\|^p + \tfrac{\rho\, p\, \eta_t}{2} \tfrac{\left\langle \hat{x}^p_t - x_t, \nabla f(x_t, \xi_t)\right\rangle}{\left\|\hat{x}^p_t - x_t\right\|^{2-p}} + \tfrac{2^{2-p}\, \rho\, \eta^p_t}{2} \left\|\nabla f(x_t, \xi_t)\right\|^p \\ &= F^p_{1/\rho}(x_t) + \tfrac{\rho\, p\, \eta_t}{2} \tfrac{\left\langle \hat{x}^p_t - x_t, \nabla f(x_t, \xi_t)\right\rangle}{\left\|\hat{x}^p_t - x_t\right\|^{2-p}} + \tfrac{2^{2-p}\, \rho\, \eta^p_t}{2} \left\|\nabla f(x_t, \xi_t)\right\|^p. \end{aligned}$$

On the other hand, using Assumption 4 and Lemma 5, we have for $\rho > \ell_p$

$$\begin{aligned} \left\langle \nabla F(x_t), \hat{x}^p_t - x_t \right\rangle &\le F(\hat{x}^p_t) - F(x_t) + \tfrac{\ell_p}{p} \left\|\hat{x}^p_t - x_t\right\|^p \\ &= F^p_{1/\rho}(x_t) - F(x_t) + \tfrac{\ell_p - \rho}{p} \left\|\hat{x}^p_t - x_t\right\|^p \le -C\, \mathcal{D}^p_{\rho + L_p}(x_t), \end{aligned}$$

where we denote $C := \frac{p-1}{p\,(\rho + L_p)^{\frac{1}{p-1}}}$. Combining the above two inequalities, and defining $\psi_t := \nabla f(x_t, \xi_t) - \nabla F(x_t)$, $\Delta_{t,p} := F^p_{1/\rho}(x_t) - F^*$, we attain

$$\Delta_{t+1,\,p} \le \Delta_{t,\,p} - \tfrac{\rho\, p\, \eta_t\, C}{2} \tfrac{\mathcal{D}^p_{\rho + L_p}(x_t)}{\left\|\hat{x}^p_t - x_t\right\|^{2-p}} + \tfrac{2^{2-p}\, \rho\, \eta^p_t}{2} \left\|\nabla f(x_t, \xi_t)\right\|^p + \tfrac{\rho\, p\, \eta_t}{2} \tfrac{\left\langle \hat{x}^p_t - x_t, \psi_t \right\rangle}{\left\|\hat{x}^p_t - x_t\right\|^{2-p}}. \tag{11}$$

It follows from Lemma 6 that for any $\rho > \frac{L_p + 2\,\ell_p}{p-1} > \ell_p$

$$\left\|\hat{x}^p_t - x_t\right\|^p \le \frac{C}{\rho - \frac{L_p + 2\,\ell_p}{p-1}} \mathcal{D}^p_{\rho + L_p}(x_t).$$

Thus we can upper bound the denominator of the negative term in (11).

$$\begin{aligned} \Delta_{t+1,\,p} \quad \le \quad & \Delta_{t,\,p} - \tfrac{\rho\, \eta_t}{2} \left(\rho - \tfrac{L + 2\,\ell}{p-1}\right)^{\frac{2-p}{p}} \left(C\, \mathcal{D}^p_{\rho + L_p}(x_t)\right)^{\frac{2(p-1)}{p}} + 2\,\rho\, \eta^p_t \left\|\nabla f(x_t, \xi_t)\right\|^p \\ & + \tfrac{\rho\, p\, \eta_t}{2} \tfrac{\left\langle \hat{x}^p_t - x_t, \psi_t \right\rangle}{\left\|\hat{x}^p_t - x_t\right\|^{2-p}}. \end{aligned}$$

Telescoping, taking the total expectation, using Assumption 1 and rearrange the bound with $\rho = \frac{2(L_p+2\,\ell_p)}{p-1}$ yields

$$\left(\tfrac{\rho}{2}\right)^{\frac{2}{p}} C^{\frac{2(p-1)}{p}} \sum_{t=1}^{T} \eta_t S^p_{\rho+L_p}(x_t) \le \Delta_{1,\,p} + 2\rho G^p \sum_{t=1}^{T} \eta_t^p,$$

where we used $S^p_{\rho+L}(x_t) = (\mathcal{D}^p_\rho(x))^{\frac{2(p-1)}{p}}$. Noticing that $\Delta_{1,\,p} \le \Delta_1 = F(x_t) - F^*$, and $\left(\tfrac{\rho}{2}\right)^{\frac{2}{p}} C^{\frac{2(p-1)}{p}} \ge \left(\frac{p-1}{p}\right)^{\frac{2(p-1)}{p}} \frac{1}{16}$ yields the claim. ■

Appendix C. Comparison of Results to Prior Work

Appendix D. Related Work

A growing body of work develops adaptive algorithms that achieve provable convergence under heavy-tailed noise. These include mirror descent variants [34, 45], Clip-SGD [31, 36, 42, 50], Normalized SGD [25], Clip-AdaGrad [7], Clip-SGD with normalization and momentum [8] and other non-linear schemes [2, 26, 38]. Some of these works go beyond in-expectation analysis and develop high-probability guarantees with polylogarithmic dependence on the inverse failure probability, $1/\delta$, thanks to adaptive step-sizes or other complex techniques such as robust distance estimation [9] and robust gradient aggregation methods [39]. Although adaptive methods often offer strong convergence guarantees, vanilla SGD may retain distinct advantages. For instance, [46] show that SGD under heavy-tailed noise can induce implicit compressibility—a property potentially lost in adaptive schemes involving clipping or normalization. Nevertheless, the theoretical understanding of vanilla SGD in the heavy-tailed setting remains limited. The only existing convergence result, due to [47], relies on restrictive technical assumptions and fails to achieve optimal rates; see Appendix D.2 for further discussion.

D.1. Convex Setting

Several existing adaptive algorithms in the literature achieve similar convergence rates as Theorem 1. For example, a scheme based on the mirror descent framework was proposed in the seminal work [34, Chapter V, Section 2.1] and later revisited in [32, 45]. The update rule of this scheme, when simplified to the unconstrained Euclidean setup, can be written as

$$\text{p-SMD:} \qquad x_{t+1} = \frac{x_t\|x_t\|^{\frac{p}{p-1}} - \eta_t \nabla f(x_t,\xi_t)}{\left\| x_t\|x_t\|^{\frac{p}{p-1}} - \eta_t \nabla f(x_t,\xi_t) \right\|^{2-p}}, \qquad \eta_t = \frac{\eta_1}{t^{1/p}}.$$

As we can see even after some conceptual simplification, this method is more complicated than vanilla SGD. First, it can be classified as an adaptive/normalized method due to the normalization of the updated point. Second, it requires computing the norm of x_t and another auxiliary vector. Third, it not only requires tuning the step-size η_t, but also requires setting the correct power for computing $\|x_t\|^{\frac{p}{p-1}}$, which is critical for convergence. This leads to unnecessary complications for implementations and computational instabilities. In fact, the implementation of the projected variant of this method requires an additional non-Euclidean projection, which can add extra computational burden.

Another modification of SGD that has become very popular in stochastic optimization and machine learning literature recently is called SGD with gradient clipping or Clip-SGD [31, 42, 50, 51]. For a predefined sequence of clipping thresholds $\{\lambda_t\}_{t\ge 1}$, the update rule of this algorithm is

$$\text{Clip-SGD:} \qquad x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \eta_t g_t\right), \quad g_t = \operatorname{clip}\left(\nabla f(x_t,\xi_t), \lambda_t\right)$$

with $\mathrm{clip}(v, \lambda) := v \cdot \min\{1, \lambda/\|v\|_2\}$. The typical choice of the sequence $\{\lambda_t\}_{t\geq 1}$ in the convex setting is increasing and has the order $\lambda_t = \lambda_1 t^{1/p}$, see e.g., [31, 36, 42].[4] While this algorithm is simpler than the previous scheme, it still requires additional hyper-parameter tuning for sequence $\{\lambda_t\}_{t\geq 1}$.

To summarize, the main advantage of our analysis in Theorem 1 is achieving similar convergence rates to above mentioned methods with a simpler algorithm – vanilla SGD.

D.2. Strongly Convex Setting

We first compare our guarantee to the literature on Clip-SGD [31, 50] (method is described in the previous section). In strongly convex setting, their recommendation for step-size is $\eta_t = {}^4\!/\!_{\mu(t+1)}$ and for clipping threshold sequence is $\lambda_t = \max\left\{2\max_{x\in\mathcal{X}}\|\nabla F(x)\|, G\,t^{1/p}\right\}$, see e.g., Theorem 9 in [31]. They obtain the rate in expected function sub-optimality

$$\mathbb{E}\left[F(\widetilde{x}_T) - F^* + \mu\|x_T - x^*\|^2\right] = \mathcal{O}\left(\frac{G^2}{\mu(T+1)^{\frac{2(p-1)}{p}}}\right) \tag{12}$$

for some weighed average point $\widetilde{x}_T$. There are two main differences compared to our result. First, our convergence rate is established for vanilla SGD without clipping. Clip-SGD has two parameter sequences which are important to tune, and the clipping threshold depends on the tail index p to achieve the optimal rate. In comparison, our SGD only has the standard step-size $\eta_t = {}^2\!/\!_{\mu t}$, which is easy to implement when the strong convexity modulus is known. The second difference is in the convergence criterion used. Our convergence criterion, $\mathbb{E}\left[(F(\bar{x}_T) - F(x^*))^{p/2} + \left(\frac{\mu}{2}\right)^{p/2}\|x_{T+1} - x^*\|^p\right]$, can be weaker than the one for Clip-SGD in (12). This means that our new analysis allows to get rid of additional clipping threshold hyper-parameter at the price of a slightly weaker convergence measure.

Now we compare our result to [47], which studies convergence of vanilla SGD under a similar infinite variance assumption. Their step-size sequence choice is arbitrarily close to the harmonic decay, i.e., $\eta_t = \eta_1/t^a$ for any $a \in (0, 1)$, and the convergence rate is presented as $\mathbb{E}\left[\|x_t - x^*\|^p\right] \leq C(d, L, \mu, \sigma)/t^{a(p-1)}$, where the constant $C(d, L, \mu, \sigma)$ hides the dependence on dimension, variance parameter, smoothness and strong convexity parameters. While this result looks similar to our Theorem 2, unfortunately, it has several important limitations. First, their analysis assumes that $F(\cdot)$ is twice differentiable function with a uniformly bounded spectral norm of the Hessian matrix $\nabla^2 F(x)$ (L-smoothness). Second, they make an additional non-standard assumption about the uniform p-positive definiteness of the Hessian matrix. To explain this concept, we let $p \geq 1$ and $\mathbf{Q}$ be a symmetric matrix. Define the signed power of a vector $\mathbf{v} \in \mathbb{R}^d$ as: $\mathbf{v}^{\langle q\rangle} := (\mathrm{sgn}(v^1)|v^1|^q, \ldots, \mathrm{sgn}(v^d)|v^d|^q)^\top$. Let $S_p = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_p = 1\}$ be the unit sphere in $\|\cdot\|_p$ norm. We say that $\mathbf{Q}$ is *p-positive definite* if for all $\mathbf{v} \in S_p$, $\mathbf{v}^\top\mathbf{Q}\mathbf{v}^{\langle p-1\rangle} > 0$. In particular, in the limit case case $p \to 1$ the assumption reduces to the diagonal dominance of the Hessian. Even simple functions such as $F(x) = x^\top \begin{pmatrix} 0.02 & -1 \\ -1 & 50.02 \end{pmatrix} x$ violated this assumption for all $p \leq 1.8$. Another limitation of this work is that the convergence rates suffers from a polynomial dimension dependence, making it unscalable with d even when the main problem constants (such as smoothness constant, strong convexity parameter and the variance) are dimension independent.

D.3. Non-Convex Setting

To the best of our knowledge, there are no algorithms in the literature working precisely under the same assumptions as our Theorem 4. However, we can compare our result to other algorithms, which use the stan-

[4] In fact the choice of exact sequence is more complicated and depends on other problem parameters, but we report the asymptotic behavior for intuition.

dard smoothness and the bounded central moment assumptions. We first compare to Clip-SGD [42, 50]. Under smoothness and p-BCM, the typical recommended order of the step-size and clipping thresholds Clip-SGD are $\eta_t = \eta_1 / t^{\frac{p}{3p-2}}$, $\lambda_t = \lambda_1 \, t^{\frac{1}{3p-2}}$. The above mentioned works derive the sample complexity of order $\mathcal{O}\left(\varepsilon^{-\frac{3p-2}{p-1}}\right)$, which is smaller that our best possible result $\mathcal{O}\left(\varepsilon^{-\frac{2p}{p-1}}\right)$ for SGD with $\eta_t = \eta_1 / \sqrt[p]{T}$ step-size when $p < 2$, and matches when $p = 2$. Several other adaptive methods achieve sample complexities similar to $\mathcal{O}\left(\varepsilon^{-\frac{3p-2}{p-1}}\right)$. For example, Normalized SGD with mini-batch or momentum [25, 33], and Normalized SGD with gradient clipping and momentum [8]. However, all these algorithms are more complex than vanilla SGD, rely on smoothness assumptions, are limited to unconstrained setting, and, similar to Clip-SGD, require at least two hyper-parameters to achieve reduced sample complexity For example, Normalized SGD with momentum [25] requires step size $\eta_t = \eta_1 \sqrt{\alpha_t / t}$ and momentum $\alpha_t = \alpha_1 / t^{\frac{p}{3p-2}}$.

In summary, we prove for the first time that non-convex SGD converges under heavy-tailed noise. Our complexity bound is worse in dependence on ε than those discussed above in the standard smooth setting for $p \in (1, 2)$, but the assumptions and the algorithms are different. This discrepancy can be caused by three potential possibilities:

i. Our analysis may be not tight.

ii. Complexity of first-order methods under Assumption 4 with $\nu = p$ is strictly worse than for smooth (i.e., $\nu = 2$).

iii. SGD is inherently slower than other adaptive methods such as Clip-SGD even if $\nu = 2$.

In Appendix E.1.2 we essentially exclude the first possibility i., by constructing an algorithm-specific lower bound for SGD with arbitrary polynomial step-sizes. This will show that our upper bound in Theorem 4 are in fact tight under our assumption. We leave the exploration of scenarios ii. and iii. for future work.

Appendix E. Complementary Results

This section contains additional results that complement our main upper-bounds presented in the main part. In Appendix E.1.1 we provide an in-probability lower-bound for SGD in the convex setting, before showing tightness of Theorem 4 in Appendix E.1.2. Appendix E.2 provides a convergence result of Minibatch-SGD under the bounded p-th *central* moment assumption, and experimental validations can be found in Appendix E.3.

E.1. Lower-Bounds

This section contains lower-bounds which complement our upper-bounds presented in the main part. Therefore let us first summarize the related work.

Several works provide sample complexity lower bounds for first-order algorithms in different heavy-tailed regimes. For Lipschitz, convex functions on a bounded domain, the seminal works [34, 40] provide a tight $\Omega\big(\varepsilon^{-p/(p-1)}\big)$ lower-bound. For L-smooth strongly-convex and non-convex functions, the $\Omega(\varepsilon^{-\frac{p}{2(p-1)}})$ and $\Omega(\varepsilon^{-\frac{3p-2}{p-1}})$ lower-bounds for the class of first-order methods are established in [50]. The works [6, 42] show in probability lower bounds, but their construction uses bounded noise and is limited to the bounded variance case. In comparison, our in probability construction is nearly tight for any $p \in (1, 2]$ and works for a large class of algorithms. Under light-tailed noise, algorithm-specific lower-bounds are established for SGD in [11] and for stochastic approximation in [27]. The lower bounds for Hölder-smooth functions are available in [3, 10, 21]. We complement these works by providing an algorithm specific lower-bounds for SGD for the heavy-tailed *and* Hölder-smooth setting.

E.1.1. IN PROBABILITY LOWER BOUND

The results in Section 2 imply, in particular, that for any $p \in (1, 2]$, SGD with constant step-size $\eta_t = \frac{\|x_1 - x^*\|}{G T^{1/p}}$ converges at the rate $\mathbb{E}\left[F\left(\widetilde{x}_T\right) - F(x^*)\right] \leq \frac{5\, G D_{\mathcal{X}}^{2-p} \|x_1 - x^*\|^{p-1}}{T^{\frac{p-1}{p}}}$. While this rate is known to be optimal [34, 40] in-expectation, it does not give us any insights about the behavior of an individual run of the algorithm. To measure the concentration of SGD around this expected convergence rate, we can use the Markov's inequality, which implies with probability at least $1 - \delta$:

$$F\left(\widetilde{x}_T\right) - F(x^*) \leq \frac{5\, G D_{\mathcal{X}}^{2-p} \|x_1 - x^*\|^{p-1}}{T^{\frac{p-1}{p}}} \frac{1}{\delta}, \qquad \text{where} \quad \widetilde{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t. \tag{13}$$

As we can see, this result has a poor dependence on $1/\delta$. In this section, we consider the large class of first-order methods with satisfying the cone condition, i.e., for any $T \geq 1$

$$x_T^{out} \in x_1 - \text{Cone}\left\{\nabla f(x_1, \xi_1), \ldots, \nabla f(x_{T-1}, \xi_{T-1})\right\}. \tag{14}$$

It is important to note that we focus on the case when the cone coefficients determining the specific algorithm are predefined/deterministic. We will show that for such algorithms, a similar polynomial dependence as in (13) is inevitable.[5] This will imply, in particular, that SGD for arbitrary step-sizes and any reasonable output strategy will suffer from such polynomial dependence on $1/\delta$.

Theorem 8 *Let $p \in (1, 2]$, $T \geq 2$, $\delta \in (0, 1/8]$, and the stochastic gradient oracle satisfies Assumption 1. Then for any algorithm satisfying the cone condition (14), there exists a convex problem (1) such that for any $\alpha > p$, with probability at least δ*

$$F(x_T^{out}) - F(x^*) \geq \frac{\|x_1 - x^*\|}{2(T-1)^{\frac{\alpha-1}{\alpha}}} \left(\frac{1}{4\,\alpha\,\delta}\right)^{1/\alpha} = \Omega\left(\frac{G\|x_1 - x^*\|}{T^{\frac{\alpha-1}{\alpha}}} \left(\frac{1}{\delta}\right)^{1/\alpha}\right),$$

where the last equality holds for $T \geq 1 + 2^{\frac{\alpha}{\alpha-1}} \left(\frac{1}{2\,\alpha\,\delta}\right)^{\frac{1}{\alpha-1}} \left(\frac{\Gamma(1-p)}{\Gamma(1-p/\alpha)}\right)^{\frac{\alpha}{\alpha-1}}$, and $\Gamma(x)$ is the Gamma function.

Proof We define a one-dimensional problem on $\mathcal{X} = \mathbb{R}$ and the stochastic gradient oracle:

$$F(x) \;=\begin{cases} -a\,x & \text{if } x \leq 0 \\ \frac{L}{2}x^2 - a\,x & \text{if } 0 \leq x \leq \frac{a}{L} \\ -\frac{a^2}{2\,L} & \text{if } x \geq \frac{a}{L}, \end{cases} \qquad \nabla f(x, \xi) \;=\begin{cases} -a + \xi & \text{if } x \leq 0 \\ L\,x - a + \xi & \text{if } 0 \leq x \leq \frac{a}{L} \\ \xi & \text{if } x \geq \frac{a}{L}, \end{cases}$$

where $a, L > 0$ will be specified later. We use a random variable $\xi \in \mathbb{R}$ such that for $\alpha > 1$, it has a characteristic function

$$\mathbb{E}\left[\exp\left(i\,s\,\xi\right)\right] = \exp\left(-|s|^{\alpha}\right).$$

This distribution is zero-mean and has bounded p-th moment for any $p < \alpha$. Namely, p-BCM holds with $\sigma := \frac{\Gamma(1-p/\alpha)}{\Gamma(1-p)}$, where $\Gamma(x)$ is a Gamma function, and (p-BM) holds with $G := 2^{p-1}(a^p + \sigma^p)^{1/p}$. By the cone assumption, for any $T \geq 1$ there exists a non-negative sequence $\{\gamma_t\}_{t \geq 1}$ such that

$$x_{T+1}^{out} = x_1 - \sum_{t=1}^{T} \gamma_t \nabla f(x_t, \xi_t) \leq x_1 + a \sum_{t=1}^{T} \gamma_t - \sum_{t=1}^{T} \gamma_t\, \xi_t, \tag{15}$$

[5] Here we say “similar” because formally our upper bound is established under bounded diameter assumption, while the lower bound construction has an unbounded domain.

where in the last inequality we used the fact that $\nabla f(x, \xi) \geq -a + \xi$ for any $x, \xi \in \mathbb{R}$. To establish an in probability lower bound, we set

$$a := \frac{\left(\sum_{t=1}^T \gamma_t^\alpha\right)^{1/\alpha}}{2\sum_{t=1}^T \gamma_t} \left(\frac{1}{4\,\alpha\,\delta}\right)^{1/\alpha}, \qquad L := \frac{1}{\sum_{t=1}^T \gamma_t}.$$

By independence of $\{\xi_t\}_{t\geq 1}$ and Theorem 7 in [5], we have

$$\Pr\left(\sum_{t=1}^T \gamma_t\,\xi_t \geq z\right) \geq \frac{1}{2}\frac{1}{2+\frac{\alpha z^\alpha}{\sum_{t=1}^T \gamma_t^\alpha}} =: \delta.$$

Using $\delta \leq 1/8$, we can bound $z^\alpha \geq \frac{1}{4\alpha\,\delta}\sum_{t=1}^T \gamma_t^\alpha$. Therefore, we have with probability at least δ

$$x_{T+1}^{out} \leq x_1 + a\sum_{t=1}^T \gamma_t - \left(\frac{1}{4\,\alpha\,\delta}\right)^{1/\alpha}\left(\sum_{t=1}^T \gamma_t^\alpha\right)^{1/\alpha} = x_1 - a\sum_{t=1}^T \gamma_t,$$

where in the last step we used the definition of a. Now multiplying both sides with $-a < 0$, selecting an arbitrary negative starting point $x_1 < 0$ and the optimum closest to the starting point $x^* = a/L$, we have

$$F(x_{T+1}^{out}) - F(x^*) \quad = \quad -ax_{T+1}^{out} + \frac{a^2}{2L} \geq a\left\|x_1 - x^*\right\| + a^2\sum_{t=1}^T \gamma_t - \frac{a^2}{2L} \geq a\left\|x_1 - x^*\right\|,$$

where the last step uses the definition of L. It remains to recall the definition of G, and notice that for any $\alpha \geq 1$, $a \geq \frac{1}{2T^{\frac{\alpha-1}{\alpha}}}\left(\frac{1}{4\,\alpha\,\delta}\right)^{1/\alpha}$, and the suboptimality $F(x_{T+1}^{out}) - F(x^*)$ is lower bounded with

$$\frac{a\,G\|x_1 - x^*\|}{2^{p-1}(a^p+\sigma^p)^{1/p}} \geq \frac{G\|x_1-x^*\|}{2}\min\left\{1; \frac{a}{\sigma}\right\} = \Omega\left(\frac{G\|x_1-x^*\|}{T^{\frac{\alpha-1}{\alpha}}}\left(\frac{1}{\delta}\right)^{1/\alpha}\right).$$

■

First, we observe that the cone assumption includes a number of first-order algorithms including SGD with last iterate output, for $\gamma_t = \eta_t$; SGD with simple average output (13), for $\gamma_t = \sum_{k=1}^{t-1}\eta_k/T$. Similarly, the majority of momentum and accelerated schemes can be expressed in this form, including heavy-ball momentum [17, 37], Nesterov's acceleration [28, 35], many regularized schemes [30]. In particular, this lower bound shows that the convergence rate of SGD will necessarily be multiplied by a polynomial in the inverse failure probability $1/\delta$, unlike recent results for high probability convergence of Clip-SGD [31, 42, 51] and Normalized-SGD [25]. Moreover, the lower bound holds for any $\alpha > p$, which means that it is nearly tight when compared to the best upper bound achieved by Theorem 1 in terms of dependence on T and p. This lower bound is also remarkably tight in δ dependence and only leaves a small gap compared to our upper bound (13) of order $\delta^{1-\frac{1}{\alpha}}$ in failure probability, which disappears in extremely heavy-tailed regime as $\delta^{1-\frac{1}{\alpha}} \to 1$ when $\alpha \to 1$. It is worth to mention that previously Sadiev et al. [42] established a high probability lower bound for SGD in strongly convex setting using bounded noise, $\Omega\left(1/\sqrt{\varepsilon\delta}\right)$, which is not tight for their setting. While their dependence on $1/\delta$ is also polynomial, our construction is different and extends to any algorithm satisfying the cone assumption.

It is worth noting that (15) in the proof of above theorem implies that if the set $\mathcal{X}$ is unbounded, the expectation $\mathbb{E}\left[F(x_T^{\text{out}})\right] = +\infty$ for any reasonable output strategy of SGD. This means that the bounded diameter assumption in Theorem 1 is necessary for any reasonable output strategy to convegence in expectation. While our lower bound uses an unbounded set $\mathcal{X} = \mathbb{R}$, we believe it is possible to modify our construction allowing a bounded set, e.g., $\mathcal{X} = [-D_{\mathcal{X}}, 0]$ and replacing the initial distance to the optimum, $\|x_1 - x^*\|$, with a sufficiently large diameter, $D_{\mathcal{X}}$. This would have allowed us to formally match the upper bound of SGD in (13). However, the two main obstacles to extend our proof to the bounded diameter case are (i) generalizing the cone assumption to such constrained setting and (ii) carefully selecting the diameter $D_{\mathcal{X}}$ to avoid hitting the projection on the left. We believe the second obstacle is manageable when considering the last iterate of projected SGD instead of a general class of algorithms and selecting sufficiently large $D_{\mathcal{X}}$. However, the obstacle (i) seems more challenging.

E.1.2. NON-CONVEX LOWER-BOUND

When choosing polynomially decaying stepsizes $\eta_t = \eta\, t^{-r}$, $\eta > 0, r \in [0, 1)$, Theorem 4 implies a sample complexity of

$$T = \tilde{\mathcal{O}}\Big(\Big(\tfrac{\Delta_1}{\eta\varepsilon^2}\Big)^{\frac{1}{1-r}} + \eta^{\frac{1}{r}}\big(\tfrac{L_\nu G^p}{\varepsilon^2}\big)^{\frac{1}{r(p-1)}}\Big) \tag{16}$$

to reach an ε-stationary point, i.e., $\mathbb{E}\,[S_\rho^p] \leq \varepsilon$. As the previously established lower-bounds in the literature do not cover our set of assumptions, we derive the tightness of this result in all parameters in the following Theorem.

Theorem 9 *Let $p, \nu \in (1, 2], \Delta_1, L_\nu \geq 0, 0 \leq \sigma \leq G, \varepsilon > 0$ and $\eta > 0, r \in [0, 1)$. Assume $\varepsilon < G/2$ and $\varepsilon^{\frac{\nu}{\nu-1}} \leq \frac{\nu}{\nu-1}\frac{\Delta_1}{4}\big(\frac{L_\nu}{2}\big)^{\frac{1}{\nu-1}}$.[6] Then, for any dimension $d \in \mathbb{N}_{\geq 1}$ and $i \in \{1, 2\}$, there exist convex, $(0, L_\nu, \nu)$-smooth and G-Lipschitz functions $F_1, F_2 \colon \mathbb{R}^d \to \mathbb{R}$ with $F_i(x_1) - \inf_{x\in\mathbb{R}^d} F_i(x) \leq \Delta_1$, and gradient oracles $\nabla f_i(x, \xi)$, that satisfy (p-BM) and (p-BCM) such that SGD with stepsizes $\eta_t = \eta\, t^{-r}$ almost surely requires at least*

$$T \geq \max\Big\{\Big(\tfrac{(1-r)\Delta_1}{8\eta\varepsilon^2}\Big)^{\frac{1}{1-r}}, \eta^{\frac{1}{r}}\Big(\tfrac{\sigma^p}{2^p L_\nu}\Big)^{\frac{1}{r(p-1)}}\big(\tfrac{L_\nu}{2\varepsilon}\big)^{\frac{p-1+\nu-1}{r(p-1)(\nu-1)}}\Big\} \tag{17}$$

iterations to reach an ε-stationary point.

In other words, the above theorem implies that for all T that do not satisfy the above inequality (17), we have $\min_{t\in[T]} \|\nabla F_1(x_t)\| > \varepsilon$ or $\min_{t\in[T]} \|\nabla F_2(x_t)\| > \varepsilon$ almost surely. In particular, the ε-dependence is given by

$$T = \Omega\Big(\varepsilon^{-\frac{2}{1-r}} + \varepsilon^{-\frac{p-1+\nu-1}{r(p-1)(\nu-1)}}\Big).$$

The proof of Theorem 9 consists of two lower-bound constructions, providing the first and second lower-bound term respectively. The following proposition provides the first term in (17), using a deterministic function construction punishing small stepsizes.

Proposition 10 *Let $\nu \in (1, 2], \varepsilon > 0, \Delta_1, L_\nu \geq 0$ and assume the target accuracy ε is sufficiently small, $\varepsilon^{\frac{\nu}{\nu-1}} \leq \frac{\nu}{\nu-1}\frac{\Delta_1}{4}\big(\frac{L_\nu}{2}\big)^{\frac{1}{\nu-1}}$. Then there exists a convex, (2ε)-Lipschitz continuous, $(0, L_\nu, \nu)$-Hölder smooth function $F \colon \mathbb{R} \to \mathbb{R}$ with $F(x_1) - \inf_{x\in\mathbb{R}} F(x) \leq \Delta_1$ such that SGD with stepsizes $(\eta_t)_{t\in\mathbb{N}_{\geq 1}} \geq 0$ requires*

$$\sum_{t=1}^{T-1} \eta_t > \tfrac{\Delta_1}{4\varepsilon^2}$$

to hold to reach an ε-stationary point, i.e. $x_T \in \mathbb{R}$ with $\|\nabla F(x_T)\| \leq \varepsilon$.

[6] Note that both assumption are mild in the sense that, whenever one of them is violated, we have convergence at x_1 or within constantly many iterations.

Proof This construction follows the idea from [24], and adapts it to SGD and Hölder-Smoothness. Similar constructions can be found in [12, Section 3.2]. W.l.o.g. assume $x_1 = 0$ and define $y_t := \sum_{\kappa=1}^{t-1} 2\varepsilon\eta_t$. Let

$$T^* := \sup\left\{T \in \mathbb{N} \mid 2\varepsilon y_T \le \tfrac{\Delta_1}{2}\right\}$$

and define $\delta_\nu := \left(\frac{2\varepsilon}{L}\right)^{\frac{1}{\nu-1}}$,

$$F(x) := \begin{cases} \Delta_1 - 2\varepsilon x, & x \le y_{T^*} \\ \Delta_1 - 2\varepsilon x + \frac{L}{\nu}(x - y_{T^*})^\nu, & y_{T^*} < x \le y_{T^*} + \delta_\nu \\ \Delta_1 - 2\varepsilon y_{T^*} - 2\varepsilon\delta_\nu + \frac{L}{\nu}\delta_\nu^\nu & y_{T^*} + \delta_\nu < x. \end{cases}$$

Note that, by the definition of T^* and our assumption $\varepsilon^{\frac{\nu}{\nu-1}} \le \frac{\nu}{\nu-1}\frac{\Delta_1}{4}\left(\frac{L}{2}\right)^{\frac{1}{\nu-1}}$, we have

$$F(x) \ge F(y_{T^*} + \delta_\nu) = \Delta_1 - 2\varepsilon y_{T^*} - 2\varepsilon\delta_\nu + \tfrac{L}{\nu}\delta_\nu^\nu \ge \tfrac{\Delta_1}{2} - 2\varepsilon\delta_\nu + \tfrac{L}{\nu}\delta_\nu^\nu \ge 0$$

and hence $F(x_1) - \inf_{x\in\mathbb{R}} F(x) \le \Delta_1$. Furthermore F is $(0, L, \nu)$-Hölder smooth and convex by construction. Next, note that

$$F'(x) := \begin{cases} -2\varepsilon, & x \le y_{T^*} \\ -2\varepsilon + L(x - y_{T^*})^{\nu-1}, & y_{T^*} < x \le y_{T^*} + \delta_\nu \\ 0 & y_{T^*} + \delta_\nu < x. \end{cases}$$

and hence we have $\|\nabla F(y_t)\| = 2\varepsilon > \varepsilon$ for all $t \le T^*$. Finally note that SGD, when started at $x_1 = 0$, observes $F'(x_t) = -2\varepsilon$ as long as $x_t \le y_{T^*}$. In particular, the iterates are given by $x_t = \sum_{\kappa=1}^{t-1} 2\varepsilon\eta_t = y_t$, and hence $|F'(x_t)| = 2\varepsilon > \varepsilon$ for all $t \le T^*$. ■

Next we focus on the second term of (17), by constructing a function and oracle that punish large stepsizes.

Proposition 11 *Let $\nu \in (1, 2], 0 \le \sigma \le G, \eta, L_\nu > 0$ and $x \in \mathbb{R}^d$. Furthermore let $d \in \mathbb{N}$ be arbitrary and*

$$F\colon \mathbb{R}^d \to \mathbb{R}, F(z) := \begin{cases} \frac{L_\nu}{2^{2-\nu}\nu}\|z\|^\nu, & \|z\| \le r_H \\ \frac{G}{2}\|z\| - C, & \|z\| > r_H, \end{cases}$$

where $r_H := \left(\frac{G}{2^{\nu-1}L_\nu}\right)^{\frac{1}{\nu-1}}$ and $C = \frac{\nu-1}{4\nu}\left(\frac{G^\nu}{L_\nu}\right)^{\frac{1}{\nu-1}}$. Then there exist a gradient oracle $\nabla f(z, \xi)$ that satisfies (p-BM) and (p-BCM) such that $y := x - \eta\nabla f(x, \xi)$ satisfies

$$\|y\| \ge \min\left\{\|x\|, \tfrac{1}{2}\left(\tfrac{\eta^{p-1}\sigma^p}{2^p L_\nu}\right)^{\frac{1}{p+\nu-2}}\right\}.$$

Proof We first note that F is differentiable, and its gradient satisfies

$$\|\nabla F(x)\| = \left\{\begin{array}{ll} \frac{L_\nu}{2^{2-\nu}}\|x\|^{\nu-1}, & \|x\| \le r_H \\ \frac{G}{2}, & \|x\| > r_H \end{array}\right\} \le \frac{G}{2}.$$

The proof hinges on constructing a gradient oracle that prevents y from entering the open ball $B_{\bar\tau}$ with radius $\bar\tau := \min\{\|x\|, \tau\}$, where $\tau := \frac{1}{2}\left(\frac{\eta^{p-1}\sigma^p}{2^p L_\nu}\right)^{\frac{1}{p+\nu-2}}$. We will differentiate two cases.

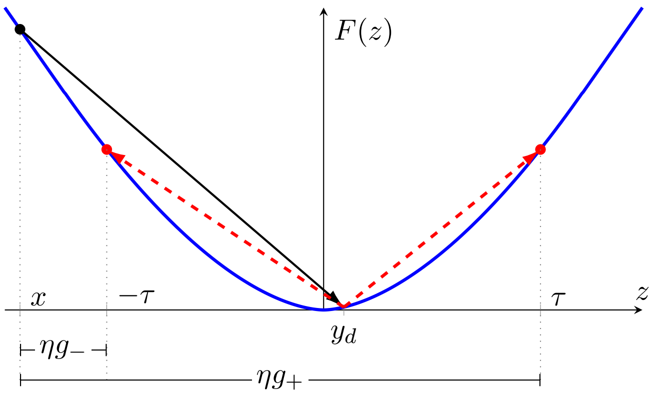


Figure 1: Visualisation of the lower bound construction in Proposition 11, for the case where the deterministic update $y_d = x - \eta\nabla F(x)$ lands within the critical radius τ. The dashed lines represent the two possible offsets introduced by the constructed gradient oracle, resulting in $y = x - \eta g_\pm u = x + \eta g_\pm = \pm\tau$.

Case I: $\|x - \eta\nabla F(x)\| \geq \bar{\tau}$**.** In this case we can use the deterministic gradient oracle $\nabla f(x,\xi) = \nabla F(x)$, as the deterministic update already lands outside $B_{\bar{\tau}}$. This oracle trivially satisfies p-BCM, and p-BM is satisfied due to $\|\nabla F(x)\| \leq \frac{G}{2}$.

Case II: $\|x - \eta\nabla F(x)\| < \bar{\tau}$**.** This case means that the deterministic update would land in $B_{\bar{\tau}}$, and we must hence construct an oracle that moves y outside it. Therefore define $r := \|x\|, u := \frac{x}{r}, \mu := \|\nabla F(x)\|, \mu' := \frac{L_\nu r^{\nu-1}}{2^{2-\nu}} \geq \mu$ and

$$g_\pm := \frac{r \pm \bar{\tau}}{\eta} \geq 0, \qquad \delta_\pm := |\mu - g_\pm|, \qquad \Delta := g_+ - g_- = \frac{2\bar{\tau}}{\eta}.$$

Note that by our case assumption, we have $g_- < \mu < g_+$. We next define the probability $\delta := \frac{\delta_-}{\Delta}$ and gradient oracle

$$\nabla f(x,\xi) := \begin{cases} g_- u, & \xi \geq \delta \\ g_+ u, & \xi \leq \delta, \end{cases}$$

where $\xi \sim \mathrm{Unif}([0,1])$. By construction we have that $\|x - \eta\nabla f(x,\xi)\| = |r - (r \pm \bar{\tau})| \, \|u\| = \bar{\tau}$ and hence it only remains to show that $\nabla f(x,\xi)$ satisfies the noise assumptions. Therefore first note that this oracle is unbiased by

$$\mathbb{E}\left[\nabla f(x,\xi)\right] = (1-\delta)g_- u + \delta g_+ u = \frac{(g_+-\mu)g_- + (\mu - g_-)g_+}{\Delta}u = \frac{\mu(g_+-g_-)}{\Delta}u = \nabla F(x),$$

where we used $\delta = \frac{\mu - g_-}{\Delta}, 1-\delta = \frac{g_+-\mu}{\Delta}$ and the definition of Δ. Next we check the bounded p-th central moment property. Therefore we first calculate

$$\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right] = (1-\delta)\delta_-^p + \delta\delta_+^p = \Delta^p((1-\delta)\delta^p + \delta(1-\delta)^p).$$

Now let $s := \min\{\delta, 1-\delta\}$ and note that $(1-\delta)\delta^p + \delta(1-\delta)^p \leq \delta^p + \delta \leq 2\delta$. Using a symmetric argument for $1-\delta$, we get that $(1-\delta)\delta^p + \delta(1-\delta)^p \leq 2s$. Next, by definition, we have $s \leq 1/2$ and

$$\delta = \frac{\mu - g_-}{\Delta} \leq \frac{\mu' - g_-}{\Delta} = \frac{\eta L_\nu r^{\nu-1}}{2^{2-\nu}2\bar{\tau}} - \frac{r-\bar{\tau}}{2\bar{\tau}} = At^{\nu-1} - \frac{t-1}{2},$$

where $A := \frac{\eta L_\nu}{2(2\bar{\tau})^{2-\nu}}$, and $t := \frac{r}{\bar{\tau}} \geq 1$. Noting that, whenever $A \leq 1/2$,

$$At^{\nu-1} - \tfrac{t-1}{2} = A + A\big(t^{\nu-1} - 1\big) - \tfrac{t-1}{2} \leq A + \tfrac{t^{\nu-1}-t}{2} \leq A$$

yields $s \leq A$ and hence

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right] \leq \Delta^p 2s \leq \left(\tfrac{2\bar{\tau}}{\eta}\right)^p \tfrac{\eta L_\nu}{(2\bar{\tau})^{2-\nu}} &= (2\bar{\tau})^{p+\nu-2}\tfrac{L_\nu}{\eta^{p-1}} \\ &\leq (2\tau)^{p+\nu-2}\tfrac{L_\nu}{\eta^{p-1}} = \tfrac{\sigma^p}{2^p}.\end{aligned}$$

By Jensen's inequality we have $\left\|\frac{v+w}{2}\right\|^p \leq \frac{\|v\|^p+\|w\|^p}{2}$ for all $v, w \in \mathbb{R}^d$, and hence $\|v+w\|^p \leq 2^{p-1}(\|v\|^p + \|w\|^p)$. Finally we use this fact to get

$$\mathbb{E}\left[\|\nabla f(x,\xi)\|^p\right] \leq 2^{p-1}(\|\nabla F(x)\|^p + \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right]) \leq \frac{G^p}{2} + \frac{\sigma^p}{2} \leq G^p.$$

Hence the gradient oracle satsifies the p-BM and p-BCM assumption, completing the proof. ■

By iteratively applying Proposition 11, we get the lower bound

$$\min_{t\in[T]} \|x_t\| \geq \min\left\{\|x_1\|, \min_{t\in[T]} \tau_t\right\}, \qquad \text{where} \qquad \tau_t := \tfrac{1}{2}\left(\tfrac{\eta_t^{p-1}\sigma^p}{2^p L_\nu}\right)^{\frac{1}{p+\nu-2}} \tag{18}$$

on the iterates. Translating it to a gradient norm lower-bound, and combining it with Proposition 10 yields Theorem 9. The formal argument can be found below.

Proof of Theorem 9 We first note that the functions in Proposition 10 and Proposition 11 are convex, $(0, L, \nu)$-smooth and G-Lipschitz by their definitions and our assumption $\varepsilon/2 < G$. Additionally note that we can lift the function F from Proposition 10 to $F_1 \colon \mathbb{R}^d \to \mathbb{R}$, by setting $F_1(x) = F(x_1)$. Hence we can use these constructions for $F_1, \nabla f_1$ and $F_2, \nabla f_2$ respectively.

Now let us first assume $T < \left(\frac{(1-r)\Delta_1}{8\eta\varepsilon^2}\right)^{\frac{1}{1-r}}$. Then we have

$$\sum_{t=1}^{T-1} \eta_t \leq \eta\bigg(1 + \int_1^{T-1} t^{-r}\bigg) = \eta\Big(1 + \tfrac{T^{1-r}-1}{1-r}\Big) \leq \tfrac{\eta T^{1-r}}{1-r} < \tfrac{\Delta_1}{8\varepsilon^2}$$

and hence $\min_{t\in[T]} \|\nabla F_1(x_t)\| > \varepsilon$ by Proposition 10.

Next we choose $F_2, \nabla f_2$ from Proposition 11 and $\|x_1\| = \left(\frac{2^{2-\nu}\nu\Delta_1}{L}\right)^{1/\nu}$, which guarantees $F_2(x_1) - \inf_x F_2(x) \leq \Delta_1$. By iteratively applying Proposition 11, we get

$$\min_{t\in[T]} \|x_t\| \geq \min\left\{\|x_1\|, \min_{t\in[T]} \tau_t\right\}, \qquad \text{where} \qquad \tau_t := \frac{1}{2}\left(\frac{\eta_t^{p-1}\sigma^p}{2^p L}\right)^{\frac{1}{p+\nu-2}}.$$

By our non-degeneration assumptions, i.e., $\varepsilon < \frac{G}{2}$ and $\varepsilon^{\frac{\nu}{\nu-1}} < \nu\Delta_1\big(\frac{L}{2^{2-\nu}}\big)^{\frac{1}{\nu-1}}$, we have $\|\nabla F(x_1)\| > \varepsilon$ and hence no ε-stationary point can be reached before the inequality $\tau_T = \min_{t\in[T]} \tau_t \leq \left(\frac{2^{2-\nu}\varepsilon}{L}\right)^{\frac{1}{\nu-1}}$ is satisfied. Rewriting this inequality yields

$$\begin{aligned}\tau_T \leq \left(\tfrac{2^{2-\nu}\varepsilon}{L}\right)^{\frac{1}{\nu-1}} &\Leftrightarrow \tfrac{\eta_T^{p-1}\sigma^p}{2^p L} \leq 2^{p+\nu-2}\left(\tfrac{2^{2-\nu}\varepsilon}{L}\right)^{\frac{p+\nu-2}{\nu-1}} \\ &\Leftrightarrow \eta T^{-r} \leq \left(\tfrac{2^p L}{\sigma^p}\right)^{\frac{1}{p-1}} 2^{\frac{p+\nu-2}{(p-1)(\nu-1)}} \left(\tfrac{\varepsilon}{L}\right)^{\frac{p+\nu-2}{(p-1)(\nu-1)}}.\end{aligned}$$

As this inequality is not satisfied whenever $T < \eta^{\frac{1}{r}}\left(\frac{\sigma^p}{2^p L}\right)^{\frac{1}{r(p-1)}}\left(\frac{L}{2\varepsilon}\right)^{\frac{p-1+\nu-1}{r(p-1)(\nu-1)}}$, we get the claim. ■

Init. Assumption	Convex	Lower-Bound	Upper-Bound
$F(x_1) - F^* \leq \Delta_1$	✗	$\Omega\big(\varepsilon^{-4}\big)$ [11]	$\mathcal{O}\big(\varepsilon^{-4}\big)$ [19]
	✓		$\mathcal{O}\big(\varepsilon^{-4}\big)$ [11]
$\operatorname{diam}(\mathcal{X}) \leq D_{\mathcal{X}}$	✗	$\Omega\big(\varepsilon^{-3}\big)$ (Theorem 9)	$\mathcal{O}\big(\varepsilon^{-4}\big)$ [20]
	✓	$\Omega\big(\varepsilon^{-3}\big)$ (Theorem 9)	$\mathcal{O}\big(\varepsilon^{-3}\big)$ [29]

Table 2: Gradient oracle complexity upper- and lower-bounds for reaching an ε-stationary point in the classical stochastic setting ($p = \nu = 2$) using (mini-batch) SGD with polynomially decaying stepsizes. The bounded-domain lower-bounds follow by choosing $\Delta_1 = \varepsilon D_{\mathcal{X}}$ in (19), guaranteeing that the iterates stay in a domain $\mathcal{X}$ with $\operatorname{diam}(\mathcal{X}) = {}^{\Delta_1}/_{\varepsilon} = D_{\mathcal{X}}$. To the best of our knowledge, the optimal sample complexity for the non-convex, bounded domain setting remains open even in this classical setting.

Discussion In the special case $p = \nu$, applying Theorem 9 with $\sigma = G$ establishes a sample complexity lower bound of

$$T = \Omega\left(\left(\frac{\Delta_1}{\eta\varepsilon^2}\right)^{\frac{1}{1-r}} + \eta^{\frac{1}{r}}\left(\frac{L_p G^p}{\varepsilon^2}\right)^{\frac{1}{r(p-1)}}\right), \tag{19}$$

confirming the tightness of our result in all parameters for unconstrained problems. Notably, our analysis demonstrates that $\eta_t \propto t^{-\frac{1}{p}}$ is the uniquely optimal polynomial decay. For constrained problems, the iterates of our lower-bound stay within a domain of diameter ${}^{\Delta_1}/_{\varepsilon}$, and hence confirm tightness of our upper-bound result whenever $D_{\mathcal{X}} \geq {}^{\Delta_1}/_{\varepsilon}$.

For domains with smaller diameter, the situation is more complicated, even in the classical $p = \nu = 2$ setting. There, our lower-bound — after choosing $\Delta_1 = \varepsilon D_{\mathcal{X}}$ to guarantee $\|x_t - x*\| \leq D_{\mathcal{X}}$ — reduces to $\Omega\big(\varepsilon^{-3}\big)$, while the upper-bound is of order $\mathcal{O}\big(\varepsilon^{-4}\big)$. For convex functions, this gap can be closed from above [29], for non-convex functions this question is still open even in this classical setting to the best of our knowledge. An overview can be found in Table 2.

Additionally note that, while Theorem 9 is derived for polynomial stepsizes for simplicity, it also holds for arbitrary stepsizes, scaling with $\sum_{t=1}^{T} \eta_t$ and $\min_{t\in[T]} \eta_t$ respectively.

Comparison to prior work The most closely related prior result established a lower bound of $\Omega(\min\left\{L_2^2\Delta_1^2, \sigma^4\right\}\varepsilon^{-4})$ iterations for SGD under standard smoothness and bounded variance assumptions [11]. We extend this prior complexity bound to the broader classes of Hölder-smooth ($\nu \in (1, 2]$) and heavy-tailed gradient noise distributions ($p \in (1, 2]$), recovering existing results in the classical $p = \nu = 2$ case as special instances. Furthermore, their construction crucially depends on a restrictive high-dimensionality assumption $d \geq \tilde{\Omega}(\sigma^2\Delta_1\varepsilon^{-4})$, limiting applicability. In contrast, our construction removes this dimensionality constraint entirely through a carefully designed gradient oracle. Finally, our analysis encompasses arbitrary stepsize sequences, whereas parts of the previous result requires specific schedules [11, Proposition 2].

Beyond SGD, the lower bounds for general first-order methods to reach an ε-stationary point have been studied extensively. In the classical ($p = \nu = 2$) setting, $\Omega(\varepsilon^{-2})$ samples are required for *convex* functions [16]. When $p \in (1, 2]$ and $\nu = 2$, [50] establish a $\Omega(\varepsilon^{-(3p-2)/(p-1)})$ lower-bound for non-convex functions, which we recover as special case when $\nu = 2$ for SGD. To the best of our knowledge, Hölder-smoothness has only been addressed in constrained convex optimization [3, 21], with bounds expressed in terms of suboptimality rather than gradient stationarity. Consequently, our results uniquely address the combination of Hölder-smoothness and heavy-tailed noise simultaneously, offering lower bounds applicable to both convex and nonconvex settings.

E.2. Upper-Bound for Central Moments

Previously, we studied vanilla SGD under Hölder smoothness and derived tight convergence rates along with a lower bound construction. While Hölder smoothness is weaker than standard smoothness ($\nu = 2$) when the set $\mathcal{X}$ is bounded, it is unclear if the standard smoothness can be utilized directly. We first recall the definition of standard smoothness.

Assumption 6 (Smoothness) *Assumption 4 holds with $\nu = 2$, i.e., there exist ℓ and L such that for all $x, y \in \mathcal{X}$*

$$-\tfrac{\ell}{2}\left\|x-y\right\|^2 \le F(x) - F(y) - \langle \nabla F(y), x-y \rangle \le \tfrac{L}{2}\left\|x-y\right\|^2 .$$

In this subsection, we consider $\nu = 2$, and will omit the subscript in constants $\ell := \ell_2$, $L := L_2$, and the superscript in $F_{1/\rho}(\cdot) := F^2_{1/\rho}(\cdot)$, $\hat{x} := \hat{x}^2$. That is, we have $F_{1/\rho}(x) := \min_{y\in\mathcal{X}} \left[F(y) + \frac{\rho}{2}\left\|y-x\right\|^2\right]$, $\hat{x} := \arg\min_{y\in\mathcal{X}} \left[F(y) + \frac{\rho}{2}\left\|y-x\right\|^2\right]$ for any $x \in \mathcal{X}$. We also refer to Definition 3 for corresponding definitions of $\mathcal{D}^2_\rho(x)$ and $\mathcal{S}^2_\rho(x)$, which coincide for $\nu = 2$. We also use the refined *central* moment noise assumption replacing (p-BM).

Assumption 7 *Let $F(\cdot)$ be differentiable on $\mathcal{X}$. We have access to stochastic gradients with $\mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla F(x)$ and there exists $p \in (1, 2]$ such that*

$$\textit{p-BCM} \qquad \mathbb{E}\left[\left\|\nabla f(x,\xi) - \nabla F(x)\right\|^p\right] \le \sigma^p \qquad \textit{for all } x \in \mathcal{X}.$$

Instead of vanilla SGD, we consider its mini-batch variant, which samples a mini-batch of i.i.d. stochastic gradients $\left\{\nabla f(x_t, \xi_t^i)\right\}_{i=1}^B$ at each iteration t and updates the decision variable via

$$\text{Mini-batch SGD:} \qquad x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t g_t), \qquad g_t = \tfrac{1}{B}\sum_{i=1}^{B} \nabla f(x_t, \xi_t^i)$$

Our analysis is based on the Lyapunov function inspired by the analysis in [15]:

$$\lambda_t := F(x_t) - F^* + \eta_{t-1}\rho(F_{1/\rho}(x_t) - F^*).$$

Theorem 12 *Let Assumptions 7, 5, and 4 hold with exponent $\nu = 2$ and curvature constants $\ell, L > 0$. Set $\rho = 2(L + 2\ell)$, and suppose Mini-batch SGD with constant step-size $\eta_t \equiv \eta = 1/2L$ is run with batch-size $B = \min\left\{1, \left(\frac{\rho\eta\sigma^2 T^{\frac{2}{p}}}{\lambda_1 L}\right)^{\frac{p}{2(p-1)}}\right\}$. Then we have $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\mathcal{S}^2_{\rho+L}(x_t)\right)^{\frac{p}{2}}\right] \le \left(\frac{36\lambda_1 L}{T}\right)^{\frac{p}{2}}$. In particular, the sample complexity to find a point x with $\mathbb{E}\left(\mathcal{S}^2_{\rho+L}(x)\right)^{\frac{p}{2}} \le \varepsilon^p$ is upper bounded by*

$$T \cdot B = \mathcal{O}\left(\frac{\lambda_1 L}{\varepsilon^2} + \left(\frac{\sigma\sqrt{(\ell+L)\lambda_1}}{\varepsilon^2}\right)^{\frac{p}{p-1}}\right).$$

Proof Define $\psi_t := g_t - \nabla F(x_t)$. By [15, Equation (15)], we have

$$\lambda_{t+1} \le \lambda_t - \tfrac{\rho\eta_t}{2(\rho+L)}\mathcal{D}^2_{\rho+L}(x_t) + \underbrace{\rho\eta_t\langle\psi_t, \hat{x}_t - x_t\rangle - \tfrac{\rho\eta_t(\rho-L)}{2}\left\|\hat{x}_t - x_t\right\|^2}_{(A)}$$
$$+ \underbrace{\rho\eta_t\langle\psi_t, x_t - x_{t+1}\rangle - \tfrac{\rho(1-\eta_t L)}{2}\left\|x_{t+1} - x_t\right\|^2}_{(B)}.$$

To control (A), first note that Young's inequality gives $\langle \psi_t, \hat{x}_t - x_t \rangle \le \frac{1}{2(\rho-L)} \|\psi_t\|^2 + \frac{\rho-L}{2} \|\hat{x}_t - x_t\|^2$ and hence

$$(A) \le \tfrac{\rho\eta_t}{2(\rho-L)} \|\psi_t\|^2 \le \tfrac{\rho\eta_t}{2L} \|\psi_t\|^2 ,$$

where we used $\rho - L \ge L$ in the last inequality. Using similar arguments and $\eta_t \le 1/2L$ also yields

$$(B) \le \tfrac{\rho\eta_t^2}{2(1-\eta_t L)} \|\psi_t\|^2 \le \tfrac{\rho\eta_t}{2L} \|\psi_t\|^2 ,$$

where the second inequality follows by Lemma 6 with $\nu = 2$. Combining the inequalities above, telescoping and using $\frac{1}{4} \le \frac{\rho}{2(\rho+L)}$ we obtain

$$\tfrac{1}{4} \sum_{t=1}^{T} \eta_t \mathcal{D}^2_{\rho+L}(x_t) \le \lambda_1 + \tfrac{\rho}{2L} \sum_{t=1}^{T} \eta_t \|\psi_t\|^2 .$$

Since the second moment of $\|\psi_t\|$ can be infinite, we cannot take expectation here. Instead we raise both sides of above inequality to power $p/2$ and derive

$$\left(\tfrac{1}{4} \sum_{t=1}^{T} \eta_t \mathcal{D}^2_{\rho+L}(x_t) \right)^{\frac{p}{2}} \le \left(\lambda_1 + \tfrac{\rho}{2L} \sum_{t=1}^{T} \eta_t \|\psi_t\|^2 \right)^{\frac{p}{2}} \le \lambda_1^{\frac{p}{2}} + \left(\tfrac{\rho}{2L}\right)^{\frac{p}{2}} \sum_{t=1}^{T} \eta_t^{\frac{p}{2}} \|\psi_t\|^p .$$

We can control the last term of above inequality in expectation using Lemma 7:

$$\mathbb{E} \|\psi_t\|^p \le \tfrac{2}{B^p} \sum_{i=1}^{B} \mathbb{E} \left\|\nabla f(x_t, \xi_t^i) - \nabla F(x_t)\right\|^p \le \tfrac{2\sigma^p}{B^{p-1}} .$$

Setting $\eta_t = \eta$ in the first, and choosing B as in the statement in a second step hence yields

$$\mathbb{E} \left(\tfrac{1}{T} \sum_{t=1}^{T} \mathcal{D}^2_{\rho+L}(x_t) \right)^{\frac{p}{2}} \le \left(\tfrac{4\lambda_1}{\eta T}\right)^{\frac{p}{2}} + \left(\tfrac{2\rho}{L}\right)^{\frac{p}{2}} \tfrac{2T\sigma^p}{T^{p/2} B^{p-1}} \le 3 \left(\frac{4\lambda_1}{\eta T} \right)^{\frac{p}{2}} .$$

The iteration complexity to reduce the stationarity measure,

$$\tfrac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[\mathcal{D}^2_{\rho+L}(x_t)^{\frac{p}{2}} \right] \le \mathbb{E} \left[\left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{D}^2_{\rho+L}(x_t) \right)^{\frac{p}{2}} \right] \le \varepsilon^p ,$$

can be upper bounded by

$$T \le \tfrac{36\lambda_1}{\eta\varepsilon^2} = 72 \tfrac{\lambda_1 L}{\varepsilon^2} ,$$

and the sample complexity directly follows, concluding the proof. ■

Discussion We can now compare this sample complexity result to previously derived iteration/sample complexity of vanilla SGD from Section 4 and Appendix E.1.2. For simplicity, we will compare the results in the unconstrained case, $\mathcal{X} = \mathbb{R}^d$ with $\ell = L$, when we have $\|\nabla F(\widetilde{x}_T)\|^p = (S^2_{\rho+L}(\widetilde{x}_T))^{p/2}$. Using Jensen's inequality, the fact that $\lambda_1 \le 3\Delta_1 = 3(F(x_1) - F^*)$, we can show that the above theorem implies that if $\widetilde{x}_T$ is sampled uniformly from the iterates of Mini-batch SGD, $\{x_t\}_{t \le T}$, then it satisfies

$$\mathbb{E} \|\nabla F(\widetilde{x}_T)\|^p \le \varepsilon^p \qquad \text{using} \qquad T \cdot B = \mathcal{O}\left(\tfrac{L\Delta_1}{\varepsilon^2} + \tfrac{(L\Delta_1)^{\frac{p}{2(p-1)}} \sigma^{\frac{p}{p-1}}}{\varepsilon^{\frac{2p}{p-1}}} \right) \qquad \text{samples.}$$

This sample complexity is in line with our guarantess for vanilla SGD from Theorem 4, which implies

$$\mathbb{E}\,\|\nabla F(\widetilde{x}_T)\|^2 \le \varepsilon^2 \qquad \text{using} \qquad T = \mathcal{O}\left(\frac{\Delta_1 L_p^{\frac{1}{p-1}} G^{\frac{p}{p-1}}}{\varepsilon^{\frac{2p}{p-1}}}\right) \qquad \text{samples.}$$

While the dependence on ε is the same, there is a potential improvement in the smoothness, L, and the moment, σ, parameters. We should also remark that the convergence criterion for SGD under Hölder smoothness is stronger compared to the one for Mini-batch SGD since $\mathbb{E}\,\|\nabla F(\widetilde{x}_T)\|^p \le \mathbb{E}\,\|\nabla F(\widetilde{x}_T)\|^2$. It is important to note that we do not have tightness result for sample complexity of Mini-batch SGD, since our oracle construction in Section E.1.2 is specifically designed for SGD, and does not extend to Mini-batch SGD. We believe our sample complexity of Mini-batch SGD above is unimprovable without use of adaptive methods, e.g., normalization and gradient clipping, but a more complex non-convex hard instance construction is required to prove tightness of this sample complexity.

Unfortunately, it remains unclear to us how to analyze in-expectation convergence of vanilla SGD (without mini-batch) under standard smoothness ($\nu = 2$). The main technical obstacle is that when $B = 1$, we need to use unbiasedness of the term (A) in the proof above. However, we cannot directly take the expectation since the term (B) may not have a finite expectation when $p < 2$.

E.3. Experiments

We consider a constrained convex (or strongly convex) optimization problem of the form

$$\min_{x\in\mathcal{X}} F(x) := \|Ax - b\|_1 + \frac{\mu}{2}\|x\|_2^2, \qquad \mathcal{X} := \{x \in \mathbb{R}^d : \|x\|_\infty \le R\},$$

where $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, $\mu \ge 0$ is a regularization parameter, $R > 0$ is a fixed radius of the ℓ_∞ ball. To simulate heavy-tailed noise, we augment the (sub)gradient of $F(\cdot)$ with synthetic noise drawn from a two-sided Pareto distribution. Specifically, we define the stochastic gradient oracle as

$$\nabla f(x_t, \xi_t) := A^\top \operatorname{sign}(Ax - b) + \mu x + \xi_t.$$

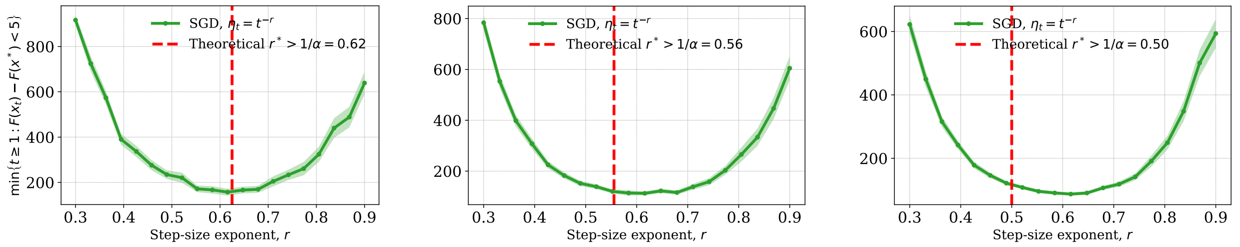
where $\xi_t \in \mathbb{R}^d$ is an i.i.d. heavy-tailed noise vector generated as

$$(\xi_t)_i = s_i \cdot u_i^{-\frac{1}{\alpha}}, \quad s_i \sim \mathrm{Unif}(\{-1, 1\}), \quad u_i \sim \mathrm{Unif}(0, 1),$$

with a tail index parameter $\alpha \in (1, 2]$. Notice that the above two-sided Pareto distribution has all moments $p \in (1, \alpha)$ finite, while all moments larger or equal to α are infinite. This distribution is chosen to simulate the noise with infinite variance satisfying Assumptions 1 and 7.

Parameters and evaluation. We fix the problem dimension to $d = 10$, generate matrix $A \sim \mathcal{N}(0, 1)^{d\times d}$ and vector $b \sim \mathcal{N}(0, 1)^d$ once for all experiments, and set the initial point to $x_0 = 100 \cdot \mathbf{1}_d$. The feasible region radius is set to $R = 10$, and experiments are run for $T = 1000$ iterations. The performance is evaluated based on 200 independent runs of each algorithm reporting the average and one standard deviation across iterations.

Experiment 1. Sensitivity of convergence rate to step-size choice. In our Theorem 1 for convex and Theorem 4 for non-convex cases, the guarantee is established for any non-negative step-size sequence. However, the reccomended step-size order (ignoring parameters G, $D_{\mathcal{X}}$, ℓ and L) is $\eta_t = 1/\sqrt[p]{t}$. In this experiment we aim to study the predictive power of our theory in a numerical experiment by varying different orders of the step-size. We set $\mu = 0$ and compare SGD with step-size $\eta_t = 1/t^r$ for 20 different values of



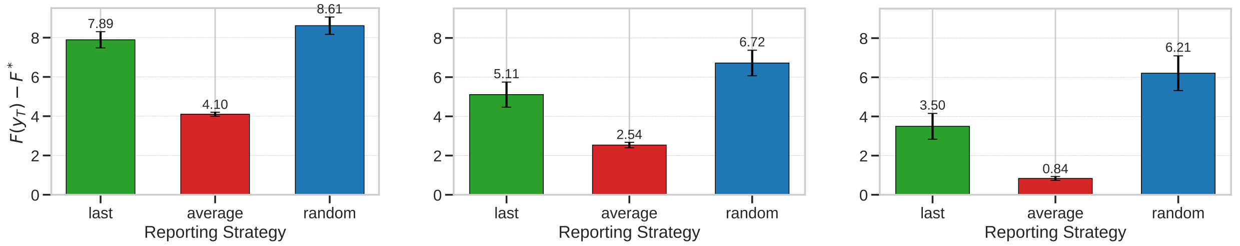
(a) $\alpha = 1.6$

(b) $\alpha = 1.8$

(c) $\alpha = 2.0$

Figure 2: Sensitivity of convergence rate to the step-size power r in $\eta_t = 1/t^r$ for different values of the heavy-tail index α. The minimal theoretical value for the optimal power $r = 1/\alpha$ is highlighted in each plot in red, and we can see that this value is often close the experimentally determined best value.

$r \in [0.3, 0.9]$. The sensitivity plot on Figure 2 shows how many iterations/samples T it takes to reach the accuracy level $F(x_T) - F(x^*) \leq 5$. vs. the step-size power r. The red dashed line indicates the minimal theoretically reccomended value $r = 1/\alpha$. All three plots show that SGD is not very sensitive to the choice of step-size order r, and the value of $r = 1/\alpha$ is often close the experimentally determined best value (with lowest possible number of iterations).



(a) $\alpha = 1.2$

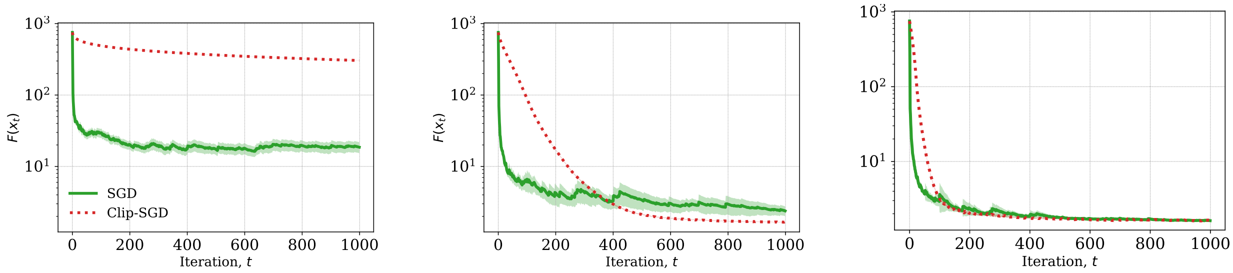
(b) $\alpha = 1.6$

(c) $\alpha = 2.0$

Figure 3: Comparison of output selection strategies in stochastic gradient descent (SGD) under convexity ($\mu = 0$) and heavy-tailed noise. We evaluate the final suboptimality $F(y_T) - F^*$, where y_T is the reported output, using three strategies: the last iterate, the average of all iterates, and a uniformly sampled iterate. Averaging the iterates yields the lowest suboptimality, significantly outperforming both the last iterate and random selection.

Experiment 2. Effect of output selection in SGD under heavy-tailed noise. As we have seen, depending on convexity assumptions, our theory suggests different output strategies for SGD. It is known that in convex setting under bounded variance assumption $p = 2$, the last iterate converges with optimal complexity [49], however, our Theorem 1 requires to output the average iterate. In strongly convex setting, to obtain the optimal convergence in function value, it is typical to output the average iterate [44], however, interestingly our Theorem 2 requires randomly sampled output for $p < 2$. To investigate the impact of these different strategies, we compare the function suboptimality of each for different noise levels. In this experiment, the step-size is set to $\eta_t = 1/\sqrt{t}$, as motivated by standard theory in the convex setting. After $T = 1000$ iterations, we compare three strategies for producing the final output y_T: (i) the last iterate x_T, (ii) the average $\widetilde{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$, and (iii) a randomly sampled iterate $\bar{x}_T \sim \mathrm{Uniform}\{x_1, \ldots, x_T\}$.

For each strategy, we measure the expected suboptimality $F(y_T) - F^*$, where F^* is the optimal objective value computed deterministically without noise. As shown in Figure 3, the average iterate significantly outperforms the other two strategies. This highlights the stabilizing effect of averaging in the presence of heavy-tailed stochastic gradients and is in line with theoretical guarantee in Theorem 1.



(a) $\alpha = 1.2$ (b) $\alpha = 1.6$ (c) $\alpha = 2.0$

Figure 4: Comparison of SGD and Clip-SGD in the strongly convex setting for different values of the heavy-tail index $\alpha \in \{1.2, 1.6, 2.0\}$. Both algorithms use the same diminishing step-size schedule $\eta_t = 1/(\mu t)$, where $\mu > 0$ is the strong convexity parameter. For Clip-SGD, the gradient clipping threshold is set as $\lambda_t = t^{\alpha-1}$ without tuning, following theoretical recommendations from [42, 50]. Results show that Clip-SGD improves the overall convergence stability under heavy-tailed noise, especially when α is small (e.g., $\alpha = 1.2, 1.6$), where standard SGD suffers from high variance. However, in the initial optimization phase, SGD often outperforms its clipped variant thanks to larger update steps.

Experiment 3. Comparison to Clip-SGD in strongly convex setting. In this experiment, we study the impact of gradient clipping in the strongly convex setting under heavy-tailed noise. Specifically, we compare standard (projected) stochastic gradient descent (SGD) with its clipped variant (Clip-SGD). The algorithms are configured with the same step-size $\eta_t = 1/(\mu t)$ as follows:

$$\text{SGD:} \qquad x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \eta_t \nabla f(x_t, \xi_t)\right),$$

$$\text{Clip-SGD:} \qquad x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \eta_t g_t\right), \quad g_t = \mathrm{clip}\left(\nabla f(x_t, \xi_t), \lambda_t\right)$$

with $\mathrm{clip}(v, \lambda) := v \cdot \min\{1, \lambda/\|v\|_2\}$. In Clip-SGD, the clipping thresholds are set as $\lambda_t = t^{\alpha-1}$ based on the theoretical analysis in [42, 50], with no tuning.

Figure 4 presents the evolution of the mean and the standard deviation of the objective value $F(x_t)$ over iterations for each method. As expected, Clip-SGD significantly reduces the variance of SGD across all noise levels. In some situations, e.g., $\alpha = 1.6$, this allows Clip-SGD to outperform SGD after $t \geq 400$ iterations, where vanilla SGD suffers from large variance and slow convergence. On the other hand, the experiment suggest that SGD is sometimes competitive and can even outperform Clip-SGD. First, we observe that when α increases toward the light-tailed regime ($\alpha = 2.0$), the performance gap narrows, confirming that clipping is especially beneficial under heavy-tailed stochasticity. Second, perhaps most surprisingly, in the most heavy-tailed regime $\alpha = 1.2$ and in the early phase of medium regime $\alpha = 1.6$, SGD can significantly outperform Clip-SGD. This happens because Clip-SGD is initially making very small steps, perhaps due to untuned clipping sequence $\lambda_t = t^{\alpha-1}$, while SGD can make large steps, quickly converging to a certain noise level.