Online Convex Optimization with Heavy Tails: Old Algorithms, New Regrets, and Applications

Anonymous Author(s)
Affiliation
Address
email

Abstract

In Online Convex Optimization (OCO), when the stochastic gradient has a finite variance, many algorithms provably work and guarantee a sublinear regret. However, limited results are known if the gradient estimate has a heavy tail, i.e., the stochastic gradient only admits a finite p-th central moment for some $\mathsf{p} \in (1, 2]$. Motivated by it, this work examines different old algorithms for OCO (e.g., Online Gradient Descent) in the more challenging heavy-tailed setting. Under the standard bounded domain assumption, we establish new regrets for these classical methods without any algorithmic modification. Remarkably, these regret bounds are fully optimal in all parameters (can be achieved even without knowing p), suggesting that OCO with heavy tails can be solved effectively without any extra operation (e.g., gradient clipping). Our new results have several applications. A particularly interesting one is the first provable convergence result for nonsmooth nonconvex optimization under heavy-tailed noise without gradient clipping.

1 Introduction

This paper studies the online learning problem with convex losses, also known as Online Convex Optimization (OCO), a widely applicable framework that learns under streaming data [4, 10, 27, 35]. OCO has tons of implications for both designing and analyzing algorithms in different areas, for example, stochastic optimization [8, 23, 14], PAC learning [3], control theory [1, 11], etc.

In an OCO problem, a learning algorithm A would interact with the environment in T rounds, where $T \in \mathbb{N}$ can be either known or unknown. Formally, in each round round t, the learner A first decides an output $\boldsymbol{x}_t \in \mathcal{X}$ from a convex feasible set $\mathcal{X} \subseteq \mathbb{R}^d$, then the environment reveals a convex loss function $\ell_t : \mathcal{X} \to \mathbb{R}$, and A incurs a loss of $\ell_t(\boldsymbol{x}_t)$. After T many rounds, the quantity measuring the algorithm's performance is called regret, defined relative to any fixed competitor $\boldsymbol{x} \in \mathcal{X}$ as follows:

$$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) \triangleq \sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}).$$

In the classical setting, instead of observing full information about ℓ_t, the learner A is only guaranteed to receive a subgradient $\nabla \ell_t(\boldsymbol{x}_t) \in \partial \ell_t(\boldsymbol{x}_t)$ at its decision, where $\partial \ell_t(\boldsymbol{x}_t)$ denotes the subdifferential set of ℓ_t at $\boldsymbol{x}_t$ [33]. This turns out to be enough for our purpose of minimizing the regret, since any OCO problem can be reduced to an Online Linear Optimization (OLO) instance via the inequality $\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}) \le \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x} \rangle$, which holds due to convexity. Under the standard bounded domain assumption, i.e., $\mathcal{X}$ has a finite diameter D, many classical algorithms, e.g., Online Gradient Descent (OGD) [50], guarantee an optimal sublinear regret $GD\sqrt{T}$ for G-Lipschitz ℓ_t. Even better, in the case that computing an exact subgradient is intractable, and one could only query a stochastic estimate $\boldsymbol{g}_t$ satisfying $\mathbb{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] \in \partial \ell_t(\boldsymbol{x}_t)$, the OGD algorithm can still solve OCO effectively with a provable $(G+\sigma)D\sqrt{T}$ regret bound in expectation if the stochastic noise $\boldsymbol{g}_t - \nabla \ell_t(\boldsymbol{x}_t)$ has a bounded second moment σ^2 for some $\sigma \geq 0$, which is called the finite variance condition.

However, many works have pointed out that even for the easier stochastic optimization (i.e., $\ell_t = F$ for a common F), the typical finite variance assumption is too optimistic and can be violated in different tasks [12, 37, 45], and their observations suggest that the stochastic gradient only admits a finite p-th central moment upper bounded by σ^{p} for some $\mathsf{p} \in (1,2]$, which is named heavy-tailed noise. This new assumption generalizes the classical finite variance condition ($\mathsf{p}=2$) and becomes challenging when $\mathsf{p} < 2$. A particular evidence is that the famous Stochastic Gradient Descent (SGD) algorithm [32] (which is exactly OGD for stochastic optimization) provably diverges [45].

Though heavy-tailed stochastic optimization has been extensively studied [18, 26, 34], limited results are known for OCO with heavy tails. The only work under this topic that we are aware of is [47], which established a parameter-free regret bound in high probability (more discussions provided later). However, their algorithm includes many nontrivial modifications like gradient clipping and significantly deviates from the existing simple OCO algorithms used in practice. Especially, consider OGD as an example. Though the heavy-tailed issue is known, OGD (or just think of it as SGD) still works (sometimes very well) in practice even without gradient clipping and is arguably one of the most popular optimizers, which seemingly contradicts the theory of unconvergence mentioned before. This indicates that, for classical OCO algorithms under heavy-tailed noise, a huge gap exists between the empirical convergence (or even the effective practical performance) and theoretical guarantees. Therefore, we are naturally led to the following question:

In what context can old OCO algorithms work under heavy tails, in what sense, and to what extent?

1.1 Contributions

Motivated by the above question, we examine three classical algorithms for OCO: Online Gradient Descent (OGD) [50], Dual Averaging (DA) [25, 43], and AdaGrad [9, 22], and answer it as follows:

Under the standard bounded domain assumption, the in-expectation regret $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]$ *is finite and optimal for any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$*, without any algorithmic modification.*

In detail, our new results for heavy-tailed OCO are summarized here:

- We prove the only and the first optimal regret bound $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$ for any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$. Remarkably, AdaGrad can achieve this result without knowing any of the Lipschitz parameter G, noise level σ, and tail index p.
- We extend the analysis of OGD to Online Strongly Convex Optimization with heavy tails and establish the first provable result $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2 \log T}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} T^{2-\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$, where $\mu > 0$ is the modulus of strong convexity and T^0 should be read as $\log T$.

Based on the new regret bounds for OCO with heavy tails, we provide the following applications:

- For nonsmooth convex optimization with heavy tails, we show the first optimal in-expectation rate $GD/\sqrt{T} + \sigma D / T^{1-1/\mathsf{p}}$ achieved without gradient clipping, which applies to both the average iterate and last iterate, demonstrating that SGD does converge once the domain is bounded.
- For nonsmooth nonconvex optimization with heavy tails, we show the first provable sample complexity of $G^2\delta^{-1}\epsilon^{-3} + \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ for finding a (δ,ϵ)-stationary point without gradient clipping. Moreover, we give the first convergence result when the problem-dependent parameters (like G, σ, and p) are unknown in advance.

1.2 Discussion on [47]

As noted, [47] is the only work for OCO with heavy tails, as far as we know. There are two major discrepancies between them and us. First, they consider the case where the feasible set $\mathcal{X}$ is unbounded and aim to establish a parameter-free regret bound, i.e., the regret bound has a linear dependency on $\|\boldsymbol{x}\|$ (up to an extra $\operatorname{polylog}\|\boldsymbol{x}\|$) for any competitor $\boldsymbol{x} \in \mathcal{X}$. Second, they focus on high-probability rather than in-expectation analysis. As such, their regret is in the form of

$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) \lesssim (G+\sigma)\|\boldsymbol{x}\| T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$ (up to extra polylogarithmic factors) with high probability. Without a doubt, their setting is harder than ours implying their bound is stronger as it can convert to an in-expectation regret $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim (G+\sigma) D T^{1/\mathsf{p}}$ for any bounded domain $\mathcal{X}$ with a diameter D.

We emphasize that the motivation behind [47] differs heavily from ours. They aim to solve heavy-tailed OCO with a new proposed method that contains many nontrivial technical tricks, including gradient clipping, artificially added regularization, and solving the additional fixed-point equation. However, their result cannot reflect why the existing simple OCO algorithms like OGD work in practice under heavy-tailed noise. In contrast, our goal is to examine whether, when, and how the classical OCO algorithms work under heavy tails, thereby filling the missing piece in the literature.

Moreover, we would like to mention two drawbacks of [47]. First, though the $T^{1/\mathsf{p}}$ regret seems tight as it matches the lower bound [24, 30, 41], this may not be the best, since an optimal bound should recover the standard $\sqrt{T}$ regret in the deterministic case (i.e., $\sigma = 0$), as one can imagine. This suggests that their bound is not entirely optimal. Second, we remark that they require knowing both problem-dependent parameters G, σ, p and time horizon T in the algorithm, which may be hard to satisfy in the online setting. In comparison, our regret bound $GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}$ is fully optimal in all parameters. Importantly, $\mathsf{AdaGrad}$ can achieve it while oblivious to the problem information.

2 Preliminary

Notation. $\mathbb{N}$ denotes the set of natural numbers (excluding 0). $[T] \triangleq \{1, \dots, T\}, \forall T \in \mathbb{N}$. $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. We write $a \lesssim b$ if $a \le Cb$ for a universal constant $C > 0$. $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ respectively represent the floor and ceiling functions. $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product and $\|\cdot\| \triangleq \sqrt{\langle \cdot, \cdot \rangle}$ is the standard 2-norm. Given $\boldsymbol{x} \in \mathbb{R}^d$ and $D > 0$, $\mathcal{B}^d(\boldsymbol{x}, D)$ is the Euclidean ball in $\mathbb{R}^d$ centered at $\boldsymbol{x}$ with a radius D. In the case $\boldsymbol{x} = \mathbf{0}$, we use the shorthand $\mathcal{B}^d(D)$. Given a nonempty closed convex set $A \subseteq \mathbb{R}^d$, Π_A is the Euclidean projection operator onto A. For a convex function f, $\partial f(\boldsymbol{x})$ denotes its subgradient set at $\boldsymbol{x}$.

Remark 1. We choose the Euclidean norm only for simplicity. Extending the results in this work to any general norm is straightforward.

This work studies OCO in the context of Assumption 1.

Assumption 1. *We consider the following series of assumptions:*

- $\mathcal{X} \subset \mathbb{R}^d$ *is a nonempty closed convex set bounded by* D, *i.e.,* $\sup_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{y}\| \le D$.
- $\ell_t : \mathcal{X} \to \mathbb{R}$ *is convex for all* $t \in [T]$.
- ℓ_t *is* G*-Lipschitz on* $\mathcal{X}$, *i.e.,* $\|\nabla \ell_t(\boldsymbol{x})\| \le G, \forall \boldsymbol{x} \in \mathcal{X}, \nabla \ell_t(\boldsymbol{x}) \in \partial \ell_t(\boldsymbol{x})$, *for all* $t \in [T]$.
- *Given a point* $\boldsymbol{x}_t \in \mathcal{X}$ *at the* t*-th iteration, one can query* $\boldsymbol{g}_t \in \mathbb{R}^d$ *satisfying* $\nabla \ell_t(\boldsymbol{x}_t) \triangleq \mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right] \in \partial \ell_t(\boldsymbol{x}_t)$ *and* $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$ *for some* $\mathsf{p} \in (1, 2]$ *and* $\sigma \ge 0$, *where* $\mathcal{F}_t \triangleq \sigma(\boldsymbol{g}_1, \dots, \boldsymbol{g}_t)$ *denotes the natural filtration and* $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla \ell_t(\boldsymbol{x}_t)$ *is the stochastic noise.*

Remark 2. D is recognized as known, like ubiquitously assumed in the OCO literature. Moreover, $\boldsymbol{x}_t$ denotes the decision/output of the online learning algorithm by default.

In Assumption 1, the first three points are standard, and the fourth is the heavy-tailed noise assumption. In particular, $\mathsf{p} = 2$ recovers the standard finite variance condition.

3 Old Algorithms under Heavy Tails

In this section, we revisit three classical algorithms for OCO: OGD, DA, and $\mathsf{AdaGrad}$, whose regret bounds are well-studied in the finite variance case but remain unknown under heavy-tailed noise.

The basic idea of proving these algorithms work under heavy tails is to leverage the boundness property of $\mathcal{X}$. We will describe it in more detail using OGD as an illustrated example. The analysis of DA follows a similar way at a high level, but differs in some details. However, though $\mathsf{AdaGrad}$ can be viewed as OGD with an adaptive stepsize, the way to utilize the boundness property is entirely different. All formal proofs are deferred to the appendix due to space limitations.

3.1 New Regret for Online Gradient Descent

Algorithm 1 Online Gradient Descent (OGD) [50]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta_t > 0$
for $t = 1$ **to** T **do**
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t)$
end for

We begin from arguably the most basic algorithm for OCO, Online Gradient Descent (OGD).

A well known analysis. The regret bound of OGD has been extensively studied [10, 27, 35]. The most well known analysis is perhaps the following one: for any $\boldsymbol{x} \in \mathcal{X}$, there is

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 = \|\Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) - \Pi_{\mathcal{X}}(\boldsymbol{x})\|^2 \le \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{x}\|^2,$$

where the inequality holds by the nonexpansive property of $\Pi_{\mathcal{X}}$. Expanding both sides and rearranging terms yield that

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \frac{\eta_t \|\boldsymbol{g}_t\|^2}{2}. \tag{1}$$

If $\boldsymbol{g}_t$ admits a finite variance, i.e., $\mathsf{p} = 2$ in Assumption 1, taking expectations on both sides, then following a standard analysis for $\eta_t = \frac{D}{(G+\sigma)\sqrt{t}}$ (or $\eta_t = \frac{D}{(G+\sigma)\sqrt{T}}$ if T is known) gives the regret

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim (G+\sigma) D\sqrt{T}, \forall \boldsymbol{x} \in \mathcal{X}.$$

However, the step of taking expectations on the R.H.S. of (1) crucially relies on the finite variance condition of $\boldsymbol{g}_t$. Therefore, one may naturally think OGD would not guarantee a finite regret if $\mathsf{p} < 2$.

A less well known analysis[1]. As discussed, the failure of the above proof under heavy-tailed noise is due to (1). Therefore, if a tighter inequality than (1) exists, then it might be possible to show that OGD still works for $\mathsf{p} < 2$. However, does it exist?

Actually, there is another less well known analysis to produce a better inequality than (1). That is, first showing for any $\boldsymbol{x} \in \mathcal{X}$, by the optimality condition of the update rule,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle \le \frac{\langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle}{\eta_t} = \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t},$$

and then obtaining

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t}. \tag{2}$$

Note that (2) is tighter than (1) as $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \le \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le \frac{\eta_t \|\boldsymbol{g}_t\|^2}{2} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t}$, where the first step is due to Cauchy-Schwarz inequality and the second one is by AM-GM inequality.

Handle $\mathsf{p} < 2$ in a simple way. Though we have tightened (1) into (2), can inequality (2) help to overcome heavy tails? The answer is surprisingly positive, and our solution is fairly simple. Instead of directly applying AM-GM inequality in the second step, we recall $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t$ and use triangle inequality to obtain

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \le \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le (\|\nabla \ell_t(\boldsymbol{x}_t)\| + \|\boldsymbol{\epsilon}_t\|) \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|. \tag{3}$$

On the one hand, by $\|\nabla \ell_t(\boldsymbol{x}_t)\| \le G$ and AM-GM inequality, there is

$$\|\nabla \ell_t(\boldsymbol{x}_t)\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le G \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le \eta_t G^2 + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t}. \tag{4}$$

[1] To clarify, the phrase "less well known" is compared to the first one. This analysis itself is also well known.

On the other hand, let $\mathsf{p}_\star \triangleq \frac{\mathsf{p}}{\mathsf{p}-1}$ and $\mathsf{C}(\mathsf{p}) \triangleq \frac{(4\mathsf{p}-4)^{\mathsf{p}-1}}{\mathsf{p}^{\mathsf{p}}}$, we have

$$\begin{aligned}
\|\boldsymbol{\epsilon}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| &= \left(\frac{4\eta_t}{\mathsf{p}_\star}\right)^{\frac{1}{\mathsf{p}_\star}} \|\boldsymbol{\epsilon}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^{1-\frac{2}{\mathsf{p}_\star}} \cdot \left(\frac{\mathsf{p}_\star \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t}\right)^{\frac{1}{\mathsf{p}_\star}} \\
&\overset{(a)}{\leq} \frac{\left(\frac{4\eta_t}{\mathsf{p}_\star}\right)^{\frac{\mathsf{p}}{\mathsf{p}_\star}} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^{\mathsf{p}-\frac{2\mathsf{p}}{\mathsf{p}_\star}}}{\mathsf{p}} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t} \\
&\overset{(b)}{\leq} \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t}, \qquad (5)
\end{aligned}$$

where (a) is by Young's inequality and (b) is due to $\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \leq D$, $\mathsf{p}_\star = \frac{\mathsf{p}}{\mathsf{p}-1}$, and $\mathsf{C}(\mathsf{p}) = \frac{(4\mathsf{p}-4)^{\mathsf{p}-1}}{\mathsf{p}^{\mathsf{p}}}$. Next, we plug (4) and (5) back into (3), then combine with (2) to know

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \qquad (6)$$

Notably, the term $\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}$ has a correct exponent p. Thus, we can safely take expectations on both sides. Finally, a standard analysis yields the following Theorem 1 (see Appendix A for a formal proof).

Theorem 1. *Under Assumption 1, taking $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

As far as we know, Theorem 1 is the first and the only provable result for OGD under heavy tails. Remarkably, it is not only tight in T [24, 30, 41] but also fully optimal in all parameters, in contrast to the bound $(G + \sigma)DT^{1/\mathsf{p}}$ of [47]. This reveals that OCO with heavy tails can be optimally solved as effectively as the finite variance case once the domain is bounded, a classical condition adapted in many existing works.

Strongly convex functions. We highlight that the above idea can also be applied to Online Strongly Convex Optimization and leads to a sublinear regret $T^{2-\mathsf{p}}$ better than $T^{1/\mathsf{p}}$. This extension can be found in Appendix A.

3.2 New Regret for Dual Averaging

Algorithm 2 Dual Averaging (DA) [25, 43]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta_t > 0$
for $t = 1$ **to** T **do**
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_1 - \eta_t \sum_{s=1}^{t} \boldsymbol{g}_s)$
end for

Remark 3. It is known that DA is a special realization of the more general Follow-the-Regularized-Leader (FTRL) framework [21]. To keep the work concise, we only focus on DA. The key idea to prove Theorem 2 can directly extend to show new regret for FTRL under heavy-tailed noise.

We turn our attention to the second candidate, the Dual Averaging (DA) algorithm, which is given in Algorithm 2. Though DA coincides with OGD when $\mathcal{X} = \mathbb{R}^d$ and $\eta_t = \eta$, these two methods in general are not equivalent and can have significant performance differences in practice. Therefore, it is also important to understand DA under heavy tails.

Despite the proof strategies for OGD and DA are in different flavors (even for $\mathsf{p} = 2$), the basic idea presented before for OGD still works here, i.e., apply the boundness property of $\mathcal{X}$ to make the term $\|\boldsymbol{\epsilon}_t\|$ have a correct exponent. Armed with this thought, we can prove the following new regret bound for DA under heavy-tailed noise. We refer the reader to Appendix B for its proof.

Theorem 2. *Under Assumption 1, taking $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ in* DA *(Algorithm 2), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

As far as we know, Theorem 2 is the first provable and optimal regret for DA under heavy tails. It guarantees the same tight bound as in Theorem 1 up to different constants.

3.3 New Regret for AdaGrad

Algorithm 3 AdaGrad [9, 22]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta > 0$
for $t = 1$ **to** T **do**
  $\eta_t = \eta V_t^{-1/2}$ where $V_t = \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2$
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t)$
end for

Remark 4. Algorithm 3 is also named AdaGrad-Norm (e.g., [42]). We simply call it AdaGrad. It is straightforward to generalize Theorem 3 below to the per-coordinate update version.

Although Theorems 1 and 2 are optimal, they both suffer from an undesired point. That is, the stepsize $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ requires knowing all problem-dependent parameters. However, it may not be easy to obtain them in an online setting. Especially, it heavily depends on the prior information about the tail index p, which is hard to know (even approximately) in advance. In other words, they both lack the adaptive property to an unknown environment.

To handle this issue, we consider AdaGrad, a classical adaptive algorithm for OCO. As can be seen, AdaGrad is just OGD with an adaptive stepsize. However, it is this adaptive stepsize that can help us to overcome the above undesired point.

Theorem 3. *Under Assumption 1, taking $\eta = D/\sqrt{2}$ in* AdaGrad *(Algorithm 3), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{AdaGrad}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Remark 5. We also establish a similar result for DA with an adaptive stepsize. See Theorem 7 in Appendix B for details.

Theorem 3 provides the first regret bound for AdaGrad under heavy tails. Impressively, it is optimal even without knowing any of G, σ, and p. This surprising result once again demonstrates the power of the adaptive method, indicating it is robust to an unknown environment and even heavy-tailed noise, which may partially explain the favorable performance of many adaptive optimizers designed based on AdaGrad like RMSProp [40] and Adam [14].

We point out that the key to establishing Theorem 3 differs from the idea used before for OGD and DA. Actually, Theorem 3 can be obtained in an embarrassingly simple way. It is known that AdaGrad with $\eta = D/\sqrt{2}$ on a bounded domain guarantees the following path-wise regret

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \lesssim D\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2}. \tag{7}$$

Observe that $\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2} \lesssim \sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \sqrt{\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^2} \le G\sqrt{T} + \left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}$, where the last step is due to $\|\cdot\|_2 \le \|\cdot\|_{\mathsf{p}}$ for any $\mathsf{p} \in [1, 2]$. After taking expectations on both sides of (7) and applying Hölder's inequality to obtain $\mathbb{E}\left[\left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \le \left(\sum_{t=1}^{T} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \le \sigma T^{\frac{1}{\mathsf{p}}}$, we conclude Theorem 3. To make the work self-consistent, we produce the formal proof of Theorem 3 in Appendix C.

4 Applications

We provide some applications based on the new regret bounds established in Section 3. The basic problem we study is optimizing a single objective F, which could be either convex or nonconvex.

4.1 Nonsmooth Convex Optimization

In this section, we consider nonsmooth convex optimization with heavy tails.

Convergence of the average iterate. First, we focus on convergence in average. By the classical online-to-batch conversion [3], the following corollary immediately holds.

Corollary 1. *Under Assumption 1 for* $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x} \rangle$ *and let* $\bar{\boldsymbol{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$*, for any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$*, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x})\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]}{T} \lesssim \frac{GD}{\sqrt{T}} + \frac{\sigma^{\mathsf{p}} D}{T^{1-\frac{1}{\mathsf{p}}}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. By convexity, $F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}) \le \frac{\sum_{t=1}^{T} F(\boldsymbol{x}_t) - F(\boldsymbol{x})}{T} \le \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})}{T}$ is valid for any OCO algorithm A. We conclude from invoking Theorems 1, 2 and 3. □

To the best of our knowledge, Corollary 1 gives the first and optimal convergence rate for these three algorithms in stochastic optimization with heavy tails. Especially, it implies that once the domain is bounded, the widely implemented SGD algorithm provably converges under heavy-tailed noise without any algorithmic change considered in many prior works, e.g., gradient clipping [18, 26].

We are only aware of two works [19, 41] based on Stochastic Mirror Descent (SMD) [24] that gave convergence results without clipping. However, they share a common drawback, i.e., their bounds are both in the form of $(G+\sigma)D/T^{1-1/\mathsf{p}}$, which cannot recover the optimal rate $GD/\sqrt{T}$ when $\sigma = 0$.

Lastly, we highlight that for $\mathsf{A} = \mathsf{AdaGrad}$, Corollary 1 is not only optimal but also adaptive to the tail index p. As far as we know, no result has achieved this property before. This once again evidences the benefit of adaptive gradient methods.

Convergence of the last iterate. Next, we consider the more challenging last-iterate convergence, which has a long history in stochastic optimization and fruitful results in the case of $\mathsf{p} = 2$ (see, e.g., [28, 36, 49]). However, less is known about heavy-tailed problems. So far, only two works [19, 29] have established the last-iterate convergence. The former is based on SMD, and the latter employs gradient clipping in SGD. Unfortunately, their rates are both in the suboptimal order $(G+\sigma)D/T^{1-1/\mathsf{p}}$.

We will provide an optimal last-iterate rate based on the following lemma, which reduces the last-iterate convergence to an online learning problem.

Lemma 1 (Theorem 1 of [7])**.** *Suppose* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ *and* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ *are two sequences of vectors satisfying* $\boldsymbol{x}_t \in \mathcal{X}$, $\boldsymbol{x}_1 = \boldsymbol{y}_1$ *and*

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \frac{T-t}{T}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right). \quad (8)$$

Given a convex function $F(\boldsymbol{x})$*, let* $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{y}_t), \boldsymbol{x} \rangle$*. Then for any online learner* A*, we have*

$$F(\boldsymbol{y}_T) - F(\boldsymbol{x}) \le \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})}{T}, \forall \boldsymbol{x} \in \mathcal{X}.$$

We emphasize that the stochastic gradient $\boldsymbol{g}_t$ received by A is an estimate of $\nabla F(\boldsymbol{y}_t)$ instead of $\nabla F(\boldsymbol{x}_t)$. This flexibility is due to the generality of the OCO framework. Moreover, for OGD, suppose there is no projection step, then (8) is equivalent to $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \frac{T-t}{T}\eta_t \boldsymbol{g}_t$, which can be viewed as SGD with a stepsize $\frac{T-t}{T}\eta_t$. For proof of Lemma 1, we refer the interested reader to [7].

Corollary 2. *Under Assumption 1 for* $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{y}_t), \boldsymbol{x} \rangle$*, where* $\boldsymbol{y}_t$ *satisfies (8), for any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$*, we have*

$$\mathbb{E}\left[F(\boldsymbol{y}_T) - F(\boldsymbol{x})\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]}{T} \lesssim \frac{GD}{\sqrt{T}} + \frac{\sigma^{\mathsf{p}} D}{T^{1-\frac{1}{\mathsf{p}}}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. Combine Lemma 1 and Theorems 1, 2 and 3 to conclude. □

As far as we know, Corollary 2 is the first optimal last-iterate convergence rate for stochastic convex optimization with heavy tails, closing the gap in existing works.

One may notice that $\boldsymbol{y}_t$ itself is not the decision made by the online learner and naturally may ask whether $\boldsymbol{x}_t$ ensures the last-iterate convergence if we simply pick $\ell_t = F$. The answer turns out to be positive at least for OGD (which is equivalent to SGD now). However, to prove this result, we rely on a technique specialized to stochastic optimization recently developed by [19, 44]. To not diverge from the topic of OCO, we defer the last-iterate convergence of OGD to Appendix D, in which Theorem 8 gives a general result for any stepsize η_t and Corollary 4 shows the last-iterate rate under the same stepsize $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ as in Theorem 1 before.

4.2 Nonsmooth Nonconvex Optimization

This section contains another application, nonsmooth nonconvex optimization with heavy tails. Due to limited space, we will provide only the necessary background. For more details, we refer the reader to [6, 13, 15, 16, 38, 39] for recent progress. We start with a new set of conditions.

Assumption 2. *We consider the following series of assumptions:*

- *The objective F is lower bounded by $F_\star \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) \in \mathbb{R}$.*
- *F is differentiable and well-behaved, i.e., $F(\boldsymbol{x}) - F(\boldsymbol{y}) = \int_0^1 \langle \nabla F(\boldsymbol{y} + t(\boldsymbol{x}-\boldsymbol{y})), \boldsymbol{x}-\boldsymbol{y}\rangle \, \mathrm{d}t$.*
- *F is G-Lipschitz on $\mathbb{R}^d$, i.e., $\|\nabla F(\boldsymbol{x})\| \le G, \forall \boldsymbol{x} \in \mathbb{R}^d$.*
- *Given $\boldsymbol{z}_t \in \mathbb{R}^d$ at the t-th iteration, one can query $\boldsymbol{g}_t \in \mathbb{R}^d$ satisfying $\mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}] = \nabla F(\boldsymbol{z}_t)$ and $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$ for some $\mathsf{p} \in (1,2]$ and $\sigma \ge 0$, where $\mathcal{F}_t$ denotes the natural filtration and $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla F(\boldsymbol{z}_t)$ is the stochastic noise.*

Remark 6. The second point is a mild regularity condition introduced by [5] and becomes standard in the literature [2, 17, 48]. See Definition 1 and Proposition 2 of [5] for more details. In the fourth point, we use the same notation $\boldsymbol{z}_t$ as in the algorithm being studied later. In fact, it can be arbitrary.

In nonsmooth nonconvex optimization, we aim to find a (δ,ϵ)-stationary point [46] (see the formal Definition 2 in Appendix E). This goal can be reduced to finding a point $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\nabla F(\boldsymbol{x})\|_\delta \le \epsilon$, where $\|\nabla F(\boldsymbol{x})\|_\delta$ is a quantity introduced by [5] as follows.

Definition 1 (Definition 5 of [5])**.** Given a point $\boldsymbol{x} \in \mathbb{R}^d$, a number $\delta > 0$ and an almost-everywhere differentiable function F, define $\|\nabla F(\boldsymbol{x})\|_\delta \triangleq \inf_{S\subset\mathcal{B}(\boldsymbol{x},\delta), \frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\boldsymbol{y}=\boldsymbol{x}} \left\| \frac{1}{|S|}\sum_{\boldsymbol{y}\in S} \nabla F(\boldsymbol{y}) \right\|$.

The only existing sample complexity under Assumption 2 is $(G+\sigma)^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ in high probability [17], where we only report the dominant term and hide the dependency on the failure probability.

However, on the theoretical side, their result cannot recover the optimal bound $G^2\delta^{-1}\epsilon^{-3}$ [5] in the deterministic case. On the practical side, their method also employs the gradient clipping step, which introduces a new clipping parameter to tune. In fact, as stated in their Section 5, they observed in experiments that their algorithm without the clipping operation (exactly the algorithm we study next) still works under heavy tails. In addition, in their Section 6, they also explicitly ask whether the requirement to know G and A can be removed.

As will be seen later, we can address these points with the new regret bounds presented before.

4.2.1 Online-to-Nonconvex Conversion under Heavy Tails

Algorithm 4 Online-to-Nonconvex Conversion (O2NC) [5]

Input: initial point $\boldsymbol{y}_0 \in \mathbb{R}^d$, $K \in \mathbb{N}$, $T \in \mathbb{N}$, online learning algorithm A.
for $n = 1$ **to** KT **do**
  Receive $\boldsymbol{x}_n$ from A
  $\boldsymbol{y}_n = \boldsymbol{y}_{n-1} + \boldsymbol{x}_n$
  $\boldsymbol{z}_n = \boldsymbol{y}_{n-1} + s_n\boldsymbol{x}_n$ where $s_n \sim \mathsf{Uniform}\,[0,1]$ i.i.d.
  Query a stochastic gradient $\boldsymbol{g}_n$ at $\boldsymbol{z}_n$
  Send $\boldsymbol{g}_n$ to A
end for

Remark 7. Note that O2NC is a randomized algorithm. Therefore, the definition of the natural filtration is adjusted to $\mathcal{F}_n \triangleq \sigma(s_1, \boldsymbol{g}_1, \ldots, s_n, \boldsymbol{g}_n, s_{n+1})$ accordingly.

We provide the Online-to-Nonconvex Conversion (O2NC) framework in Algorithm 4, which serves as a meta algorithm. Roughly speaking, Algorithm 4 reduces a nonconvex optimization problem to an OCO (in fact, OLO) problem, for which the K-shifting regret (see (9)) of the online learner A crucially affects the final convergence rate. However, the existing Theorem 8 of [5], a general convergence result for the above reduction, cannot directly apply to heavy-tailed noise, since its proof relies on the finite variance condition on $\boldsymbol{g}_n$ (see Appendix E for more details).

Theorem 4. *Under Assumption 2 and let* $\boldsymbol{v}_k \triangleq -D\frac{\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)}{\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|}, \forall k \in [K]$ *for arbitrary* $D > 0$*, then for any online learning algorithm* A *in* O2NC *(Algorithm 4), we have*

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{K}\left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right] \lesssim \frac{F(\boldsymbol{y}_0)-F_\star}{DKT}+\frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_K)\right]}{DKT}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_K)$ in Theorem 4 is called *K-shifting regret* [5], defined as follows:

$$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_K) \triangleq \sum_{k=1}^{K}\sum_{n=(k-1)T+1}^{kT}\ell_n(\boldsymbol{x}_n)-\ell_n(\boldsymbol{v}_k) \quad \text{where} \quad \ell_n(\boldsymbol{x}) \triangleq \langle \boldsymbol{g}_n, \boldsymbol{x}\rangle. \tag{9}$$

Theorem 4 here provides a new and the first theoretical guarantee for O2NC under heavy tails. Especially, it recovers Theorem 8 of [5] when $\mathsf{p}=2$. A remarkable point is that the O2NC algorithm itself does not need any information about p. The proof of Theorem 4 can be found in Appendix E.

4.2.2 Convergence Rates

Theorem 4 enables us to apply the results presented in Section 3. Concretely, for $\mathcal{X}=\mathcal{B}^d(D)$ and any $\mathsf{A}\in\{\mathsf{OGD},\mathsf{DA},\mathsf{AdaGrad}\}$, if we reset the stepsize in A after every T iterations, there will be $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_K)\right] \lesssim GDK\sqrt{T}+\sigma DKT^{1/\mathsf{p}}$ by our new regret bounds, since $\boldsymbol{v}_k\in\mathcal{X}$. With a carefully picked D, we obtain the following Theorem 5. Its proof is deferred to Appendix E.

Theorem 5. *Under Assumption 2 and let* $\Delta \triangleq F(\boldsymbol{y}_0)-F_\star$ *and* $\bar{\boldsymbol{z}}_k \triangleq \frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\boldsymbol{z}_n, \forall k\in[K]$*, setting any* $\mathsf{A}\in\{\mathsf{OGD},\mathsf{DA},\mathsf{AdaGrad}\}$ *in* O2NC *(Algorithm 4) with a domain* $\mathcal{X}=\mathcal{B}^d(D)$ *for* $D=\delta/T$ *and resetting the stepsize in* A *after every* T *iterations, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\bar{\boldsymbol{z}}_k)\right\|_\delta\right] \lesssim \frac{\Delta}{\delta K}+\frac{G}{\sqrt{T}}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Notably, this is the first time confirming that gradient clipping is indeed unnecessary for the O2NC framework, matching the experimental observation of [17].

Corollary 3. *Under the same setting of Theorem 5, suppose we have* $N\geq 2$ *stochastic gradient budgets, taking* $K=\lfloor N/T\rfloor$ *and* $T=\lceil N/2\rceil \wedge \left(\left\lceil(\delta GN/\Delta)^{\frac{2}{3}}\right\rceil \vee \left\lceil(\delta\sigma N/\Delta)^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\right\rceil\right)$*, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\bar{\boldsymbol{z}}_k)\right\|_\delta\right] \lesssim \frac{G}{\sqrt{N}}+\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}}+\frac{\Delta}{\delta N}+\frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}}+\frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

Corollary 3 is obtained by optimizing K and T in Theorem 5. It implies a sample complexity of $G^2\delta^{-1}\epsilon^{-3}+\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ for finding a (δ,ϵ)-stationary point, improved over the previous bound $(G+\sigma)^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ [17]. Furthermore, leveraging the adaptive feature of AdaGrad, Corollary 5 in Appendix E shows how to set K and T without G, σ, and p, resulting in the first provably rate for O2NC when no problem information is known in advance, which solves the problem asked by [17].

5 Conclusion and Limitation

This paper shows that three classical OCO algorithms, OGD, DA, and AdaGrad, can achieve the optimal in-expectation regret under heavy tails without any algorithmic modification if the feasible set is bounded, and provides some applications in stochastic optimization. The main limitation of our work is that all the proof crucially relies on the bounded domain assumption, which may not always be suitable in practice. Finding a weaker sufficient condition, under which the classical OCO algorithms work with heavy tails provably, is a direction worth studying in the future.

References

[1] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 111–119. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/agarwal19c.html`.

[2] Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94909–94933. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/ac8ec9b4d94c03f0af8c4fe3d5fad4fd-Paper-Conference.pdf`.

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339.

[4] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[5] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6643–6670. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/cutkosky23a.html`.

[6] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6692–6703. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/2c8d9636f74d0207ff4f65956010f450-Paper-Conference.pdf`.

[7] Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[10] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL `http://dx.doi.org/10.1561/2400000013`.

[11] Elad Hazan and Karan Singh. Introduction to online control, 2025. URL `https://arxiv.org/abs/2211.09619`.

[12] Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pages 4262–4274. PMLR, 2021.

[13] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4570–4597. PMLR, 12–15 Jul 2023. URL `https://proceedings.mlr.press/v195/jordan23a.html`.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022. URL http://jmlr.org/papers/v23/21-1507.html.

[16] Guy Kornowski and Ohad Shamir. On the complexity of finding small subgradients in nonsmooth optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. URL https://openreview.net/forum?id=SaRQ4oTqWbP.

[17] Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32122–32138. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bo.html.

[18] 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.

[19] Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

[20] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NKotdPUc3L.

[21] Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 525–533, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/mcmahan11b.html.

[22] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

[23] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, pages 244–256. Omnipress, 2010.

[24] Arkadi Nemirovski and David Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

[25] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[26] 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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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.

[27] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[28] Francesco Orabona. Last iterate of sgd converges (even in unbounded domains). 2020. URL https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/.

[29] Daniela Angela Parletta, Andrea Paudice, and Saverio Salzo. An improved analysis of the clipped stochastic subgradient method under heavy-tailed noise, 2025. URL https://arxiv.org/abs/2410.00573.

[30] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–510, 2009. doi: 10.1109/ALLERTON.2009.5394945.

[31] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[32] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

[33] R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[34] 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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sadiev23a.html.

[35] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL http://dx.doi.org/10.1561/2200000018.

[36] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/shamir13.html.

[37] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

[38] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

[39] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21360–21379. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tian22a.html.

[40] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[41] 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 Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 65–102. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/vural22a.html.

[42] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ward19a.html.

[43] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

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

[45] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

[46] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11173–11182. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20p.html.

[47] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8000–8012. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/349956dee974cfdcbbb2d06afad5dd4a-Paper-Conference.pdf.

[48] Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58780–58799. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24k.html.

[49] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[50] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Missing Proofs for Online Gradient Descent

This section provides missing proofs for regret bounds of OGD. Before showing the formal proof, we recall the following core inequality that holds for any $\boldsymbol{x} \in \mathcal{X}$ given in (6):

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{10}$$

The key to establishing the above result is showing

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t} \le \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{11}$$

the proof of which is by combining (3), (4), and (5) established in the main text.

A.1 Proof of Theorem 1

Proof. For any $\boldsymbol{x} \in \mathcal{X}$, sum up (10) from $t = 1$ to T and drop the term $-\frac{\|\boldsymbol{x}_{T+1} - \boldsymbol{x}\|^2}{2\eta_T}$ to obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \\
\le& \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2\eta_1} + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}} && (12)\\
\le& \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}, && (13)
\end{aligned}$$

where the last step is due to $\|\boldsymbol{x}_t - \boldsymbol{x}\| \le D, \forall t \in [T]$ and $\eta_{t+1} \le \eta_t, \forall t \in [T-1]$.

Taking expectations on both sides of (13) yields that

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \le \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1}\sigma^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{14}$$

where for the L.H.S., we use $\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \mid \mathcal{F}_{t-1}\right]\right]$ and

$$\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \mid \mathcal{F}_{t-1}\right] = \langle \mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right], \boldsymbol{x}_t - \boldsymbol{x} \rangle = \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x} \rangle \ge \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}), \tag{15}$$

for the R.H.S., we use $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$.

Finally, we plug $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}, \forall t \in [T]$ into (14), then use $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \lesssim \sqrt{T}$ and $\sum_{t=1}^{T} \frac{1}{t^{1-1/\mathsf{p}}} \lesssim T^{1/\mathsf{p}}$ to conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

A.2 Extension to Online Strongly Convex Optimization

Next, we extend Theorem 1 to the strongly convex case, i.e., $\exists \mu > 0$ such that for all $t \in [T]$,

$$\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \langle \nabla \ell_t(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \ell_t(\boldsymbol{y}) \le \ell_t(\boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \nabla \ell_t(\boldsymbol{y}) \in \partial \ell_t(\boldsymbol{y}). \tag{16}$$

In this setting, it is well known that OGD achieves a logarithmic regret bound when $\mathsf{p} = 2$ [10, 27]. Theorem 6 below provides the first provable result for $\mathsf{p} < 2$.

Theorem 6. *Under Assumption 1 and additionally assuming (16), taking $\eta_t = \frac{1}{\mu t}$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2(1 + \log T)}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} \times \begin{cases} T^{2-\mathsf{p}} & \mathsf{p} \in (1,2) \\ 1 + \log T & \mathsf{p} = 2 \end{cases}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Theorem 6 shows that under strongly convexity, OGD for $\mathsf{p} \in (1,2)$ achieves a better sublinear regret $T^{2-\mathsf{p}}$ than $T^{1/\mathsf{p}}$ in Theorem 1 as $2-\mathsf{p} \le 1/\mathsf{p}, \forall \mathsf{p} > 0$. One point we highlight here is that the stepsize $\eta_t = \frac{1}{\mu t}$ is commonly used in the OCO literature and is independent of the tail index p.

However, in contrast to Theorem 1, we suspect Theorem 6 is not tight in T for $\mathsf{p} \in (1,2)$. The reason is that for nonsmooth strongly convex optimization with heavy tails (i.e., $\ell_t = F, \forall t \in [T]$ where F is strongly convex), Theorem 6 can convert to a convergence rate only in the order of $1/T^{\mathsf{p}-1}$, which is worse than the lower bound $1/T^{2-2/\mathsf{p}}$ [45]. Therefore, we conjecture that a way to obtain a better regret bound than $T^{2-\mathsf{p}}$ exists, which we leave as future work.

Proof of Theorem 6. For any $\boldsymbol{x} \in \mathcal{X}$, we take expectations on both sides of (12) to have

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \le \left(\frac{1}{\eta_1} - \mu\right)\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \mu\right)\frac{\mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2\right]}{2}$$
$$+ \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1}\sigma^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{17}$$

where for the L.H.S., we follow a similar step of reasoning out (15) but instead using

$$\langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \ge \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}\|^2,$$

for the R.H.S., we use $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$.

Next, we plug $\eta_t = \frac{1}{\mu t}, \forall t \in [T]$ into (17) to obtain

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \sum_{t=1}^{T} \frac{G^2}{\mu t} + \frac{\sigma^{\mathsf{p}} D^{2-\mathsf{p}}}{\mu^{\mathsf{p}-1} t^{\mathsf{p}-1}}$$
$$\lesssim \frac{G^2(1+\log T)}{\mu} + \frac{\sigma^{\mathsf{p}} D^{2-\mathsf{p}}}{\mu^{\mathsf{p}-1}} \times \begin{cases} T^{2-\mathsf{p}} & \mathsf{p} \in (1,2) \\ 1 + \log T & \mathsf{p} = 2 \end{cases}.$$

Lastly, it is known that if ℓ_t is G-Lipschitz and μ-strongly convex on a domain $\mathcal{X}$ with a diameter D, then it satisfies $D \lesssim \frac{G}{\mu}$ (e.g., see Lemma 2 of [31]). Therefore, when $\mathsf{p} \in (1,2)$,

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2(1+\log T)}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} T^{2-\mathsf{p}}.$$

□

B Missing Proofs for Dual Averaging

This section provides missing proofs for regret bounds of DA.

B.1 Proof of Theorem 2

Proof. Let $L_t(\boldsymbol{x}) \triangleq \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_{t-1}} + \sum_{s=1}^{t-1}\langle \boldsymbol{g}_s, \boldsymbol{x}\rangle, \forall t \in [T+1]$, where $\eta_0 \triangleq \eta_1$. Then DA can be equivalently written as

$$\boldsymbol{x}_t = \mathrm{argmin}_{\boldsymbol{x} \in \mathcal{X}} L_t(\boldsymbol{x}), \forall t \in [T+1].$$

By Lemma 7.1 of [27], for any $\boldsymbol{x} \in \mathcal{X}$,

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle = \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + L_{T+1}(\boldsymbol{x}_{T+1}) - L_{T+1}(\boldsymbol{x}) + \sum_{t=1}^{T} L_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle - L_{t+1}(\boldsymbol{x}_{t+1})$$
$$\le \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} L_t(\boldsymbol{x}_t) - L_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle,$$

where the inequality holds by $L_{T+1}(\boldsymbol{x}_{T+1}) \leq L_{T+1}(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}$ due to $\boldsymbol{x}_{T+1} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} L_{T+1}(\boldsymbol{x})$. Note that for any $t \in [T]$,

$$\begin{aligned}
&L_t(\boldsymbol{x}_t) - L_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \\
=&L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle + \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_1\|^2}{2\eta_{t-1}} - \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_1\|^2}{2\eta_t} \\
\overset{(a)}{\leq}& L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \\
\overset{(b)}{\leq}& \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}},
\end{aligned}$$

where (a) is by $\eta_t \leq \eta_{t-1}, \forall t \in [T]$ and (b) is holds because L_t is $\frac{1}{\eta_{t-1}}$-strongly convex and $\boldsymbol{x}_t = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} L_t(\boldsymbol{x})$, which together imply

$$L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) \leq \langle \nabla L_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \leq -\frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}.$$

Therefore, we have

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}. \tag{18}$$

By the same argument as proving (11) but replacing η_t with η_{t-1}, there is

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \leq \eta_{t-1} G^2 + \mathsf{C}(\mathsf{p}) \eta_{t-1}^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}.$$

As such, we know

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \eta_{t-1} G^2 + \mathsf{C}(\mathsf{p}) \eta_{t-1}^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}.$$

Finally, following similar steps in proving Theorem 1 in Appendix A, we conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

B.2 Dual Averaging with an Adaptive Stepsize

We show that DA with an adaptive stepsize can also achieve the optimal regret $GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}$.

Theorem 7. *Under Assumption 1, taking $\eta_t = 2DV_t^{-1/2}$ and $V_t = \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2$ in* DA *(Algorithm 2), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. For any $\boldsymbol{x} \in \mathcal{X}$, we have

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \overset{(18)}{\leq} \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}, \tag{19}$$

where $\eta_0 \triangleq \eta_1$. On the one hand, we can use AM-GM inequality to bound

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \leq \frac{\eta_{t-1} \|\boldsymbol{g}_t\|^2}{2}.$$

On the other hand, we know

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \leq \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \leq \|\boldsymbol{g}_t\| D, \tag{20}$$

where the second step is by Cauchy-Schwarz inequality. Therefore, for any $t \geq 2$,

$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} &\leq \frac{\eta_{t-1}\|\boldsymbol{g}_t\|^2}{2} \wedge \|\boldsymbol{g}_t\| D \overset{(a)}{\leq} \frac{2}{\frac{2}{\eta_{t-1}\|\boldsymbol{g}_t\|^2} + \frac{1}{\|\boldsymbol{g}_t\|D}} \\ &\overset{(b)}{=} \frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2} + \|\boldsymbol{g}_t\|} \overset{(c)}{\leq} \frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}}, \end{aligned} \tag{21}$$

where (a) is due to $x \wedge y \leq \frac{2}{x^{-1}+y^{-1}}, \forall x, y > 0$, (b) is by $\eta_{t-1} = \frac{2D}{\sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2}}$, and (c) holds because of $\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2} \leq \sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2} + \|\boldsymbol{g}_t\|$. Note that (21) is also true for $t = 1$ by (20).

Combine (19) and (21) and use $\|\boldsymbol{x} - \boldsymbol{x}_1\| \leq D$ to obtain

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}} = \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\eta_t\|\boldsymbol{g}_t\|^2,$$

which only differs from (22) by a constant. Hence, by a similar proof for (24), there is

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \lesssim D\left[\sqrt{\sum_{t=1}^{T}\|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right],$$

implying

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

C Missing Proofs for AdaGrad

This section provides missing proofs for regret bounds of AdaGrad.

C.1 Proof of Theorem 3

Proof. As mentioned, AdaGrad can be viewed as OGD with a stepsize $\eta_t = \frac{\eta}{\sqrt{V_t}} = \frac{\eta}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}}$. Therefore, we can use (1) for AdaGrad to know for any $\boldsymbol{x} \in \mathcal{X}$,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \frac{\eta_t\|\boldsymbol{g}_t\|^2}{2}.$$

Sum up the above inequality from $t = 1$ to T and drop the term $-\frac{\|\boldsymbol{x}_{T+1} - \boldsymbol{x}\|^2}{2\eta_T}$ to have

$$\begin{aligned}\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle &\leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2\eta_1} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T}\frac{\eta_t\|\boldsymbol{g}_t\|^2}{2} \\ &\leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t\|\boldsymbol{g}_t\|^2}{2}, \end{aligned} \tag{22}$$

where the last step is by $\|\boldsymbol{x}_t - \boldsymbol{x}\| \leq D, \forall t \in [T]$ and $\eta_{t+1} \leq \eta_t, \forall t \in [T-1]$.

Next, observe that for any $t \in [T]$,

$$\|\boldsymbol{g}_t\|^2 = \frac{\eta^2}{\eta_t^2} - \frac{\eta^2}{\eta_{t-1}^2} = \eta^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\left(\frac{1}{\eta_t} + \frac{1}{\eta_{t-1}}\right) \leq \frac{2\eta^2}{\eta_t}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right),$$

where $1/\eta_0$ should be read as 0. The above inequality implies

$$\sum_{t=1}^{T}\frac{\eta_t\|\boldsymbol{g}_t\|^2}{2} \leq \eta^2\sum_{t=1}^{T}\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{\eta^2}{\eta_T}. \tag{23}$$

Combine (22) and (23) to have

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t-\boldsymbol{x}\rangle \leq \frac{D^2}{2\eta_T}+\frac{\eta^2}{\eta_T}=\left(\frac{D^2}{2\eta}+\eta\right)\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{g}_t\right\|^2}.$$

Note that there is

$$\begin{aligned}\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{g}_t\right\|^2} &\leq \sqrt{\sum_{t=1}^{T}2\left\|\nabla\ell_t(\boldsymbol{x}_t)\right\|^2+2\left\|\boldsymbol{\epsilon}_t\right\|^2} \leq \sqrt{2\sum_{t=1}^{T}\left\|\nabla\ell_t(\boldsymbol{x}_t)\right\|^2}+\sqrt{2\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^2}\\ &\leq \sqrt{2\sum_{t=1}^{T}\left\|\nabla\ell_t(\boldsymbol{x}_t)\right\|^2}+\sqrt{2}\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},\end{aligned}$$

where the last step is due to $\left\|\cdot\right\|_2 \leq \left\|\cdot\right\|_{\mathsf{p}}$ for any $\mathsf{p}\in[1,2]$. Hence, we obtain

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t-\boldsymbol{x}\rangle \leq \sqrt{2}\left(\frac{D^2}{2\eta}+\eta\right)\left[\sqrt{\sum_{t=1}^{T}\left\|\nabla\ell_t(\boldsymbol{x}_t)\right\|^2}+\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right]. \tag{24}$$

We take expectations on both sides of (24), then apply Hölder's inequality to have

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \leq \left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \leq \sigma T^{\frac{1}{\mathsf{p}}},$$

and finally plug in $\eta = D/\sqrt{2}$ to conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{AdaGrad}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T}+\sigma D T^{1/\mathsf{p}}.$$

□

D Missing Proofs for Applications: Nonsmooth Convex Optimization

We prove the following last-iterate convergence result for SGD (i.e., OGD for stochastic optimization) under heavy-tailed noise. The proof of Theorem 8 is inspired by [19, 44].

Theorem 8. *Under Assumption 1 for $\ell_t(\boldsymbol{x})=F(\boldsymbol{x})$, for any stepsize $\eta_t>0$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[F(\boldsymbol{x}_T)-F(\boldsymbol{x})\right] \lesssim \frac{D^2}{\sum_{t=1}^{T}\eta_t}+G^2\sum_{t=1}^{T}\frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}+\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\sum_{t=1}^{T}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}.$$

Proof. Given $\boldsymbol{x}\in\mathcal{X}$, we recursively define

$$\boldsymbol{y}_0 \triangleq \boldsymbol{x} \quad \text{and} \quad \boldsymbol{y}_t \triangleq \left(1-\frac{w_{t-1}}{w_t}\right)\boldsymbol{x}_t+\frac{w_{t-1}}{w_t}\boldsymbol{y}_{t-1}, \forall t\in[T], \tag{25}$$

in which

$$w_t \triangleq \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}, \forall t\in\{0\}\cup[T-1] \quad \text{and} \quad w_T \triangleq w_{T-1}=1. \tag{26}$$

Equivalently, $\boldsymbol{y}_t$ can be written into a convex combination of $\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ as

$$\boldsymbol{y}_t = \frac{w_0}{w_t}\boldsymbol{x}+\sum_{s=1}^{t}\frac{w_s-w_{s-1}}{w_t}\boldsymbol{x}_s, \forall t\,\{0\}\cup[T]. \tag{27}$$

Therefore, $\boldsymbol{y}_t$ also falls into $\mathcal{X}$ and satisfies $\boldsymbol{y}_t\in\mathcal{F}_{t-1}$.

We invoke (10) for $\boldsymbol{y}_t$ to obtain

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{y}_t \rangle \leq \frac{\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{28}$$

Since $\boldsymbol{x}_t, \boldsymbol{y}_t \in \mathcal{F}_{t-1}$, there is

$$\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right], \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] = \mathbb{E}\left[\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] \geq \mathbb{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right],$$

where the last step is due to the convexity of F. As such, we can take expectations on both sides of (28) to have

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right] &\leq \frac{\mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2\right]}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1}\sigma^{\mathsf{p}} D^{2-\mathsf{p}} \\
&\leq \frac{\mathbb{E}\left[\frac{w_{t-1}}{w_t}\left\|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\right\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2\right]}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1}\sigma^{\mathsf{p}} D^{2-\mathsf{p}},
\end{aligned} \tag{29}$$

where the second step is due to $\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2 \leq \left(1 - \frac{w_{t-1}}{w_t}\right)\|\boldsymbol{x}_t - \boldsymbol{x}_t\|^2 + \frac{w_{t-1}}{w_t}\left\|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\right\|^2 = \frac{w_{t-1}}{w_t}\left\|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\right\|^2$ by (25) and the convexity of $\|\boldsymbol{x}_t - \cdot\|^2$. Mutiply both sides of (29) by $w_t\eta_t$ and sum up from $t = 1$ to T to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right)\right] \\
\leq &\frac{w_0\|\boldsymbol{x}_1 - \boldsymbol{y}_0\|^2 - \mathbb{E}\left[w_T\|\boldsymbol{x}_{T+1} - \boldsymbol{y}_T\|^2\right]}{2} + \sum_{t=1}^{T} w_t\eta_t^2 G^2 + \mathsf{C}(\mathsf{p})w_t\eta_t^{\mathsf{p}}\sigma^{\mathsf{p}} D^{2-\mathsf{p}} \\
\leq &\frac{w_0 D^2}{2} + \sum_{t=1}^{T} w_t\eta_t^2 G^2 + \mathsf{C}(\mathsf{p})w_t\eta_t^{\mathsf{p}}\sigma^{\mathsf{p}} D^{2-\mathsf{p}}.
\end{aligned} \tag{30}$$

Now observe that

$$\begin{aligned}
F(\boldsymbol{y}_t) - F(\boldsymbol{x}) &\overset{(27)}{\leq} \frac{w_0}{w_t}\left(F(\boldsymbol{x}) - F(\boldsymbol{x})\right) + \sum_{s=1}^{t} \frac{w_s - w_{s-1}}{w_t}\left(F(\boldsymbol{x}_s) - F(\boldsymbol{x})\right) \\
&= \sum_{s=1}^{t} \frac{w_s - w_{s-1}}{w_t}\left(F(\boldsymbol{x}_s) - F(\boldsymbol{x})\right),
\end{aligned}$$

which implies

$$\begin{aligned}
\sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{y}_t) - F(\boldsymbol{x})\right) &\leq \sum_{t=1}^{T}\sum_{s=1}^{t}\left(w_s - w_{s-1}\right)\eta_t\left(F(\boldsymbol{x}_s) - F(\boldsymbol{x})\right) \\
&= \sum_{t=1}^{T}\left(w_t - w_{t-1}\right)\left(\sum_{s=t}^{T}\eta_s\right)\left(F(\boldsymbol{x}_t) - F(\boldsymbol{x})\right).
\end{aligned}$$

Thus, we can lower bound the L.H.S. of (30) by

$$\begin{aligned}
\sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right) &= \sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{x}_t) - F(\boldsymbol{x})\right) - w_t\eta_t\left(F(\boldsymbol{y}_t) - F(\boldsymbol{x})\right) \\
&\geq \sum_{t=1}^{T}\left[w_t\eta_t - \left(w_t - w_{t-1}\right)\left(\sum_{s=t}^{T}\eta_s\right)\right]\left(F(\boldsymbol{x}_t) - F(\boldsymbol{x})\right) \\
&= w_T\eta_T\left(F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right),
\end{aligned} \tag{31}$$

where the last step is due to, for $t \in [T-1]$,

$$w_t\eta_t - (w_t - w_{t-1})\left(\sum_{s=t}^{T}\eta_s\right) \overset{(26)}{=} \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t - \left(\frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s} - \frac{\eta_T}{\sum_{s=t}^{T}\eta_s}\right)\left(\sum_{s=t}^{T}\eta_s\right)$$
$$= \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t - \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t = 0,$$

and $w_T \overset{(26)}{=} w_{T-1} = 1$.

We plug (31) back into (30) and divide both sides by $w_T\eta_T$ to obtain

$$\mathbb{E}\left[F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right] \le \frac{w_0 D^2}{2w_T\eta_T} + \sum_{t=1}^{T}\frac{w_t\eta_t^2}{w_T\eta_T}G^2 + \mathsf{C}(\mathsf{p})\frac{w_t\eta_t^{\mathsf{p}}}{w_T\eta_T}\sigma^{\mathsf{p}}D^{2-\mathsf{p}}$$
$$\overset{(26)}{\lesssim} \frac{D^2}{\sum_{t=1}^{T}\eta_t} + G^2\sum_{t=1}^{T}\frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T}\eta_s} + \sigma^{\mathsf{p}}D^{2-\mathsf{p}}\sum_{t=1}^{T}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}.$$

□

Equipped with Theorem 8, we show the following anytime last-iterate convergence rate for SGD/OGD. As far as we know, this is the first and the only provable result demonstrating that the last iterate of SGD can converge in heavy-tailed stochastic optimization without gradient clipping. Compared to Corollary 2, the difference is up to an extra logarithmic factor. Therefore, it is nearly optimal.

Corollary 4. *Under Assumption 1 for $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x})$, taking $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right] \lesssim \frac{GD\left(1+\log T\right)}{\sqrt{T}} + \frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Proof. By Theorem 8, we have

$$\begin{aligned}&\mathbb{E}\left[F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right]\\ \lesssim&\frac{D^2}{\sum_{t=1}^{T}\eta_t} + G^2\sum_{t=1}^{T}\frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T}\eta_s} + \sigma^{\mathsf{p}}D^{2-\mathsf{p}}\sum_{t=1}^{T}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}\\ =&\frac{D^2}{\sum_{t=1}^{T}\eta_t} + G^2\left(\eta_T + \sum_{t=1}^{T-1}\frac{\eta_t^2}{\sum_{s=t+1}^{T}\eta_s}\right) + \sigma^{\mathsf{p}}D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1} + \sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=t+1}^{T}\eta_s}\right).\end{aligned}$$

For any $t \in \{0\} \cup [T-1]$, observe that by Cauchy-Schwarz inequality

$$(T-t)^2 \le \left(\sum_{s=t+1}^{T}\frac{1}{\eta_s}\right)\left(\sum_{s=t+1}^{T}\eta_s\right) \Rightarrow \frac{1}{\sum_{s=t+1}^{T}\eta_s} \le \frac{\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}.$$

Thus, there is

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right] \lesssim&\frac{D^2}{T^2}\sum_{t=1}^{T}\frac{1}{\eta_t} + G^2\left(\eta_T + \sum_{t=1}^{T-1}\frac{\eta_t^2\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right)\\ &+ \sigma^{\mathsf{p}}D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1} + \sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right).\end{aligned} \tag{32}$$

We first bound

$$\sum_{t=1}^{T}\frac{1}{\eta_t} = \sum_{t=1}^{T}\frac{G\sqrt{t}}{D} \vee \frac{\sigma t^{1/\mathsf{p}}}{D} \le \sum_{t=1}^{T}\frac{G\sqrt{t}}{D} + \frac{\sigma t^{1/\mathsf{p}}}{D} \lesssim \frac{G}{D}T^{3/2} + \frac{\sigma}{D}T^{1+1/\mathsf{p}},$$

which implies

$$\frac{D^2}{T^2}\sum_{t=1}^{T}\frac{1}{\eta_t}\lesssim\frac{GD}{\sqrt{T}}+\frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{33}$$

Next, we know

$$\begin{aligned}\eta_T+\sum_{t=1}^{T-1}\frac{\eta_t^2\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2} &\overset{(a)}{\leq}\frac{D}{G\sqrt{T}}+\sum_{t=1}^{T-1}\left[\frac{D}{G}\cdot\frac{\sum_{s=t+1}^{T}\sqrt{s}}{t(T-t)^2}+\frac{\sigma D}{G^2}\cdot\frac{\sum_{s=t+1}^{T}s^{1/\mathsf{p}}}{t(T-t)^2}\right]\\ &\overset{\text{Fact 1}}{\lesssim}\frac{D}{G\sqrt{T}}+\frac{D\left(1+\log T\right)}{G\sqrt{T}}+\frac{\sigma D\left(1+\log T\right)}{G^2T^{1-\frac{1}{\mathsf{p}}}},\end{aligned}$$

where (a) is by $\eta_t\leq\frac{D}{G\sqrt{t}}$ and $\frac{1}{\eta_s}\leq\frac{G\sqrt{s}}{D}\vee\frac{\sigma s^{1/\mathsf{p}}}{D}$. Hence, there is

$$G^2\left(\eta_T+\sum_{t=1}^{T-1}\frac{\eta_t^2\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right)\lesssim\frac{GD\left(1+\log T\right)}{\sqrt{T}}+\frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{34}$$

Similarly, we can bound

$$\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1}+\sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right)\lesssim\frac{GD\left(1+\log T\right)}{\sqrt{T}}+\frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{35}$$

Finally, we plug (33), (34) and (35) back into (32) to conclude. □

E Missing Proofs for Applications: Nonsmooth Nonconvex Optimization

E.1 (δ,ϵ)-Stationary Points

Definition 2 (Definition 4 of [5]). A point $\boldsymbol{x}\in\mathbb{R}^d$ is a (δ,ϵ)-stationary point of an almost-everywhere differentiable function F if there is a finite subset $S\subset\mathcal{B}^d(\boldsymbol{x},\delta)$ such that for $\boldsymbol{y}$ selected uniformly at random from S, $\mathbb{E}\left[\boldsymbol{y}\right]=\boldsymbol{x}$ and $\left\|\mathbb{E}\left[\nabla F(\boldsymbol{y})\right]\right\|\leq\epsilon$.

The concept of the (δ,ϵ)-stationary point presented here is due to [5], which is mildly more stringent than the notion of [46], since the latter does not require $\mathbb{E}\left[\boldsymbol{y}\right]=\boldsymbol{x}$. For more discussions, see Section 2.1 of [5].

E.2 Proof of Theorem 4

In this section, our ultimate goal is to prove Theorem 4 for the O2NC algorithm, extending Theorem 8 of [5] from $\mathsf{p}=2$ to any $\mathsf{p}\in(1,2]$. Notably, our new result does not require any modification to the O2NC method, but is obtained only from a more careful analysis, indicating that O2NC is a robust and powerful algorithmic framework.

We begin with Lemma 2, which lies as the cornerstone for establishing the convergence of O2NC.

Lemma 2 (Theorem 7 of [5]). *Under Assumption 2 (only need the second point and the unbiased part in the fourth point), for any sequence of vectors $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{KT}\in\mathbb{R}^d$,* O2NC *(Algorithm 4) guarantees*

$$\mathbb{E}\left[F(\boldsymbol{y}_{KT})\right]=F(\boldsymbol{y}_0)+\mathbb{E}\left[\sum_{n=1}^{KT}\langle\boldsymbol{g}_n,\boldsymbol{x}_n-\boldsymbol{u}_n\rangle\right]+\mathbb{E}\left[\sum_{n=1}^{KT}\langle\boldsymbol{g}_n,\boldsymbol{u}_n\rangle\right]. \tag{36}$$

To relate Lemma 2 to the concept of K-shifting regret introduced before (see (9)), suppose now a sequence of vectors $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_K$ is given, if we set $\boldsymbol{u}_n=\boldsymbol{v}_k$ for all $n\in\{(k-1)T+1,\ldots,kT\}$ and $k\in[K]$, then the second term on the R.H.S. of (36) can be written as $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_K)\right]$, and the third term can be simplified into $\sum_{k=1}^{K}\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]$.

Same as [5], we pick $\boldsymbol{v}_k \triangleq -D\frac{\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)}{\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|}$ for some constant $D>0$, which gives us

$$\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]=\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n,\boldsymbol{v}_k\right\rangle\right]-D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right]$$
$$\leq D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]-D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right].$$

If $\boldsymbol{\epsilon}_n$ has a finite variance (i.e., $\mathsf{p}=2$), then like [5], one can invoke Hölder's inequality and use the fact $\mathbb{E}\left[\langle\boldsymbol{\epsilon}_m,\boldsymbol{\epsilon}_n\rangle\right]=0,\forall m\neq n\in[KT]$ to obtain for any $k\in[K]$,

$$\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]\leq\sqrt{\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|^2\right]}=\sqrt{\sum_{n=(k-1)T+1}^{kT}\mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^2\right]}\leq\sigma\sqrt{T}.$$

However, this argument immediately fails when $\mathsf{p}<2$ as $\mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^2\right]$ can be $+\infty$. To handle this potential issue, we require the following Lemma 3.

Lemma 3 (Lemma 4.3 of [20])**.** *Given a vector-valued martingale difference sequence $\boldsymbol{w}_1,\dots,\boldsymbol{w}_T$, there is*

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{w}_t\right\|\right]\leq 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right],\forall\mathsf{p}\in[1,2].$$

Equipped with Lemmas 2 and 3, we are ready to formally prove Theorem 4, demonstrating that the $\mathsf{O2NC}$ framework provably works under heavy-tailed noise.

Proof of Theorem 4. We invoke Lemma 2 with $\boldsymbol{u}_n=\boldsymbol{v}_{\lceil n/T\rceil},\forall n\in[KT]$ (equivalently, $\boldsymbol{u}_n=\boldsymbol{v}_k$ if $n\in\{(k-1)T+1,\dots,kT\}$) and use the definition of K-shifting regret (see (9)) to obtain

$$\mathbb{E}\left[F(\boldsymbol{y}_{KT})\right]=F(\boldsymbol{y}_0)+\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_K)\right]+\sum_{k=1}^{K}\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right].\tag{37}$$

Recall that $\boldsymbol{g}_n=\nabla F(\boldsymbol{z}_n)+\boldsymbol{\epsilon}_n$, which implies for any $k\in[K]$,

$$\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]=\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n,\boldsymbol{v}_k\right\rangle\right]+\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n),\boldsymbol{v}_k\right\rangle\right]$$
$$\leq\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\|\boldsymbol{v}_k\|\right]+\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n),\boldsymbol{v}_k\right\rangle\right]$$
$$=D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]-D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right],\tag{38}$$

where the second step is by Cauchy-Schwarz inequality and the last equation holds due to

$$\boldsymbol{v}_k=-D\frac{\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)}{\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|},\forall k\in[K].\tag{39}$$

Combine (37) and (38), apply $F(\boldsymbol{y}_{KT})\geq F_\star$, and rearrange terms to have

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{K}\left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right]$$
$$\leq\frac{F(\boldsymbol{y}_0)-F_\star}{DKT}+\frac{\mathbb{E}\left[R_T^{\mathsf{A}}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_K)\right]}{DKT}+\frac{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]}{KT}.\tag{40}$$

For any fixed $k \in [K]$, we apply Lemma 3 with $\boldsymbol{w}_t = \boldsymbol{\epsilon}_{(k-1)T+t}, \forall t \in [T]$ to know

$$\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n\right\|\right] \le 2\sqrt{2}\mathbb{E}\left[\left(\sum_{n=(k-1)T+1}^{kT} \|\boldsymbol{\epsilon}_n\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right]$$
$$\le 2\sqrt{2}\left(\sum_{n=(k-1)T+1}^{kT} \mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \le 2\sqrt{2}\sigma T^{\frac{1}{\mathsf{p}}}, \tag{41}$$

where the second step is by Hölder's inequality (note that $\mathsf{p} > 1$). Finally, we conclude the proof after plugging (41) back into (40). □

E.3 Proof of Theorem 5

Proof. By Theorem 4, there is

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{K}\left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)\right\|\right] \lesssim \frac{F(\boldsymbol{y}_0) - F_\star}{DKT} + \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K)\right]}{DKT} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{42}$$

Note that A has the domain $\mathcal{X} = \mathcal{B}^d(D)$ and $s_n \sim \mathsf{Uniform}\,[0,1]$. Thus, for any $n \in [KT]$,

$$\|\boldsymbol{x}_n\| \le D \quad \text{and} \quad s_n \in [0,1]. \tag{43}$$

We first lower bound the L.H.S. of (42). Given $k \in [K]$, for any $m < n \in \{(k-1)T+1, \ldots, kT\}$, observe that

$$\|\boldsymbol{z}_n - \boldsymbol{z}_m\| = \|\boldsymbol{y}_{n-1} + s_n\boldsymbol{x}_n - \boldsymbol{y}_{m-1} - s_m\boldsymbol{x}_m\| = \left\|s_n\boldsymbol{x}_n - s_m\boldsymbol{x}_m + \sum_{i=m}^{n-1}\boldsymbol{x}_i\right\|$$
$$\le s_n\|\boldsymbol{x}_n\| + (1-s_m)\|\boldsymbol{x}_m\| + \sum_{i=m+1}^{n-1}\|\boldsymbol{x}_i\| \overset{(43)}{\le} (n-m+1)\,D \le DT.$$

Recall that $\bar{\boldsymbol{z}}_k = \frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\boldsymbol{z}_n$ and $D = \delta/T$ now, then the above inequality implies

$$\|\boldsymbol{z}_n - \bar{\boldsymbol{z}}_k\| \le DT = \delta, \forall n \in \{(k-1)T+1, \ldots kT\}, \tag{44}$$

which means

$$\boldsymbol{z}_n \in \mathcal{B}^d(\bar{\boldsymbol{z}}_k, \delta), \forall n \in \{(k-1)T+1, \ldots kT\}.$$

By the definition of $\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta$ (see Definition 1), there is

$$\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta \le \left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|. \tag{45}$$

Next, we upper bound the R.H.S. of (42). By the definition of K-shifting regret (see (9)), there is

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K)\right] = \sum_{k=1}^{K}\mathbb{E}\left[\sum_{n=(k-1)T+1}^{kT}\langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k\rangle\right].$$

Note that we reset the stepsize in A after every T iterations and $\boldsymbol{v}_k \in \mathcal{B}^d(D)$ by its definition (see (39)). Then for any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$, we can invoke its regret bound[2] (i.e., Theorems 1, 2 and 3) to obtain

$$\mathbb{E}\left[\sum_{n=(k-1)T+1}^{kT}\langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k\rangle\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall k \in [K],$$

[2] A minor point here is that the current function $\ell_n(\boldsymbol{x}) = \langle \boldsymbol{g}_n, \boldsymbol{x}\rangle$ does not entirely fit Assumption 1. We clarify that one does not need to worry about it, since all results proved in Section 3 hold under this change. For example, in the proof of Theorem 1, we can safely replace the L.H.S. of (14) with $\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle\right]$.

which implies

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_K)\right]\lesssim GDK\sqrt{T}+\sigma DKT^{1/\mathsf{p}}. \tag{46}$$

Finally, we plug (45) and (46) back into (42), then use $D=\delta/T$ and $\Delta=F(\boldsymbol{y}_0)-F_\star$ to have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{\Delta}{\delta K}+\frac{G}{\sqrt{T}}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

□

E.4 Proof of Corollary 3

Proof. Recall that we pick

$$K=\left\lfloor\frac{N}{T}\right\rfloor\quad\text{and}\quad T=\left\lceil\frac{N}{2}\right\rceil\wedge\left(\left\lceil\left(\frac{\delta GN}{\Delta}\right)^{\frac{2}{3}}\right\rceil\vee\left\lceil\left(\frac{\delta\sigma N}{\Delta}\right)^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\right\rceil\right),$$

where $\Delta=F(\boldsymbol{y}_0)-F_\star$. We invoke Theorem 5 and use $KT\geq N/4$ (see Fact 2) to obtain

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{\Delta T}{\delta N}+\frac{G}{\sqrt{T}}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

By the definition of T, we know

$$\frac{\Delta T}{\delta N}\lesssim\frac{\Delta}{\delta N}\left[1+\left(\frac{\delta GN}{\Delta}\right)^{\frac{2}{3}}+\left(\frac{\delta\sigma N}{\Delta}\right)^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\right]=\frac{\Delta}{\delta N}+\frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}}+\frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}},$$

and

$$\frac{G}{\sqrt{T}}\lesssim\frac{G}{\sqrt{N}}+\frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}},\qquad\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}\lesssim\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}}+\frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

Therefore, there is

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{G}{\sqrt{N}}+\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}}+\frac{\Delta}{\delta N}+\frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}}+\frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

□

E.5 Extension to the Case of Unknown Problem-Dependent Parameters

In Corollary 5, we show how to set K and T when all problem-dependent parameters are unknown. It is particularly meaningful for AdaGrad. As in that case, the rate is achieved without knowing any problem-dependent parameter. This kind of result is the first to appear for nonsmooth nonconvex optimization with heavy tails. However, the rate is not as good as Corollary 3. It is currently unclear whether the same bound $1/(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}$ as in Corollary 3 can be obtained when no information about the problem is known.

Corollary 5. *Under the same setting of Theorem 5, suppose we have $N\geq 2$ stochastic gradient budgets, taking $K=\lfloor N/T\rfloor$ and $T=\lceil N/2\rceil\wedge\left\lceil(\delta N)^{\frac{2}{3}}\right\rceil$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}}+\frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}}+\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

Proof. We invoke Theorem 5 and use $KT\geq N/4$ (see Fact 2) to obtain

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{\Delta T}{\delta N}+\frac{G}{\sqrt{T}}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

By the definition of T, we know

$$\frac{\Delta T}{\delta N}\lesssim\frac{\Delta}{\delta N}\left[1+(\delta N)^{\frac{2}{3}}\right]\lesssim\frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}}.$$

and

$$\frac{G}{\sqrt{T}}\lesssim\frac{G}{\sqrt{N}}+\frac{G}{(\delta N)^{\frac{1}{3}}}\lesssim\frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}},$$
$$\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}\lesssim\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}}+\frac{\sigma}{(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}\lesssim\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

Therefore, there is

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right]\lesssim\frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}}+\frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}}+\frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

□

F Algebraic Facts

We give two useful algebraic facts in this section.

Fact 1. *For any $T\in\mathbb{N}$ and $a\in(0,1)$, there is*

$$\sum_{t=1}^{T-1}\frac{\sum_{s=t+1}^{T}s^a}{t(T-t)^2}\lesssim\frac{1+\log T}{T^{1-a}}.$$

Proof. Note that $\sum_{s=t+1}^{T}s^a\le(T-t)T^a$, which implies

$$\sum_{t=1}^{T-1}\frac{\sum_{s=t+1}^{T}s^a}{t(T-t)^2}\le\sum_{t=1}^{T-1}\frac{T^a}{t(T-t)}=\frac{1}{T^{1-a}}\sum_{t=1}^{T-1}\frac{1}{t}+\frac{1}{T-t}=\frac{2}{T^{1-a}}\sum_{t=1}^{T-1}\frac{1}{t}\lesssim\frac{1+\log T}{T^{1-a}}.$$

□

Fact 2. *Given $2\le N\in\mathbb{N}$, $K=\lfloor N/T\rfloor$ and $T\in\mathbb{N}$ satisfying $T\le\lceil N/2\rceil$, there is $KT\ge N/4$.*

Proof. Note that $KT=\lfloor N/T\rfloor T\ge N-T\ge(N-1)/2\ge N/4$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed because this paper is purely theoretical.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.