General Tail Bounds for Non-Smooth Stochastic Mirror Descent

# General Tail Bounds for Non-Smooth Stochastic Mirror Descent

**Khaled Eldowa** **Andrea Paudice**
Università degli Studi di Milano, Milan, Italy

## Abstract

In this paper, we provide novel tail bounds on the optimization error of Stochastic Mirror Descent for convex and Lipschitz objectives. Our analysis extends the existing tail bounds from the classical light-tailed Sub-Gaussian noise case to heavier-tailed noise regimes. We study the optimization error of the last iterate as well as the average of the iterates. We instantiate our results in two important cases: a class of noise with exponential tails and one with polynomial tails. A remarkable feature of our results is that they do not require an upper bound on the diameter of the domain. Finally, we support our theory with illustrative experiments that compare the behavior of the average of the iterates with that of the last iterate in heavy-tailed noise regimes.

## 1 INTRODUCTION

*Stochastic Mirror Descent* (SMD) and its more popular Euclidean counterpart *Stochastic (sub-)Gradient Descent* (SGD) are at the core of modern machine learning. For example, they are widely used for performing large-scale optimization tasks, as in the case of empirical (or regularized) risk minimization, and for minimizing the statistical risk in kernel methods. In this paper, we study the performance of SMD in the general problem of minimizing a (non-smooth) convex and Lipschitz function given only noisy oracle access to its (sub-)gradients. SGD was first introduced by Ermol'ev (1969), who studied the convergence of the iterates for convex Lipschitz objectives. Subsequent studies focused on deriving in-expectation bounds on the optimization error of the average of the iterates. Denoting with $T$ the number of iterations, these bounds are of the order of $1/\sqrt{T}$. In their seminal work, (Nemirovski et al., 2009) introduced SMD as a non Euclidean generalization of SGD and showed that it enjoys the same $1/\sqrt{T}$ bound. The shortcoming of in-expectation bounds is that they do not offer guarantees on individual runs of the algorithm. This is especially limiting when multiple runs of the algorithm are not possible, as in large scale problems, or when the data arrives in a stream. Tail bounds offer stronger guarantees that apply to individual runs of the algorithms. For a fixed confidence level $\delta \in (0, 1)$, a straightforward application of Markov's inequality, gives a bound of the order $1/(\delta\sqrt{T})$ that holds with probability at least $1 - \delta$. This bound is much worse than its in-expectation counterpart, even for moderately small $\delta$. Tighter tail bounds with only an overhead of order $\sqrt{\log(1/\delta)}$ have been obtained under a sub-Gaussian assumption on the noise (Liu et al., 2023).

Recent works (Zhang et al., 2020) show that in some settings, the sub-Gaussian assumption is not appropriate, and the noise is better modelled by heavier tailed distributions. Most works studying tail bounds for SMD (SGD) under heavy-tailed noise consider the extreme cases where the noise is only assumed to have finite variance or lower order moments (e.g., Gorbunov et al. (2020); Nguyen et al. (2023)). Under these assumptions, it is necessary to employ some form of truncation of the (sub-)gradients to obtain a poly-logarithmic dependence on $1/\delta$. Instead, we consider less-studied intermediate regimes for the noise, including two classes of sub-Weibull and polynomially tailed distributions. The former is a class of random variables with exponentially decaying tails (including sub-Gaussian and sub-exponential distributions), which has been shown to be relevant in practical applications (Vladimirova et al., 2020), and has been studied in various machine learning and optimization problems (Madden et al., 2021; Kim et al., 2022; Li and Liu, 2022; Li and Jordan, 2023; Wood and Dall'Anese, 2023). The latter class, which includes some Pareto and power law distributions, has also recently captured interest in the machine learning community (Bakhshizadeh et al., 2023; Lou et al., 2022). Moreover, we study the performance of SMD in its

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

plain form (i.e., without truncation), which, in practice, is the more widely used approach. Also, truncation introduces at least one additional parameter, the truncation level, further complicating the tuning of the algorithm in practice. Ideally then, one would like to avoid truncation unless the noise is extremely heavy-tailed.

On a different thread, we notice that most results in the non-smooth case concern the average of the iterates generated by SMD during its execution. However, in practice, taking as solution just the last iterate is by far the preferred heuristic. Consequently, more recent works (Shamir and Zhang, 2013; Harvey et al., 2019; Jain et al., 2021) have focused on developing an understanding of the theoretical performance of this approach. In this case, state-of-the-art tails bounds are of almost (up to $\log(T)$ factors) the same order as for the average of the iterates, though the analyses are still restricted to the sub-Gaussian noise regime.

Motivated by these facts, we derive novel and general tail bounds for both the average of the iterates (Section 4) and the last iterate (Section 5). In their most general form, our results require controlling the tails of certain martingales depending only on the noise. We then show how to instantiate these bounds in the two considered noise models. Unlike most tail bounds in the (non-smooth) convex and Lipschitz setting, our results do not require a bound on the diameter of the domain. On the technical side, we extend existing analysis techniques and concentration results to cope with the challenges posed by our more general problem setting. In particular, the combination of the heavy-tailed noise with the unbounded domain and the peculiar recurrences arising in the analysis of the last iterate. Finally, some of our results for the average of the iterates show an intriguing *two-regime* phenomenon (also observed in (Lou et al., 2022) in a different and more specific setting), where the terms accounting for the heavy-tailed behavior of the noise decay more quickly with the horizon $T$. As our results for the last iterate do not exhibit this behavior, we investigate further this separation in the experiments (Section 6).

## 2 RELATED WORKS

In the case of *sub-Gaussian noise*, the performance of SGD and SMD has been analyzed in (Harvey et al., 2019; Jain et al., 2021) and (Liu et al., 2023) respectively. In (Harvey et al., 2019), the authors consider the setting with a *bounded domain*. They provide tail bounds for the average of the iterates and the last iterate of the order $\sqrt{\log(1/\delta)/T}$ and $\sqrt{\log(1/\delta)/T}\cdot\log(T)$ respectively. Jain et al. (2021) show that when the time horizon is known in advance, the last iterate enjoys the same tail bound as the average of the iterates as long as a carefully designed step-size schedule is used. Notably, this result does hold for *unbounded domains*. Liu et al. (2023) consider the more general framework of SMD with *unbounded domains*, and analyze the performance of the average of the iterates. The authors prove tail bounds of the order of $\sqrt{\log(1/\delta)/T}$ and $\sqrt{\log(1/\delta)/T\cdot\log(T)}$ for the case of known and unknown $T$ respectively.

On the other extreme of the spectrum, another research line considers very general models where the noise is only assumed to posses moments of order at most $p \in (1, 2]$. In this setting, the optimal in-expectation rates are of the order $T^{(1-p)/p}$, see (Vural et al., 2022). To obtain high-probability analogues with only a $\log(1/\delta)$ overhead, existing works consider modifications of the standard SMD algorithm where the oracle answers are pre-processed via some form of *truncation*. For $p = 2$, Parletta et al. (2022) provide tail bounds for several averaging schemes under the assumption of a *bounded domain*, where both the cases of known and unknown $T$ are considered. The *unbounded domain* setting is analyzed in (Gorbunov et al., 2021), although only in the case when $T$ is known. Similar results have been obtained for *smooth* convex objectives (Nazin et al., 2019; Gorbunov et al., 2020; Holland, 2022; Nguyen et al., 2023), where both *bounded* and *unbounded domains* have been considered. Our work is conceptually close to that of Lou et al. (2022), which explores the limits of plain SGD in the specific problem of least-squares regression with linear models. In that paper, the authors derive tails bounds for the average of the iterates under polynomially-tailed noise. We recover similar results in our more general problem setting, including the *two-regime* behavior highlighted therein.

## 3 PROBLEM SETTING

We consider the problem of minimizing a convex function $f\colon \mathcal{X} \to \mathbb{R}$, where the domain $\mathcal{X} \subseteq \mathbb{R}^d$ is a non-empty, closed, and convex set over which $f$ admits a minimum. For any $x \in \mathcal{X}$, let $\partial f(x)$ denote the sub differential at $x$. Access to the function $f$ is provided through a noisy first-order oracle. At each step $t$, the learner queries the oracle with a point $x_t \in \mathcal{X}$ and receives $\hat{g}_t \in \mathbb{R}^d$ such that $\hat{g}_t = g_t - \xi_t$, where $g_t \in \partial f(x_t)$ and $\mathbb{E}[\xi_t \,|\, \xi_1, \dots, \xi_{t-1}] = 0$.

In the following, we use $\|\cdot\|$ to refer to a fixed arbitrary norm in $\mathbb{R}^d$. Let $\psi\colon \mathbb{R}^d \to (-\infty, +\infty]$ be a convex function, and define $\mathrm{dom}(\psi) \coloneqq \{x \in \mathbb{R}^d \colon \psi(x) < +\infty\}$. For any $y \in \mathrm{dom}(\psi)$ at which $\psi$ is differentiable, the Bregman divergence at $y$ induced by $\psi$ is defined as

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla\psi(y)\rangle\,,$$

**Algorithm 1** Stochastic Mirror Descent

**input:** regularizer $\psi$ satisfying Assumption 1, non-increasing sequence of positive learning rates $(\eta_t)_t$
**initialization:** choose $x_1 \in \operatorname{int}(\operatorname{dom}(\psi))$
**for** $t = 1, \dots$ **do**
  output $x_t$ and receive $\hat{g}_t$
  set $x_{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \langle \hat{g}_t, x \rangle + \frac{1}{\eta_t} B_\psi(x, x_t)$
**end for**

for any $x \in \operatorname{dom}(\psi)$. For some $\lambda \geq 0$, $\psi$ is said to be $\lambda$-strongly convex with respect to $\|\cdot\|$ if $\psi(x) \geq \psi(y) + \langle x - y, g \rangle + (\lambda/2)\|x - y\|^2$ for any $x, y \in \operatorname{dom}(\psi)$ and $g \in \partial\psi(y)$. This directly implies that $B_\psi(x, y) \geq (\lambda/2)\|x - y\|^2$ if $\psi$ is differentiable at $y$. To specify an instance of the mirror descent framework (see Algorithm 1), one needs to select a regularizer function $\psi$, which we will assume to satisfy the following:[1]

**Assumption 1.** *The regularizer function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ is closed, differentiable on $\operatorname{int}(\operatorname{dom}(\psi))$, 1-strongly convex with respect to $\|\cdot\|$, and satisfies $\mathcal{X} \subseteq \operatorname{dom}(\psi)$ and $\operatorname{int}(\operatorname{dom}(\psi)) \neq \{\}$. Moreover, it satisfies at least one of the following: (i) $\lim_{t\to\infty} \|\nabla\psi(x_t)\|_2 \to \infty$, for any sequence $(x_t)_t$ in $\operatorname{int}(\operatorname{dom}(\psi))$ with $\lim_{t\to\infty} x_t \to x \in \partial \operatorname{dom}(\psi)$; (ii) $\mathcal{X} \subseteq \operatorname{int}(\operatorname{dom}(\psi))$.*

This is a standard assumption (see (Beck and Teboulle, 2003) or (Orabona, 2023, Section 6.4)) that serves to insure that the iterates $(x_t)_t$ returned by the mirror descent algorithm are well-defined. Denote by $\|\cdot\|_*$ the dual norm of $\|\cdot\|$, that is $\|\cdot\|_* := \sup_{\|w\| \leq 1} \langle \cdot, w \rangle$. The following assumption implies that $f$ is Lipschitz with respect to $\|\cdot\|$.

**Assumption 2.** *There exists a constant $G > 0$ such that for all $x \in \mathcal{X}$ and $g \in \partial f(x)$, $\|g\|_* \leq G$.*

Let $f^* = \min_{x \in \mathcal{X}} f(x)$ and $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$. For any $x \in \mathcal{X}$, we define the optimization error at $x$ as $f(x) - f^*$. For some time horizon $T$, our goal in this work is to prove high probability bounds on the optimization error of the average iterate $\bar{x}_T = (1/T)\sum_{t=1}^T x_t$ and the last iterate $x_T$ produced by Algorithm 1. Towards that end, we impose some restrictions on the noise vectors $(\xi_t)_t$. For what follows, let $\mathcal{F}_t$ be the sigma algebra generated by $(\xi_1, \dots, \xi_{t-1})$. Moreover, we will use $\mathbb{E}_t[\cdot]$ to denote $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. The following assumption provides a bound on the conditional second moment of $\|\xi_t\|_*$.

**Assumption 3.** *There exists a constant $\sigma > 0$ such that for every step $t \geq 1$, it holds that $\mathbb{E}_t[\|\xi_t\|_*^2] \leq \sigma^2$.*

This assumption is sufficient for proving in-expectation bounds, and tail bounds, but only of the order $1/(\delta\sqrt{T})$. We only use this as a base assumption when stating general facts. Instead, we will instantiate our results under two different (stronger) assumptions on the noise terms $(\xi_t)_t$. The first assumption involves the class of sub-Weibull random variables (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2022), which generalizes the notions of sub-Gaussian and sub-exponential random variables. For $\theta > 0$ and $\phi > 0$, we say that a random variable $X$ is sub-Weibull$(\theta, \phi)$ if it satisfies $\mathbb{E}[\exp((|X|/\phi)^{1/\theta})] \leq 2$. At $\theta = 1/2$, we recover the definition of a sub-Gaussian random variable, and at $\theta = 1$, we recover that of a sub-exponential random variable (Vershynin, 2018, Chapter 2). Via Markov's inequality, one can show that $X$ being sub-Weibull$(\theta, \phi)$ implies that for $t \geq 0$, $P(|X| \geq t) \leq 2\exp(-(t/\phi)^{1/\theta})$. In this work, our focus is on the heavy-tailed regime where $\theta \geq 1$, though we also consider the canonical case of $\theta = 1/2$ for comparison. In particular, we will consider the following assumption:

**Assumption 4.** *For some $\theta \geq 1$, there exists a constant $\phi > 0$ such that for every step $t \geq 1$, $\|\xi_t\|_*$ is sub-Weibull$(\theta, \phi)$ conditioned on $\mathcal{F}_{t-1}$; that is,*

$$\mathbb{E}\Big[\exp\Big((\|\xi_t\|_*/\phi)^{1/\theta}\Big) \mid \mathcal{F}_{t-1}\Big] \leq 2\,.$$

Alternatively, we also consider the following assumption.

**Assumption 5.** *For some $p > 4$, there exists a constant $\phi > 0$ such that for every step $t \geq 1$,*

$$\mathbb{E}\Big[(\|\xi_t\|_*/\phi)^p \mid \mathcal{F}_{t-1}\Big] \leq 1\,.$$

The above implies, via Markov's inequality, that $X$ satisifies the following polynomially decaying tail bound: $P(|X| \geq t) \leq (\phi/t)^p$ for any $t > 0$. We only consider $p > 4$ as the analyses in the sequel require studying the concentration properties of terms involving $\|\xi_t\|_*^2$.

## 4 AVERAGE ITERATE ANALYSIS

When one's concern is studying the error of the average of the iterates $\bar{x}_T$ at some time horizon $T$, a fairly standard analysis under Assumptions 1–3 yields that

$$f(\bar{x}_T) - f^* \leq \frac{1}{\eta_T T}\Bigg(B_\psi(x^*, x_1) + \sum_{t=1}^T \eta_t^2(G^2 + \sigma^2) + \underbrace{\sum_{t=1}^T \eta_t \langle \xi_t, x_t - x^* \rangle}_{:=U} + \underbrace{\sum_{t=1}^T \eta_t^2(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2)}_{:=V}\Bigg)\,.$$

[1] For a set $S \subseteq \mathbb{R}$, $\operatorname{int}(S)$ and $\partial S$ refer to its interior and boundary respectively.

It is easy to verify that $\mathbb{E}U = \mathbb{E}V = 0$, which immediately yields a bound on the error in expectation. Proving a high-probability bound, on the other hand, requires controlling both terms in high probability. For $V$, this solely depends on the assumed statistical properties of $\|\xi_t\|_*$. Whereas for $U$, one also needs to control the terms $\|x_t - x^*\|$. This presents a major obstacle if one would like to avoid scaling with a bound on the diameter of the domain in terms of $\|\cdot\|$, which might not exist in some cases. In the recent work of Liu et al. (2023), a more careful analysis distills this problem, roughly speaking, to bounding a term of the form

$$\sum_{t=1}^{T} w_t \eta_t \langle \xi_t, x_t - x^* \rangle - v_t \|x_t - x^*\|^2,$$

where $(w_t)_t$ and $(v_t)_t$ are two carefully chosen sequences of weights. Assuming that the terms $\|\xi_t\|_*$ are conditionally sub-Gaussian, as done in (Liu et al., 2023), and applying the standard Chernoff method to bound this term in high probability, this refinement has the effect of normalizing the vectors $x_t - x^*$. Unfortunately, this "white-box" approach does not readily extend beyond the light-tailed case. For instance, if the noise terms are sub-exponential, it is not clear how to deal with the additional hurdle that the moment-generating function of $\langle \xi_t, x_t - x^* \rangle$ is only bounded in a constrained range, whose diameter is inversely proportional to $\|x_t - x^*\|$.

In the more recent work of Nguyen et al. (2023), a different weighting scheme is proposed for the purpose of analyzing a clipped version of SMD in a setting where it is only assumed that the $p$-th moment of the noise is bounded for $p \in (1, 2]$. However, as presented, their analysis is still a "white-box" one, which leverages the properties of the clipped gradient estimate. In what follows, we demonstrate that a similar weighting scheme can be utilized in our setting to isolate the effect of the vectors $x_t - x^*$ in a "black-box" manner, independently of the assumed statistical properties of the noise. For $t \geq 1$, let

$$D_t = \max\Big\{\gamma, \sqrt{B_\psi(x^*, x_1)}, \ldots, \sqrt{B_\psi(x^*, x_t)}\Big\}, \quad (1)$$

where $\gamma > 0$ is a constant that will be dictated by the analysis. Normalizing per-iterate quantities with $(D_t)_t$ is a natural choice as it is a non-decreasing sequence, predictable with respect to $(\mathcal{F}_t)_t$, and most notably, it holds that $\sqrt{2}D_t \geq \sqrt{2B_\psi(x^*, x_t)} \geq \|x_t - x^*\|$. The following theorem provides a high probability bound on the error of the average of the iterates without requiring an upper bound on the diameter of the domain, as long as one can control the tails of two martingales essentially depending only on the noise.

**Theorem 1.** *Let $Y_1, Y_2 : (0,1) \times [0,\infty)^T \to (0,\infty)$ be two functions such that for any $\delta \in (0,1)$,*

$$P\bigg(\max_{s \leq T} \sum\nolimits_{t=1}^{s} \eta_t \Big\langle \xi_t, \frac{x_t - x^*}{\sqrt{2}D_t} \Big\rangle > Y_1(\delta, (\eta_t)_{t=1}^T)\bigg) \leq \delta$$

*and*

$$P\bigg(\sum\nolimits_{t=1}^{T} \eta_t^2 (\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2) > Y_2(\delta, (\eta_t)_{t=1}^T)\bigg) \leq \delta\,,$$

*where $D_t$ is as defined in (1) with $\gamma$ chosen as $\sqrt{Y_2(\delta/2, (\eta_t)_{t=1}^T) + \sum_{t=1}^T \eta_t^2(G^2 + \sigma^2)}$. Then, under Assumptions 1–3, Algorithm 1 satisfies the following with probability at least $1 - \delta$:*

$$f(\bar{x}_T) - f^* \leq \frac{3}{\eta_T T}\bigg(B_\psi(x^*, x_1) + \sum_{t=1}^{T} \eta_t^2 (G^2 + \sigma^2) + 2Y_1(\delta/2, (\eta_t)_{t=1}^T)^2 + Y_2(\delta/2, (\eta_t)_{t=1}^T)\bigg).$$

*Proof.* Lemma 5 in Appendix A with $z = x^*$ and $w_t = 1/D_t$ yields that for any $s \in [T]$

$$\frac{B_\psi(x^*, x_{s+1})}{D_s} + \sum_{t=1}^{s} \frac{\eta_t}{D_t}(f(x_t) - f^*) \leq \frac{B_\psi(x^*, x_1)}{D_1} + \sum_{t=1}^{s} \frac{\eta_t^2}{2D_t}\|\hat{g}_t\|_*^2 + \sum_{t=1}^{s} \eta_t \Big\langle \xi_t, \frac{x_t - x^*}{D_t} \Big\rangle.$$

For brevity, define $d_t = \sqrt{B_\psi(x^*, x_t)}$. Taking the previous inequality further, we have that

$$\frac{d_{s+1}^2}{D_s} + \sum_{t=1}^{s} \frac{\eta_t}{D_t}(f(x_t) - f^*) \leq d_1 + \gamma \vee \bigg(\frac{1}{\gamma}\sum_{t=1}^{T} \frac{\eta_t^2}{2}\|\hat{g}_t\|_*^2\bigg) + \sqrt{2}\bigg(\max_{s \leq T} \sum_{t=1}^{s} \eta_t \Big\langle \xi_t, \frac{x_t - x^*}{\sqrt{2}D_t} \Big\rangle\bigg)_+, \quad (2)$$

where $x \vee y = \max\{x, y\}$ and $x_+ = \max\{0, x\}$. Define $B_T$ as the right-hand side of the last inequality. Consequently, we have that $d_{s+1}^2 \leq D_s B_T$, and thanks to the non-negativity of the last term in the right-hand side of (2), we have that $D_1 = \max\{d_1, \gamma\} \leq B_T$. Moreover, if $D_s \leq B_T$ for some $s \in [T]$, then

$$D_{s+1} = \max\{D_s, d_{s+1}\} \leq \max\{D_s, \sqrt{D_s B_T}\} \leq B_T\,.$$

Thus, via induction, $D_s \leq B_T$ for all $s \in [T]$. Since $(\eta_t)_t$ and $(D_t)_t$ are non-increasing and non-decreasing respectively, we can conclude from (2) and the convexity of $f$ that

$$f(\bar{x}_T) - f^* \leq \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f^*) \leq \frac{D_T B_T}{\eta_T T} \leq \frac{B_T^2}{\eta_T T}\,. \quad (3)$$

Utilizing Assumptions 2 and 3, we have that

$$\sum_{t=1}^{T}\frac{\eta_t^2}{2}\|\hat{g}_t\|_*^2=\sum_{t=1}^{T}\frac{\eta_t^2}{2}\|g_t-\xi_t\|_*^2\le\sum_{t=1}^{T}\eta_t^2(\|g_t\|_*^2+\|\xi_t\|_*^2)$$
$$=\sum_{t=1}^{T}\eta_t^2(\|g_t\|_*^2+\mathbb{E}_t\|\xi_t\|_*^2)+\sum_{t=1}^{T}\eta_t^2(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)$$
$$\le\sum_{t=1}^{T}\eta_t^2(G^2+\sigma^2)+\sum_{t=1}^{T}\eta_t^2(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\,.$$

Combining this with the assumed tail bounds and plugging in the value of $\gamma$ yields that

$$B_T\le d_1+\sqrt{Y_2(\delta/2,(\eta_t)_{t=1}^T)+\sum_{t=1}^{T}\eta_t^2(G^2+\sigma^2)}+\sqrt{2}Y_1(\delta/2,(\eta_t)_{t=1}^T)\,,$$

with probability at least $1-\delta$, which, combined with (3), allows us to conclude the proof after simple calculations. □

Theorem 1 provides a modular bound, turning which into a concrete convergence rate requires applying suitable martingale concentration results, depending on the adopted noise model. Starting with the sub-Weibull case, Proposition 2 in Appendix E, a more versatile version of a result in Proposition 11 in (Madden et al., 2021), provides a maximal concentration inequality for martingales with conditionally sub-Weibull increments. Utilizing this results leads to the following corollary.

**Corollary 1.** *For any $\delta\in(0,1)$ and $\eta>0$, Algorithm 1, under Assumptions 1, 2 and 4, satisfies the following with probability at least $1-\delta$.*

*(i) If $\eta_t=\eta$, $f(\bar{x}_T)-f^*$ is bounded by*

$$\frac{C}{T}\left(\frac{B_\psi(x^*,x_1)}{\eta}+\eta(G^2+\phi^2\log(e/\delta))T+\eta\phi^2\log^{2\theta}(eT/\delta)\right)$$

*(ii)If $\eta_t=\frac{\eta}{\sqrt{t}}$, $f(\bar{x}_T)-f^*$ is bounded by*

$$\frac{C\log(eT)}{\sqrt{T}}\left(\frac{B_\psi(x^*,x_1)}{\eta}+\eta\left(G^2+\phi^2\log^{2\theta}(e/\delta)\right)\right)$$

*where $C$ is a constant depending only on $\theta$.*

A proof is provided in Appendix B. Firstly, we remark that these bounds can also be shown to hold in the sub-Gaussian setting (with $\theta=1/2$), where they recover the corresponding results in (Liu et al., 2023). Also notice that, regardless of $\theta$, as $\phi$ goes to zero, we recover the standard bounds for the deterministic setting. In the case when $\eta_t=\eta$ (the known time horizon setting), the bound exhibits what we will refer to as a *two-regime* behaviour. To better illustrate this, consider that an optimal tuning of $\eta$ yields a bound of order

$$\sqrt{B_\psi(x^*,x_1)}\left(\sqrt{\frac{G^2+\phi^2\log(e/\delta)}{T}}+\frac{\phi\log^{\theta}(eT/\delta)}{T}\right).$$

The first term in the brackets is the standard sub-Gaussian rate, while the second depends on the assumed shape of the noise. The key observation here is that as the horizon grows longer, the sub-Gaussian term will eventually come to dominate, masking the heavy-tailed behaviour of the noise. This turning point depends, most importantly, on the required confidence level $1-\delta$ and the shape parameter $\theta$. Specifically, it can be shown to be $\mathcal{O}(2^{2\theta}\log^{2\theta}((2\theta)^{2\theta}e/\delta))$. It is also noteworthy that the second term is primarily the contribution of the noise at a single step, a phenomenon inherited from the Freedman-style concentration inequalities on which this result is based.

In the case when $\eta_t=\eta/\sqrt{t}$ (the anytime setting), the bound in Corollary 1 is akin, in form, to results presented in (Madden et al., 2021; Li and Liu, 2022) in the non-convex setting under different assumptions.[2] However, we avoid the extra dependence on $\log^{2\theta}(T)$ featured in these works thanks to the general form of Proposition 2, which allows one to take advantage of the fact that the learning rate schedule is imbalanced to retain the same dependence on $T$ as in the light-tailed case. On the other hand, this imbalance also means that for both martingales featured in Theorem 1, the effect of the noise in the beginning (when $\eta_t$ is large) is, in a sense, comparable to that of the whole sequence. On the surface, this explains why the bound we presented in the anytime case does not exhibit the two-regime behaviour enjoyed by the first bound. The deeper cause is that the analysis relies on controlling the maximum of the terms $\|x_t-x^*\|$ in high probability, which seems to naturally result in the dominance of the heavy-tailed regime. In fact, it is not difficult (see Appendix C for the proof of a stronger statement) to show that under the assumption that $\max_t\sqrt{B_\psi(x^*,x_t)}\le D$, one can obtain a bound of order

$$\frac{1}{\sqrt{T}}\left(\frac{D^2}{\eta}+\eta\left(G^2+\phi^2\left(\log(e/\delta)+\frac{\log^{2\theta}(eT/\delta)}{\sqrt{T}}\right)\right)\right).$$

Even if one cannot generally tune $\eta$ optimally (as $T$ is unknown), the message is that as $T$ grows, the bound

[2] In these works, they consider *smooth* (possibly) non-convex objectives and provide rates for the average norm of the gradients (or the optimization error under an additional strong Polyak-Lojasiewicz condition).

approaches its sub-Gaussian counterpart. Deriving a similar guarantee without assuming a bound on the diameter remains an interesting problem.

Under Assumption 5, one can use Fuk-Nagaev type concentration inequalities (see, e.g., Rio (2017)) to control the tails of the martingales in question. Doing so, we arrive at the following corollary, whose proof is provided in Appendix B.

**Corollary 2.** *For any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1, under Assumptions 1, 2 and 5, satisfies the following with probability at least $1-\delta$.*

*(i) If $\eta_t = \eta$, $f(\bar{x}_T) - f^*$ is bounded by*

$$\frac{C}{T}\left(\frac{B_\psi(x^*, x_1)}{\eta} + \eta(G^2 + \phi^2 \log(e/\delta))T + \eta\phi^2(T/\delta)^{2/p}\right)$$

*(ii) If $\eta_t = \frac{\eta}{\sqrt{t}}$, $f(\bar{x}_T) - f^*$ is bounded by*

$$\frac{C\log(eT)}{\sqrt{T}}\left(\frac{B_\psi(x^*, x_1)}{\eta} + \eta\Big(G^2 + \phi^2(1/\delta)^{2/p}\Big)\right)$$

*where $C$ is a constant depending only on $p$.*

The bounds are analogous to the sub-Weibull case, except that the terms accounting for the heavy tailed behaviour feature a polynomial (instead of logarithmic) dependence on $1/\delta$. A suitable tuning of $\eta$ in the first case leads to a bound of order

$$\sqrt{B_\psi(x^*, x_1)}\left(\sqrt{\frac{G^2 + \phi^2\log(e/\delta)}{T}} + \frac{\phi(1/\delta)^{1/p}}{T^{1-1/p}}\right).$$

Notice that $T^{1-1/p} > T^{3/4}$; hence, also in this case, the sub-Gaussian term can dominate if the horizon is long enough, with the turning point being $\mathcal{O}((1/\delta)^{2/(p-2)})$. A similar bound was reported in Lou et al. (2022) for the particular setting of a linear regression problem with the squared loss,[3] where the two-regime behaviour of the bound was also highlighted.

In the anytime setting, similar to the sub-Weibull case, the bound retains the same dependence on $T$ as in the sub-Gaussian case, but only exhibits heavy-tailed behaviour. Analogously to the sub-Weibull case, when $\max_t \sqrt{B_\psi(x^*, x_t)} \le D$, one can prove (see Appendix C) a bound of order

$$\frac{1}{\sqrt{T}}\left(\frac{D^2}{\eta} + \eta\left(G^2 + \phi^2\left(\log(e/\delta) + \frac{(1/\delta)^{2/p}}{\sqrt{T}}\right)\right)\right)$$

The question of deriving a similar bound (for general convex and Lipschitz functions) without assuming a bound on the diameter is more pressing in this case, as the steeper polynomial dependence on $1/\delta$ would otherwise call for the use of truncation.

[3] In their setting, it was only assumed that $p > 2$.

# 5 LAST ITERATE ANALYSIS

Focusing on the anytime case, a typical last iterate analysis in the non-smooth setting (Shamir and Zhang, 2013; Harvey et al., 2019) starts with a bound of the following form:[4]

**Lemma 1.** *Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ for some constant $\eta > 0$ satisfies*

$$f(x_T) - f^* \le \frac{2}{T}\sum_{t=\lceil T/2\rceil}^{T}(f(x_t) - f^*) + \sum_{t=\lceil T/2\rceil}^{T}\langle \xi_t, w_t\rangle + \frac{\eta}{\sqrt{2T}}\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2 + \frac{\sqrt{2}}{\eta\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T} z_t\,,$$

*where, for $j < T$, $\alpha_j = \frac{1}{(T-j)(T-j+1)}$, and for any time-step $t \ge \lceil T/2\rceil$,*

$$w_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j(x_t - x_j)\,, \quad z_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j B_\psi(x_j, x_t)\,,$$

$$\text{and} \quad \rho_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j\,.$$

The first term in the bound can be dealt with using the techniques of the previous section, the third term appears in the analysis of the previous section (albeit with different weights) and can be handled similarly, while the last term is usually handled using a uniform bound on the divergence terms, though this is not necessary as we will see. It is not difficult then to show that these three terms decay at a rate of at most $\log(T)/\sqrt{T}$ with high probability. The main obstacle in the way of proving a tail bound for the error is showing that the second (martingale) term enjoys a similar rate. Naively bounding the norms of the vectors $w_t$ using a diameter bound is not sufficient. Instead, one needs to exploit the peculiar structure of this term. For the following, define the martingale sequence $(Q_s)_{s=\lceil T/2\rceil}^{T}$ where $Q_s = \sum_{t=\lceil T/2\rceil}^{s}\langle \xi_t, w_t\rangle$, and denote by $\langle Q\rangle_s$ its total conditional variance (TCV), i.e., $\langle Q\rangle_s = \sum_{t=\lceil T/2\rceil}^{s}\mathbb{E}_t\langle \xi_t, w_t\rangle^2$. Via the convexity of $\|\cdot\|^2$ and the fact that $\|x_t - x_j\|^2 \le 2B_\psi(x_j, x_t)$, it holds that $\|w_t\|^2 \le 2\rho_t z_t$. Thus, under Assumption 3, one can verify that $\langle Q\rangle_s \le 2\sigma^2\sum_{t=\lceil T/2\rceil}^{s}\rho_t z_t$. The key observation of Harvey et al. (2019) is that this sum can be bounded with an affine function of the martingale itself. Via a generalized version of Freedman's inequality, the authors exploit the resulting fact that $\langle Q\rangle_T$ is upper bounded with a suitable affine function of $Q_T$ to

[4] Proofs for the results presented in this section can be found in Appendix D.

arrive at the desired tail bound. This inequality, however, is once again specific to the sub-Gaussian noise setting, beyond which one usually needs finer control on the individual $w_t$ terms, as argued in the previous section. Hence, once again, we seek an approach through which we can disentangle the vectors $w_t$ from the noise terms $\xi_t$. The following lemma provides a starting point by showing that $z^* \coloneqq \max_{\lceil T/2 \rceil \le s \le T} z_s$ can itself be related to $(Q_s)_s$.

**Lemma 2.** *In the same setting as Lemma 1, it holds that*

$$z^* \le \frac{6\sqrt{2}\eta}{T\sqrt{T}} \sum\nolimits_{t=\lceil T/2 \rceil}^{T} (f(x_t) - f^*) + \frac{3\sqrt{2}\eta}{\sqrt{T}} Q_{n^*} + \frac{3\eta^2}{T} \sum\nolimits_{t=\lceil T/2 \rceil}^{T} \rho_t \|\hat{g}_t\|_*^2 ,$$

*where* $n^* = \min\{n : n \in \arg\max_{\lceil T/2 \rceil \le s \le T} Q_s\}$.

A nice implication of this lemma is that the last term in the bound of Lemma 1 can be related to the preceding terms. However, at this point, this lemma does not provide a tight (high probability) bound on the $z_t$ (or $\|w_t\|$) terms due to the dependence on $Q_{n^*}$. Thus, techniques relying on such a bound, like the averaging scheme of the previous section or extensions of the concentration result of Harvey et al. (2019) to sub-Weibull random variables in (Madden et al., 2021, Proposition 11),[5] are not easily utilizable. Instead, the real advantage of this lemma is that it allows one to relate not only the TCV but also the total quadratic variation (TQV) of $Q_T$, given by $[Q]_T = \sum_{t=\lceil T/2 \rceil}^{T} \langle \xi_t, w_t \rangle^2$, back to the martingale itself through $z^*$:

**Lemma 3.** *In the same setting as Lemma 1, it holds under Assumption 3 that*

$$\langle Q \rangle_T + [Q]_T \le 4\sigma^2 z^* \log(4T) + 2z^* \sum\nolimits_{t=\lceil T/2 \rceil}^{T} \rho_t (\|\xi_t\|_*^2 - \mathbb{E}_t \|\xi_t\|_*^2) .$$

The sum in the second term occurs also when bounding the third term in the bound of Lemma 1, and has been encountered in the average iterate analysis. Notice that, trivially, the left hand side of Lemma 3 is also a bound for the sum of the TCV and TQV at any step, particularly at $n^*$. Being able to bound this sum allows one to derive powerful concentration results with few assumptions. In the next proposition, we extend one such result, Theorem 2.1 in (Bercu and Touati, 2008), in the spirit of Theorem 3.3 in (Harvey et al., 2019).

**Proposition 1.** *Let $(M_t)_{t=0}^n$ be a square integrable martingale adapted to filtration $(\mathcal{F}_t)_{t=0}^n$ with $M_0 = 0$. Then, for all $x, \beta > 0$ and $\alpha \ge 0$,*

$$P\left( \bigcup_{t=1}^{n} \{M_t \ge x \text{ and } \langle M \rangle_t + [M]_t \le \alpha M_t + \beta\} \right) \le \exp\left( - \min\left\{ \frac{x^2}{8\beta}, \frac{x}{6\alpha} \right\} \right) .$$

Utilizing this tool, together with the preceding lemmas, we arrive at the following general bound for the last iterate.

**Theorem 2.** *Let $\Xi_1, \Xi_2 : (0,1) \to (0,\infty)$ be two functions such that for any $\delta \in (0,1)$,*

$$P\left( \frac{1}{\sqrt{T}} \sum\nolimits_{t=1}^{T} (f(x_t) - f^*) > \Xi_1(\delta) \right) \le \delta$$

*and*

$$P\left( \sum\nolimits_{t=\lceil T/2 \rceil}^{T} \rho_t (\|\xi_t\|_*^2 - \mathbb{E}_t \|\xi_t\|_*^2) > \Xi_2(\delta) \right) \le \delta .$$

*Then, under Assumptions 1–3, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies the following with probability at least $1-\delta$:*

$$f(x_T) - f^* \le \frac{35}{\sqrt{T}} \Bigg( 2\Xi_1(\delta/3) + \sqrt{2}\eta G^2 \log(4T) + 9\sqrt{2}\eta \Big( \Xi_2(\delta/3) + 2\sigma^2 \log(4T) \Big) \log(3/\delta) \Bigg) .$$

To obtain a concrete bound, one needs a tail bound for the error of the average iterate and a similar bound for a by-now-familiar martingale term. The following corollary provides concrete bounds for our two noise models.

**Corollary 3.** *For any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies the following with probability at least $1-\delta$, where $C_1$ and $C_2$ are constant depending solely on, respectively, $\theta$ and $p$.*
*(i) Under Assumptions 1, 2 and 4, $f(x_T) - f^*$ is bounded by*

$$\frac{C_1 \log(eT)}{\sqrt{T}} \left( \frac{B_\psi(x^*, x_1)}{\eta} + \eta \left( G^2 + \phi^2 \log^{2\theta+1}(e/\delta) \right) \right)$$

*(ii) Under Assumptions 1, 2 and 5, $f(x_T) - f^*$ is bounded by*

$$\frac{C_2 \log(eT)}{\sqrt{T}} \left( \frac{B_\psi(x^*, x_1)}{\eta} + \eta \Big( G^2 + \phi^2 (1/\delta)^{2/p} \log(e/\delta) \Big) \right)$$

Firstly, these bounds retain the same decay rate in $T$ as that in the deterministic case, whose bounds are recovered as the noise vanishes. Compared to their counterparts in the average case, both bounds contain an

[5] The latter would actually require an almost sure bound.

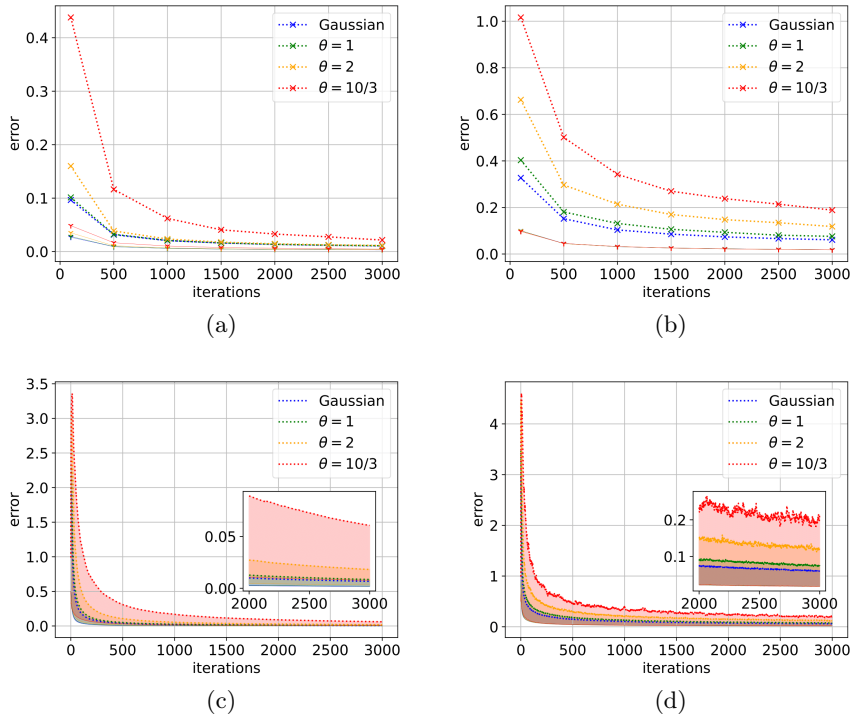



Figure 1: The performance of the average iterate and the last iterate are reported in the plots on the left and the right columns respectively. Solid lines show the average of the error across runs and dotted lines show its 99-percentile. The top row refers to the case of $\eta = 1/\sqrt{T}$, while the bottom row refers to the case of $\eta = 1/\sqrt{t}$. The zoomed plots highlight the performance in the last 1k iterations.

extra $\log(1/\delta)$ factor, an artifact of the general disentanglement technique we adopt. While this means that the exact sub-Gaussian rate is not recovered, this factor is arguably negligible for heavier noise. Although we focused on the anytime learning rates $\eta/\sqrt{t}$, similar results can be straightforwardly verified to hold when using a constant learning rate. Interestingly, for either schedule, the bounds obtainable from this analysis do not assume the two-regime form. The main obstacle for this is encountered as early as the fairly standard Lemma 1, and is manifested in the third term therein. This term leads to the dominance of the heavy-tailed regime, primarily through the contribution of the noise in the final iterates, where $\rho_t$ is $\Theta(1)$. Beyond the standard step-size choices, extending the analysis of the scheme proposed by Jain et al. (2021) to heavy-tailed noise is an interesting problem.

## 6 EXPERIMENTS

We present two experiments comparing the performance of the average of the iterates with that of the last iterate when using Algorithm 1 to minimize $f(x) = |x|$ over $\mathbb{R}$ with $\psi(x) = 1/2\|x\|_2^2$ (i.e., classical SGD). For the noise, we consider the Gaussian distribution with variance 1 and three different Weibull distributions with $\theta = 1, 2, 10/3$ respectively (see Appendix F for an additional experiment concerning polynomially-tailed noise). For a fair comparison, the Weibull distributions are scaled to have unit variance. In each experiment, we run the algorithm for 3k iterations, repeated 20k times. We report the average and the 99-percentile of the optimization errors. In the first experiment, we use $1/\sqrt{T}$ as a fixed step-size and run the algorithm for seven values of $T$ ranging from 100 to 3k, reporting only the errors at the end of each run. The results for the average iterate and the last iterate are reported in plots (a) and (b) respectively. While in both plots the average error is almost the same across noise levels (due to the normalization), the 99-percentile curves show a significant difference in behaviour between the two plots. In particular, for the average iterate, the curves for the heavy-tailed noise distributions approach the Gaussian level as the horizon grows, as predicted by the two-regime bounds. Whereas for the last iterate, the different noise levels exhibit a clear separation for all values of $T$, indicating higher sensitivity to heavy-tailed noise. In the second experiment, we set $\eta_t = 1/\sqrt{t}$ and report the evolution of the error through the 3k iterations for the average

iterate and the last iterate in plots (c) and (d) respectively. We observe once again that the 99-percentile curves for the last iterate remain well separated across the entire run. On the other hand, in the average iterate case, the very small scale of the $y$-axis in the zoomed plot and the steeper slope of the 99-percentile curves (with respect to the Gaussian one) seem to hint towards a two-regime behaviour in the anytime case as well.

## References


Milad Bakhshizadeh, Arian Maleki, and Victor H de la Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3):1655–1685, 2023.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175, 2003.

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5): 1848–1869, 2008.

Yu. M. Ermol'ev. On the method of generalized stochastic gradients and quasi-Féjer sequences. *Cybernetics*, 5:208–220, 1969.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pages 15042–15053, 2020.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the 32nd Conference on Learning Theory*, pages 1579–1613, 2019.

Matthew J. Holland. Anytime guarantees under heavy-tailed data. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 6918–6925, 2022.

Prateek Jain, Dheeraj M. Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. *SIAM J. Optim.*, 31(2):1108–1130, 2021.

Seunghyun Kim, Liam Madden, and Emiliano Dall'Anese. Online stochastic gradient methods under sub-weibull noise and the polyak-Łojasiewicz condition. In *Proceedings of the IEEE 61st Conference on Decision and Control (CDC)*, pages 3499–3506, 2022.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11 (4):1389–1456, 2022.

Chris Junchi Li and Michael I Jordan. Nonconvex stochastic scaled gradient descent and generalized eigenvector problems. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1230–1240, 2023.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12931–12963, 2022.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21884–21914, 2023.

Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *The Journal of Machine Learning Research*, 23(1):2227–2248, 2022.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2021.

Alexander V. Nazin, Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Ta Duy Nguyen, Alina Ene, and Huy L Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2023.

Daniela A. Parletta, Andrea Paudice, Massimiliano Pontil, and Saverio Salzo. High probability bounds

for stochastic subgradient schemes with heavy tailed noise. *arXiv preprint arXiv:2208.08567*, 2022.

Emmanuel Rio. About the constants in the Fuk-Nagaev inequalities. *Electronic Communications in Probability*, 22:1–12, 2017.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

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 *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 65–102. PMLR, 2022.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511813658.

Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *SIAM Journal on Optimization*, 33(3): 1943–1967, 2023.

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 *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pages 15383–15393, 2020.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A BASIC RESULTS FOR STOCHASTIC MIRROR DESCENT

The following lemma is a standard result for mirror descent; see, for example, (Orabona, 2023, Lemma 6.9).

**Lemma 4.** *For any $z \in \mathcal{X}$, the iterates $(x_t)_t$ output by Algorithm 1 satisfy*

$$f(x_t) - f(z) \leq \frac{1}{\eta_t} B_\psi(z, x_t) - \frac{1}{\eta_t} B_\psi(z, x_{t+1}) + \langle \xi_t, x_t - z \rangle + \frac{\eta_t}{2} \|\hat{g}_t\|_*^2 .$$

*Proof.* Since $x_{t+1}$ is the minimizer of the convex function $\Phi_t(x) = \langle \hat{g}_t, x \rangle + \frac{1}{\eta_t} B_\psi(x, x_t)$ in $\mathcal{X}$, it satisfies that for any $z \in \mathcal{X}$,

$$\langle \hat{g}_t + (1/\eta_t)\nabla\psi(x_{t+1}) - (1/\eta_t)\nabla\psi(x_t), x_{t+1} - z \rangle = \langle \nabla\Phi_t(x_{t+1}), x_{t+1} - z \rangle \leq 0 . \tag{4}$$

Hence,

$$\begin{aligned}
\eta_t \langle \hat{g}_t, x_t - z \rangle &= \eta_t \langle \hat{g}_t, x_t - x_{t+1} \rangle + \eta_t \langle \hat{g}_t, x_{t+1} - z \rangle \\
&= \eta_t \langle \hat{g}_t, x_t - x_{t+1} \rangle + \langle \eta_t \hat{g}_t + \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_{t+1} - z \rangle + \langle \nabla\psi(x_t) - \nabla\psi(x_{t+1}), x_{t+1} - z \rangle \\
&\overset{(a)}{\leq} \eta_t \langle \hat{g}_t, x_t - x_{t+1} \rangle + \langle \nabla\psi(x_t) - \nabla\psi(x_{t+1}), x_{t+1} - z \rangle \\
&\overset{(b)}{=} \eta_t \langle \hat{g}_t, x_t - x_{t+1} \rangle + B_\psi(z, x_t) - B_\psi(z, x_{t+1}) - B_\psi(x_{t+1}, x_t) \\
&\overset{(c)}{\leq} B_\psi(z, x_t) - B_\psi(z, x_{t+1}) + \eta_t \|\hat{g}_t\|_* \|x_t - x_{t+1}\| - \frac{1}{2}\|x_t - x_{t+1}\|^2 \\
&\overset{(d)}{\leq} B_\psi(z, x_t) - B_\psi(z, x_{t+1}) + \frac{\eta_t^2}{2} \|\hat{g}_t\|_*^2 ,
\end{aligned}$$

where (a) holds via (4), (b) holds via (Beck and Teboulle, 2003, Lemma 4.1), (c) holds by the 1-strong convexity of $\psi$ and the fact that (by the definition of the dual norm) $\|\hat{g}_t\|_* = \sup_{x \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \langle \hat{g}_t, x/\|x\| \rangle$, and (d) holds since $ax - (1/2)x^2 \leq (1/2)a^2$ for $x, a \in \mathbb{R}$. After dividing by $\eta_t$, the lemma follows using that $\hat{g}_t = g_t - \xi_t$ and the fact that $\langle g_t, x_t - z \rangle \geq f(x_t) - f(z)$ as $g_t \in \partial f(x_t)$. □

**Lemma 5.** *For any $z \in \mathcal{X}$ and any non-increasing sequence of positive weights $(w_t)_t$, Algorithm 1 satisfies that for any $s \geq 1$,*

$$w_s B_\psi(z, x_{s+1}) + \sum_{t=1}^{s} w_t \eta_t (f(x_t) - f(z)) \leq w_1 B_\psi(z, x_1) + \sum_{t=1}^{s} \frac{w_t \eta_t^2}{2} \|\hat{g}_t\|_*^2 + \sum_{t=1}^{s} w_t \eta_t \langle \xi_t, x_t - z \rangle .$$

*Proof.* Since both $\eta_t$ and $w_t$ are non-negative, it follows from Lemma 4 that

$$w_t \eta_t (f(x_t) - f(z)) \leq w_t B_\psi(z, x_t) - w_t B_\psi(z, x_{t+1}) + w_t \eta_t \langle \xi_t, x_t - z \rangle + \frac{w_t \eta_t^2}{2} \|\hat{g}_t\|_*^2 .$$

Using that $(w_t)_t$ is a non-increasing sequence, we have that

$$\begin{aligned}
\sum_{t=1}^{s} w_t (B_\psi(z, x_t) - B_\psi(z, x_{t+1})) &= w_1 B_\psi(z, x_1) - w_s B_\psi(z, x_{s+1}) + \sum_{t=2}^{s} B_\psi(z, x_t)(w_t - w_{t-1}) \\
&\leq w_1 B_\psi(z, x_1) - w_s B_\psi(z, x_{s+1}) ,
\end{aligned}$$

which entails that

$$\sum_{t=1}^{s} w_t \eta_t (f(x_t) - f(z)) \leq w_1 B_\psi(z, x_1) - w_s B_\psi(z, x_{s+1}) + \sum_{t=1}^{s} w_t \eta_t \langle \xi_t, x_t - z \rangle + \sum_{t=1}^{s} \frac{w_t \eta_t^2}{2} \|\hat{g}_t\|_*^2 .$$

□

**Lemma 6.** *Let $j$ and $r$ be two time indices such that $j \leq r$, and define $\tilde{\eta}_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}$ and $\tilde{\eta}_1 = \frac{1}{\eta_1}$. Then, Algorithm 1 satisfies that*

$$\frac{1}{\eta_r} B_\psi(x_j, x_{r+1}) + \sum_{t=j}^{r} (f(x_t) - f(x_j)) \leq \sum_{t=j}^{r} \langle \xi_t, x_t - x_j \rangle + \frac{1}{2} \sum_{t=j}^{r} \eta_t \|\hat{g}_t\|_*^2 + \sum_{t=j}^{r} \tilde{\eta}_t B_\psi(x_j, x_t) .$$

*Proof.* For $t \geq j$, Lemma 4 implies that

$$f(x_t) - f(x_j) \leq \frac{1}{\eta_t} B_\psi(x_j, x_t) - \frac{1}{\eta_t} B_\psi(x_j, x_{t+1}) + \langle \xi_t, x_t - x_j \rangle + \frac{\eta_t}{2} \|\hat{g}_t\|_*^2 .$$

Notice that,

$$\begin{aligned}\sum_{t=j}^{r} \left( \frac{1}{\eta_t} B_\psi(x_j, x_t) - \frac{1}{\eta_t} B_\psi(x_j, x_{t+1}) \right) &= \frac{1}{\eta_j} B_\psi(x_j, x_j) - \frac{1}{\eta_r} B_\psi(x_j, x_{r+1}) + \sum_{t=j+1}^{r} \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) B_\psi(x_j, x_t) \\ &= -\frac{1}{\eta_r} B_\psi(x_j, x_{r+1}) + \sum_{t=j}^{r} \tilde{\eta}_t B_\psi(x_j, x_t) ,\end{aligned}$$

where we have used that $B_\psi(x_j, x_j) = 0$. Thus, we conclude that

$$\sum_{t=j}^{r} (f(x_t) - f(x_j)) \leq \sum_{t=j}^{r} \langle \xi_t, x_t - x_j \rangle + \frac{1}{2} \sum_{t=j}^{r} \eta_t \|\hat{g}_t\|_*^2 - \frac{1}{\eta_r} B_\psi(x_j, x_{r+1}) + \sum_{t=j}^{r} \tilde{\eta}_t B_\psi(x_j, x_t) .$$

□

# B PROOFS OF SECTION 4

Before proving Corollaries 1 and 2, we state two lemmas specializing Propositions 2 and 3 in Appendix E to the two martingales we encounter when analyzing SMD.

**Lemma 7.** *Let $(\omega_t)_{t=1}^T$ be a sequence of positive (deterministic) weights with $\omega_*$ denoting their maximum. Additionally, let $(u_t)_{t=1}^T$ be a sequence of vectors in $\mathbb{R}^d$ such that $u_t$ is $\mathcal{F}_{t-1}$-measurable and $\|u_t\| \leq 1$. Then, under Assumption 4, the following holds for any $\delta \in (0,1)$ and $s \geq 0$.*

*(i)*

$$P\left( \max_{k \in [T]} \sum_{t=1}^{k} \omega_t \langle \xi_t, u_t \rangle \geq \phi \sqrt{C_1 \sum_{t=1}^{T} \omega_t^2 \log(2/\delta)} + 4\phi\omega_* C_2 \log^\theta \left( \frac{2e \sum_{t=1}^T \omega_t^s}{\omega_*^s \delta} \right) \right) \leq \delta ,$$

*where $C_1 = 2^{3\theta+1}\Gamma(3\theta+1)$ and $C_2 = \max\{1, (s\theta - s)^{\theta-1}\}$.*

*(ii)*

$$P\left( \max_{k \in [T]} \sum_{t=1}^{k} \omega_t (\|\xi_t\|_*^2 - \mathbb{E}_t \|\xi_t\|_*^2) \geq C_3 \phi^2 \sqrt{C_1 \sum_{t=1}^{T} \omega_t^2 \log(2/\delta)} + 4 C_2 C_3 \phi^2 \omega_* \log^{2\theta} \left( \frac{2e \sum_{t=1}^T \omega_t^s}{\omega_*^s \delta} \right) \right) \leq \delta ,$$

*where $C_1 = 2^{6\theta+1}\Gamma(6\theta+1)$, $C_2 = \max\{1, (2s\theta - s)^{2\theta-1}\}$, and $C_3 = 2^{2\theta+1}\Gamma(2\theta+1)/\ln^{2\theta}(2)$.*

*Proof. (i)* Since $\|u_t\| \leq 1$, the definition of the dual norm implies that $|\omega_t \langle \xi_t, u_t \rangle| \leq \omega_t \|u_t\| \|\xi_t\|_* \leq \omega_t \|\xi_t\|_*$, yielding that $\omega_t \langle \xi_t, u_t \rangle$ is sub-Weibull$(\theta, \omega_t \phi)$ conditioned on $\mathcal{F}_{t-1}$. The result then follows from Proposition 2(ii).

*(ii)* Using the definition of the sub-Weibull property, one can easily verify that if a random variable X is sub-Weibull$(\theta, \phi)$; then, $X^2$ is sub-Weibull$(2\theta, \phi^2)$. Using this along with Lemma 13 yields that $\omega_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2)$ is sub-Weibull$(2\theta, c_\theta \omega_t \phi^2)$ conditioned on $\mathcal{F}_{t-1}$, where $c_\theta = 2^{2\theta+1}\Gamma(2\theta+1)/\ln^{2\theta}(2)$. Hence, the result once more follows from Proposition 2(ii). □

**Lemma 8.** *Let $(\omega_t)_{t=1}^T$ be a sequence of positive (deterministic) weights with $\omega_*$ denoting their maximum. Additionally, let $(u_t)_{t=1}^T$ be a sequence of vectors in $\mathbb{R}^d$ such that $u_t$ is $\mathcal{F}_{t-1}$-measurable and $\|u_t\| \leq 1$. Then, under Assumption 5, the following holds for any $\delta \in (0,1)$.*

*(i)*

$$P\left( \max_{k \in [T]} \sum_{t=1}^{k} \omega_t \langle \xi_t, u_t \rangle > \phi \sqrt{2 \sum_{t=1}^{T} \omega_t^2 \log(1/\delta)} + (2 + (p/3))\phi \left( \sum_{t=1}^{T} \omega_t^p / \delta \right)^{1/p} \right) \leq \delta .$$

*(ii)*

$$P\left(\max_{k\in[T]}\sum_{t=1}^{k}\omega_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2) > 2\phi^2\sqrt{2\sum_{t=1}^{T}\omega_t^2\log(1/\delta)} + 2(2+(p/6))\phi^2\left(\sum_{t=1}^{T}\omega_t^{p/2}/\delta\right)^{2/p}\right) \le \delta\,.$$

*Proof. (i)* From the definition of the dual norm and the fact that $\|u_t\|\le 1$, we have that

$$\mathbb{E}\Big[|\omega_t\langle\xi_t,u_t\rangle|^p \,|\, \mathcal{F}_{t-1}\Big] \le \omega_t^p\mathbb{E}\Big[\|u_t\|^p\|\xi_t\|_*^p \,|\, \mathcal{F}_{t-1}\Big] \le \omega_t^p\mathbb{E}\Big[\|\xi_t\|_*^p \,|\, \mathcal{F}_{t-1}\Big] \le (\omega_t\phi)^p\,,$$

where the last inequality follows form Assumption 5. The result then follows from Proposition 3.

*(ii)* On the other hand,

$$\mathbb{E}\Big[|\omega_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)|^{p/2} \,|\, \mathcal{F}_{t-1}\Big] \le 2^{p/2}\omega_t^{p/2}\mathbb{E}\Big[\|\xi_t\|_*^p \,|\, \mathcal{F}_{t-1}\Big] \le (2\omega_t\phi^2)^{p/2}\,,$$

where the first inequality follows from Lemma 15 and the second follows from Assumption 5. Consequently, the result follows once more from Proposition 3. □

### B.1 Proof of Corollary 1

**Corollary 1.** *For any $\delta\in(0,1)$ and $\eta>0$, Algorithm 1, under Assumptions 1, 2 and 4, satisfies the following with probability at least $1-\delta$.*

*(i) If $\eta_t=\eta$, $f(\bar{x}_T)-f^*$ is bounded by*

$$\frac{C}{T}\left(\frac{B_\psi(x^*,x_1)}{\eta} + \eta(G^2+\phi^2\log(e/\delta))T + \eta\phi^2\log^{2\theta}(eT/\delta)\right)$$

*(ii)If $\eta_t=\frac{\eta}{\sqrt{t}}$, $f(\bar{x}_T)-f^*$ is bounded by*

$$\frac{C\log(eT)}{\sqrt{T}}\left(\frac{B_\psi(x^*,x_1)}{\eta} + \eta\left(G^2+\phi^2\log^{2\theta}(e/\delta)\right)\right)$$

*where $C$ is a constant depending only on $\theta$.*

*Proof.* For $t\in[T]$, let $u_t=(x_t-x^*)/(\sqrt{2}D_t)$, while for $k\in[T]$, we define

$$W_k=\sum_{t=1}^{k}\eta_t\langle\xi_t,u_t\rangle \qquad\text{and}\qquad V_k=\sum_{t=1}^{k}\eta_t^2(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\,.$$

As argued before, it holds that $\sqrt{2}D_t\ge\|x_t-x^*\|$, implying that $\|u_t\|\le 1$. For what follows, we will use $C, C_1, C_2, \ldots$ to denote positive constants—depending only on $\theta$—whose values may change between steps.

**Case *(i)*:** $\eta_t=\eta$
Starting with $(W_k)$, we invoke Lemma 7(i) with $s=0$ and $\omega_t=\eta$ obtaining that

$$P\left(\max_{k\in[T]}W_k \ge C_1\eta\phi\sqrt{T\log(2/\delta)} + C_2\eta\phi\log^{\theta}(2eT/\delta)\right)\le\delta\,.$$

For $(V_k)$, we invoke Lemma 7(ii) with $s=0$ and $\omega_t=\eta^2$ to get that

$$P\left(\max_{k\in[T]}V_k \ge C_1\eta^2\phi^2\sqrt{T\log(2/\delta)} + C_2\eta^2\phi^2\log^{2\theta}(2eT/\delta)\right)\le\delta\,.$$

With these tail bounds, Theorem 1 implies that

$$\begin{aligned}\frac{\eta T}{3}(f(\bar{x}_T)-f^*) &\le B_\psi(x^*,x_1)+\eta^2(G^2+C_1\phi^2)T + C_2\eta^2\phi^2\Big(T\log(4/\delta)+\log^{2\theta}(4eT/\delta)\Big)\\ &\qquad + C_3\eta^2\phi^2\Big(\sqrt{T\log(4/\delta)}+\log^{2\theta}(4eT/\delta)\Big)\\ &\le B_\psi(x^*,x_1)+\eta^2(G^2+C_1\phi^2)T + C_2\eta^2\phi^2\Big(T\log(4/\delta)+\log^{2\theta}(4eT/\delta)\Big)\,,\end{aligned}$$

where we have used the fact that Assumption 4 implies Assumption 3 with $\sigma^2 = 2\Gamma(2\theta+1)\phi^2$ thanks to Lemma 12. Subsequently, we have that

$$f(\bar{x}_T) - f^* \leq \frac{C}{T}\left(\frac{1}{\eta}B_\psi(x^*, x_1) + \eta(G^2 + \phi^2 \log(e/\delta))T + \eta\phi^2 \log^{2\theta}(eT/\delta)\right).$$

**Case *(ii)*:** $\eta_t = \eta/\sqrt{t}$
For $(W_k)$, we use Lemma 7(i) with $s = 3$ and $\omega_t = \eta/\sqrt{t}$, while for $(V_k)$, we use the Lemma 7(ii) with $s = 2$ and $\omega_t = \eta^2/t$ yielding that

$$P\left(\max_{k\in[T]} W_k \geq C_1\eta\phi\sqrt{\sum_{t=1}^{T}(1/t)\log(2/\delta)} + C_2\eta\phi\log^\theta\left(2e\sum_{t=1}^{T}(1/t)^{3/2}/\delta\right)\right) \leq \delta\,,$$

and

$$P\left(\max_{k\in[T]} V_k \geq C_1\eta^2\phi^2\sqrt{\sum_{t=1}^{T}(1/t)^2\log(2/\delta)} + C_2\eta^2\phi^2\log^{2\theta}\left(2e\sum_{t=1}^{T}(1/t)^2/\delta\right)\right) \leq \delta\,.$$

Combining this with the facts that

$$\sum_{t=1}^{T}\frac{1}{t} \leq \log(eT)\,, \qquad \sum_{t=1}^{T}\frac{1}{t^{3/2}} \leq 3\,, \qquad \text{and} \qquad \sum_{t=1}^{T}\frac{1}{t^2} \leq 2\,,$$

implies via Theorem 1 that

$$\begin{aligned}\frac{\eta\sqrt{T}}{3}(f(\bar{x}_T) - f^*) &\leq B_\psi(x^*, x_1) + \eta^2(G^2 + C_1\phi^2)\log(eT) + C_2\eta^2\phi^2\Big(\log(eT)\log(4/\delta) + \log^{2\theta}(12e/\delta)\Big) \\ &\qquad + C_3\eta^2\phi^2\Big(\sqrt{\log(4/\delta)} + \log^{2\theta}(8e/\delta)\Big) \\ &\leq B_\psi(x^*, x_1) + \eta^2(G^2 + C_1\phi^2)\log(eT) + C_2\eta^2\phi^2\Big(\log(eT)\log(4/\delta) + \log^{2\theta}(12e/\delta)\Big) \\ &\leq B_\psi(x^*, x_1) + \eta^2(G^2 + C_1\phi^2)\log(eT) + C_2\eta^2\phi^2\log(eT)\log^{2\theta}(12e/\delta)\,,\end{aligned}$$

where we have again used Lemma 12 to bound $\mathbb{E}_t\|\xi_t\|_*^2$ in terms of $\phi^2$ (in place of $\sigma^2$) under Assumption 4. Hence, we conclude that

$$f(\bar{x}_T) - f^* \leq \frac{C\log(eT)}{\sqrt{T}}\left(\frac{1}{\eta}B_\psi(x^*, x_1) + \eta\left(G^2 + \phi^2\log^{2\theta}(e/\delta)\right)\right).$$

□

### B.2 Proof of Corollary 2

**Corollary 2.** *For any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1, under Assumptions 1, 2 and 5, satisfies the following with probability at least $1-\delta$.*

*(i) If $\eta_t = \eta$, $f(\bar{x}_T) - f^*$ is bounded by*

$$\frac{C}{T}\left(\frac{B_\psi(x^*, x_1)}{\eta} + \eta(G^2 + \phi^2\log(e/\delta))T + \eta\phi^2(T/\delta)^{2/p}\right)$$

*(ii) If $\eta_t = \frac{\eta}{\sqrt{t}}$, $f(\bar{x}_T) - f^*$ is bounded by*

$$\frac{C\log(eT)}{\sqrt{T}}\left(\frac{B_\psi(x^*, x_1)}{\eta} + \eta\Big(G^2 + \phi^2(1/\delta)^{2/p}\Big)\right)$$

*where $C$ is a constant depending only on $p$.*

*Proof.* Similar to the proof of Corollary 1, we define $u_t = (x_t - x^*)/(\sqrt{2}D_t)$ (which satisfies $\|u_t\| \leq 1$), and consider once again the two martingale terms

$$W_k = \sum_{t=1}^{k} \eta_t \langle \xi_t, u_t \rangle \qquad \text{and} \qquad V_k = \sum_{t=1}^{k} \eta_t^2 (\|\xi_t\|_*^2 - \mathbb{E}_t \|\xi_t\|_*^2) \,.$$

For $(W_k)$, we use Lemma 8(i) with $\omega_t = \eta_t$, while for $(V_k)$, we use the Lemma 8(ii) with $\omega_t = \eta_t^2$ yielding that

$$P\left( \max_{k \in [T]} W_k > \phi \sqrt{2 \sum_{t=1}^{T} \eta_t^2 \log(1/\delta)} + (2 + (p/3)) \phi \left( \sum_{t=1}^{T} \eta_t^p / \delta \right)^{1/p} \right) \leq \delta \,,$$

and

$$P\left( \max_{k \in [T]} V_k > 2\phi^2 \sqrt{2 \sum_{t=1}^{T} \eta_t^4 \log(1/\delta)} + 2(2 + (p/6)) \phi^2 \left( \sum_{t=1}^{T} \eta_t^p / \delta \right)^{2/p} \right) \leq \delta \,.$$

For what follows, we will use $C$ to denote a positive constant—depending only on $p$—whose value may change between steps.

**Case *(i)*:** $\eta_t = \eta$
Theorem 1 with the tail bounds above yields that

$$\begin{aligned}
\frac{\eta T}{3}(f(\bar{x}_T) - f^*) &\leq B_\psi(x^*, x_1) + \eta^2 (G^2 + \phi^2) T + 4\eta^2 \phi^2 \Big( 2T \log(2/\delta) + (2 + (p/3))^2 (2T/\delta)^{2/p} \Big) \\
&\qquad + 2\eta^2 \phi^2 \Big( \sqrt{2T \log(2/\delta)} + (2 + (p/6))(2T/\delta)^{2/p} \Big) \\
&\leq B_\psi(x^*, x_1) + \eta^2 (G^2 + \phi^2) T + 6\eta^2 \phi^2 \Big( 2T \log(2/\delta) + (2 + (p/3))^2 (2T/\delta)^{2/p} \Big) \,,
\end{aligned}$$

where we have used the fact that Assumption 5 implies Assumption 3 with $\sigma^2 = \phi^2$. Subsequently, we have that

$$\begin{aligned}
f(\bar{x}_T) - f^* &\leq \frac{3}{T} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \Big( G^2 + \phi^2 (1 + 12 \log(2/\delta)) \Big) T + 6\eta \phi^2 (2 + (p/3))^2 (2T/\delta)^{2/p} \right) \\
&\leq \frac{C}{T} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta (G^2 + \phi^2 \log(e/\delta)) T + \eta \phi^2 (T/\delta)^{2/p} \right) .
\end{aligned}$$

**Case *(ii)*:** $\eta_t = \eta/\sqrt{t}$
Using that

$$\sum_{t=1}^{T} \eta_t^2 = \eta^2 \sum_{t=1}^{T} \frac{1}{t} \leq \eta^2 \log(eT) \,, \qquad \sum_{t=1}^{T} \eta_t^4 = \eta^4 \sum_{t=1}^{T} \frac{1}{t^2} \leq 2\eta^4 \,, \qquad \text{and} \qquad \sum_{t=1}^{T} \eta_t^p = \eta^p \sum_{t=1}^{T} \frac{1}{t^{p/2}} \leq 2\eta^p$$

as $p > 4$ and $t \geq 1$, Theorem 1 implies that

$$\begin{aligned}
\frac{\eta \sqrt{T}}{3}(f(\bar{x}_T) - f^*) &\leq B_\psi(x^*, x_1) + \eta^2 (G^2 + \phi^2) \log(eT) + 4\eta^2 \phi^2 \Big( 2 \log(eT) \log(2/\delta) + (2 + (p/3))^2 (4/\delta)^{2/p} \Big) \\
&\qquad + 4\eta^2 \phi^2 \Big( \sqrt{\log(2/\delta)} + (2 + (p/6))(4/\delta)^{2/p} \Big) \\
&\leq B_\psi(x^*, x_1) + \eta^2 (G^2 + \phi^2) \log(eT) + 8\eta^2 \phi^2 \Big( 2 \log(2/\delta) + (2 + (p/3))^2 (4/\delta)^{2/p} \Big) \log(eT) \\
&\leq B_\psi(x^*, x_1) + \eta^2 (G^2 + \phi^2) \log(eT) + 8\eta^2 \phi^2 \Big( p(2/\delta)^{2/p} + (2 + (p/3))^2 (4/\delta)^{2/p} \Big) \log(eT) \,,
\end{aligned}$$

where we have used that $\log(2/\delta) \leq (p/2)(2/\delta)^{2/p}$, and once again used $\phi^2$ in place of $\sigma^2$ by virtue of Assumption 5. Hence, we conclude that

$$\begin{aligned}
f(\bar{x}_T) - f^* &\leq \frac{3 \log(eT)}{\sqrt{T}} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \Big( G^2 + \phi^2 + 16(2 + p)^2 \phi^2 (4/\delta)^{2/p} \Big) \right) \\
&\leq \frac{C \log(eT)}{\sqrt{T}} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \Big( G^2 + \phi^2 (1/\delta)^{2/p} \Big) \right) .
\end{aligned}$$

□

# C BOUNDS FOR THE AVERAGE ITERATE UNDER A BOUNDED DOMAIN ASSUMPTION

In this section, we consider again the case when $\eta_t = \eta/\sqrt{t}$ and prove, under a bounded domain assumption, error bounds for the average iterate that assume a two-regime form. We start with following standard error bound.

**Lemma 9.** *Assume that there exits $D > 0$ such that $\sqrt{B_\psi(x,y)} \le D$ for any $(x,y) \in \mathrm{dom}(\psi) \times \mathrm{int}(\mathrm{dom}(\psi))$. Then, under Assumptions 1–3, Algorithm 1 satisfies*

$$f(\bar{x}_T) - f^* \le \frac{1}{T}\left(\frac{D^2}{\eta_T} + \sum_{t=1}^T \eta_t(G^2+\sigma^2) + \sum_{t=1}^T \langle \xi_t, x_t - x^*\rangle + \sum_{t=1}^T \eta_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2)\right).$$

*Proof.* Lemma 4 with $z = x^*$ yields that

$$f(x_t) - f^* \le \frac{1}{\eta_t}B_\psi(x^*, x_t) - \frac{1}{\eta_t}B_\psi(x^*, x_{t+1}) + \langle \xi_t, x_t - x^*\rangle + \frac{\eta_t}{2}\|\hat{g}_t\|_*^2.$$

Summing this inequality we obtain that

$$\begin{aligned}
\sum_{t=1}^T (f(x_t) - f^*) &\le \sum_{t=1}^T \frac{1}{\eta_t}B_\psi(x^*, x_t) - \sum_{t=1}^T \frac{1}{\eta_t}B_\psi(x^*, x_{t+1}) + \sum_{t=1}^T \langle \xi_t, x_t - x^*\rangle + \frac{1}{2}\sum_{t=1}^T \eta_t\|\hat{g}_t\|_*^2 \\
&= \frac{1}{\eta_1}B_\psi(x^*, x_1) - \frac{1}{\eta_T}B_\psi(x^*, x_{T+1}) + \sum_{t=2}^T B_\psi(x^*, x_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \\
&\qquad + \sum_{t=1}^T \langle \xi_t, x_t - x^*\rangle + \frac{1}{2}\sum_{t=1}^T \eta_t\|\hat{g}_t\|_*^2 \\
&\le \frac{D^2}{\eta_1} + D^2\sum_{t=2}^T\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \sum_{t=1}^T \langle \xi_t, x_t - x^*\rangle + \frac{1}{2}\sum_{t=1}^T \eta_t\|\hat{g}_t\|_*^2 \\
&= \frac{D^2}{\eta_T} + \sum_{t=1}^T \langle \xi_t, x_t - x^*\rangle + \frac{1}{2}\sum_{t=1}^T \eta_t\|\hat{g}_t\|_*^2.
\end{aligned}$$

The required result then follows using the fact that $f(\bar{x}_T) - f^* \le \frac{1}{T}\sum_{t=1}^T (f(x_t) - f^*)$ and that

$$\|\hat{g}_t\|_*^2 = \|g_t - \xi_t\|_*^2 \le 2(\|g_t\|_*^2 + \|\xi_t\|_*^2) = 2(\|g_t\|_*^2 + \mathbb{E}_t\|\xi_t\|_*^2) + 2(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2) \le 2(G^2+\sigma^2) + 2(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2),$$

where we used Assumptions 2 and 3 in the last step. □

We then state the two following corollaries specializing the result of the last lemma under Assumptions 4 and 5 respectively.

**Corollary 4.** *Assume that there exits $D > 0$ such that $\sqrt{B_\psi(x,y)} \le D$ for any $(x,y) \in \mathrm{dom}(\psi) \times \mathrm{int}(\mathrm{dom}(\psi))$. Then, for any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies, under Assumptions 1, 2 and 4, that with probability at least $1-\delta$,*

$$\begin{aligned}
f(\bar{x}_T) - f^* &\le \frac{C_1}{\sqrt{T}}\left(\frac{D^2}{\eta} + \eta G^2 + \phi D\left(\sqrt{\log(e/\delta)} + \frac{\log^\theta(eT/\delta)}{\sqrt{T}}\right) + \eta\phi^2\left(1 + \sqrt{\frac{\log(eT)\log(e/\delta)}{T}} + \frac{\log^{2\theta}(e/\delta)}{\sqrt{T}}\right)\right) \\
&\le \frac{C_2}{\sqrt{T}}\left(\frac{D^2}{\eta} + \eta G^2 + \eta\phi^2\left(\log(e/\delta) + \frac{\log^{2\theta}(e/\delta)}{\sqrt{T}} + \frac{\log^{2\theta}(eT/\delta)}{T}\right)\right),
\end{aligned}$$

*where $C_1$ and $C_2$ are constants depending only on $\theta$.*

*Proof.* For $t \in [T]$, let $u_t = (x_t - x^*)/(\sqrt{2}D)$, while for $k \in [T]$, we define

$$W_k = \sum_{t=1}^k \langle \xi_t, u_t\rangle \qquad \text{and} \qquad V_k = \sum_{t=1}^k \eta_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2).$$

Since $\sqrt{2}D \geq \sqrt{2B_\psi(x^*, x_t)} \geq \|x_t - x^*\|$, it holds that $\|u_t\| \leq 1$. For what follows, we will use $C, C_1, C_2, \ldots$ to denote positive constants—depending only on $\theta$—whose values may change between steps. For the first martingale $(W_k)$, we invoke Lemma 7(i) with $s = 0$ and $\omega_t = 1$ obtaining that

$$P\left(\max_{k\in[T]} W_k \geq C_1\phi\sqrt{T\log(2/\delta)} + C_2\phi\log^\theta(2eT/\delta)\right) \leq \delta\,.$$

while for $(V_k)$, we use the Lemma 7(ii) with $s = 3$ and $\omega_t = \eta/\sqrt{t}$ yielding that

$$P\left(\max_{k\in[T]} V_k \geq C_1\eta\phi^2\sqrt{\sum_{t=1}^{T}(1/t)\log(2/\delta)} + C_2\eta\phi^2\log^{2\theta}\left(2e\sum_{t=1}^{T}(1/t)^{3/2}/\delta\right)\right) \leq \delta\,.$$

Since $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, $\sum_{t=1}^{T}(1/t) \leq \log(eT)$, and $\sum_{t=1}^{T}(1/t)^{3/2} \leq 3$, Lemma 9 implies via a union bound that with probability at least $1-\delta$,

$$\begin{aligned}T(f(\bar{x}_T) - f^*) \leq \frac{D^2\sqrt{T}}{\eta} + C_1\eta(G^2+\phi^2)\sqrt{T} + C_2\phi D\Big(\sqrt{T\log(4/\delta)} + \log^\theta(4eT/\delta)\Big)\\ + C_3\eta\phi^2\Big(\sqrt{\log(eT)\log(4/\delta)} + \log^{2\theta}(12e/\delta)\Big)\,,\end{aligned}$$

where we have used that Assumption 4 implies Assumption 3 with $\sigma^2 = 2\Gamma(2\theta+1)\phi^2$ thanks to Lemma 12. This proves the first inequality in the statement. Going further, we can use the fact that $2ab = \inf_{r>0} a^2/r + rb^2$ for any $a, b > 0$ to get that

$$\begin{aligned}2\phi D\Big(\sqrt{T\log(4/\delta)} + \log^\theta(4eT/\delta)\Big) &\leq \frac{D^2}{\eta_T} + \eta_T\phi^2\Big(\sqrt{T\log(4/\delta)} + \log^\theta(4eT/\delta)\Big)^2\\ &\leq \frac{D^2\sqrt{T}}{\eta} + \frac{2\eta}{\sqrt{T}}\phi^2\Big(T\log(4/\delta) + \log^{2\theta}(4eT/\delta)\Big)\,,\end{aligned}$$

implying that

$$\begin{aligned}T(f(\bar{x}_T) - f^*) &\leq C_1\frac{D^2\sqrt{T}}{\eta} + C_2\eta(G^2+\phi^2)\sqrt{T} + C_3\eta\phi^2\left(\sqrt{T}\log(4/\delta) + \frac{\log^{2\theta}(4eT/\delta)}{\sqrt{T}}\right)\\ &\quad + C_4\eta\phi^2\Big(\sqrt{\log(eT)\log(4/\delta)} + \log^{2\theta}(12e/\delta)\Big)\\ &\leq C_1\frac{D^2\sqrt{T}}{\eta} + C_2\eta(G^2+\phi^2)\sqrt{T} + C_3\eta\phi^2\left(\sqrt{T}\log(4/\delta) + \frac{\log^{2\theta}(4eT/\delta)}{\sqrt{T}} + \log^{2\theta}(12e/\delta)\right).\end{aligned}$$

□

**Corollary 5.** *Assume that there exits $D > 0$ such that $\sqrt{B_\psi(x,y)} \leq D$ for any $(x,y) \in \mathrm{dom}(\psi) \times \mathrm{int}(\mathrm{dom}(\psi))$. Then, for any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies, under Assumptions 1, 2 and 5, that with probability at least $1-\delta$,*

$$\begin{aligned}f(\bar{x}_T) - f^* &\leq \frac{C_1}{\sqrt{T}}\left(\frac{D^2}{\eta} + \eta G^2 + \phi D\left(\sqrt{\log(e/\delta)} + \frac{(1/\delta)^{1/p}}{T^{1/2-1/p}}\right) + \eta\phi^2\left(1 + \sqrt{\frac{\log(eT)\log(e/\delta)}{T}} + \frac{(1/\delta)^{2/p}}{\sqrt{T}}\right)\right)\\ &\leq \frac{C_2}{\sqrt{T}}\left(\frac{D^2}{\eta} + \eta G^2 + \eta\phi^2\left(\log(e/\delta) + \frac{(1/\delta)^{2/p}}{\sqrt{T}}\right)\right),\end{aligned}$$

*where $C_1$ and $C_2$ are constants depending only on $p$.*

*Proof.* Similar to the proof of Corollary 4, we define $u_t = (x_t - x^*)/(\sqrt{2}D)$ (which satisfies $\|u_t\| \leq 1$), and consider once again the two martingale terms

$$W_k = \sum_{t=1}^{k}\langle\xi_t, u_t\rangle \qquad \text{and} \qquad V_k = \sum_{t=1}^{k}\eta_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2)\,.$$

For $(W_k)$, we use Lemma 8(i) with $\omega_t = 1$, while for $(V_k)$, we use the Lemma 8(ii) with $\omega_t = \eta/\sqrt{t}$ yielding that

$$P\left(\max_{k\in[T]} W_k > \phi\sqrt{2T\log(1/\delta)} + (2+(p/3))\phi\,(T/\delta)^{1/p}\right) \le \delta\,,$$

and

$$P\left(\max_{k\in[T]} V_k > 2\eta\phi^2\sqrt{2\sum_{t=1}^{T}(1/t)\log(1/\delta)} + 2(2+(p/6))\eta\phi^2\left(\sum_{t=1}^{T}(1/t)^{p/4}/\delta\right)^{2/p}\right) \le \delta\,.$$

Since $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$, $\sum_{t=1}^{T}(1/t) \le \log(eT)$, and $\sum_{t=1}^{T}(1/t)^{p/4} \le p/(p-4)$, Lemma 9 implies via a union bound that with probability at least $1-\delta$,

$$\begin{aligned} T(f(\bar{x}_T) - f^*) \le \frac{D^2\sqrt{T}}{\eta} + 2\eta(G^2+\phi^2)\sqrt{T} + \sqrt{2}\phi D\Big(\sqrt{2T\log(2/\delta)} + (2+(p/3))\,(2T/\delta)^{1/p}\Big) \\ + 2\eta\phi^2\Big(\sqrt{2\log(eT)\log(2/\delta)} + (2+(p/6))(p/(p-4))^{2/p}\,(2/\delta)^{2/p}\Big)\,, \end{aligned}$$

where we have used that Assumption 5 implies Assumption 3 with $\sigma^2 = \phi^2$. This proves the first inequality in the corollary's statement. For the second, we use once again that $2ab = \inf_{r>0} a^2/r + rb^2$ for any $a, b > 0$, which implies that

$$\begin{aligned} \sqrt{2}\phi D\Big(\sqrt{2T\log(2/\delta)} + (2+(p/3))\,(2T/\delta)^{1/p}\Big) &\le \frac{D^2}{\eta_T} + \frac{1}{2}\eta_T\phi^2\Big(\sqrt{2T\log(2/\delta)} + (2+(p/3))\,(2T/\delta)^{1/p}\Big)^2 \\ &\le \frac{D^2\sqrt{T}}{\eta} + \frac{\eta}{\sqrt{T}}\phi^2\Big(2T\log(2/\delta) + (2+(p/3))^2\,(2T/\delta)^{2/p}\Big)\,, \end{aligned}$$

using which we obtain that

$$\begin{aligned} T(f(\bar{x}_T) - f^*) &\le \frac{2D^2\sqrt{T}}{\eta} + 2\eta(G^2+\phi^2)\sqrt{T} + \eta\phi^2\Big(2\sqrt{T}\log(2/\delta) + (2+(p/3))^2 T^{(4-p)/(2p)}\,(2/\delta)^{2/p}\Big) \\ &\quad + 2\eta\phi^2\Big(\sqrt{2\log(eT)\log(2/\delta)} + (2+(p/6))(p/(p-4))^{2/p}\,(2/\delta)^{2/p}\Big) \\ &\le \frac{2D^2\sqrt{T}}{\eta} + 2\eta(G^2+\phi^2)\sqrt{T} + 3\eta\phi^2\Big(2\sqrt{T}\log(2/\delta) + (2+(p/3))^2(p/(p-4))^{2/p}\,(2/\delta)^{2/p}\Big)\,, \end{aligned}$$

where in the second step we used that $p > 4$. □

# D PROOFS OF SECTION 5

## D.1 Proof of Proposition 1

**Proposition 1.** *Let $(M_t)_{t=0}^n$ be a square integrable martingale adapted to filtration $(\mathcal{F}_t)_{t=0}^n$ with $M_0 = 0$. Then, for all $x, \beta > 0$ and $\alpha \ge 0$,*

$$P\left(\bigcup_{t=1}^{n}\{M_t \ge x \text{ and } \langle M\rangle_t + [M]_t \le \alpha M_t + \beta\}\right) \le \exp\left(-\min\left\{\frac{x^2}{8\beta}, \frac{x}{6\alpha}\right\}\right).$$

*Proof.* For any $\lambda \in \mathbb{R}$ and $0 \le t \le n$, define[6]

$$V_t(\lambda) = \exp\left(\lambda M_t - \frac{\lambda^2}{2}(\langle M\rangle_t + [M]_t)\right).$$

By Lemma B.1 in (Bercu and Touati, 2008), $(V_t(\lambda))_{t=0}^n$ is a (non-negative) supermartingale (with $V_0(\lambda) = 1$). For $t \in [n]$, define the event $A_t = \{M_t \ge x \text{ and } \langle M\rangle_t + [M]_t \le \alpha M_t + \beta\}$. From the proof of Theorem 3.3 in (Harvey

[6] One can set $\langle M\rangle_0 = [M]_0 = M_0$.

et al., 2019), if we fix some $\lambda \in (0, 1/(2\alpha))$, then there exists $c = c(\lambda, \alpha) \in (0, 2]$ such that $(\lambda + c\lambda^2\alpha)^2 = 2c\lambda^2$. With this in mind, we have that for any $t \in [n]$ and any $\lambda \in (0, 1/(2\alpha))$:

$$\begin{aligned}
\mathbb{I}\{A_t\} &\leq \exp\Big((\lambda + c\lambda^2\alpha)M_t - c\lambda^2(\langle M\rangle_t + [M]_t) - \lambda x + c\lambda^2\beta\Big) \\
&= \exp(-\lambda x + c\lambda^2\beta)\exp\Big((\lambda + c\lambda^2\alpha)M_t - c\lambda^2(\langle M\rangle_t + [M]_t)\Big) \\
&= \exp(-\lambda x + c\lambda^2\beta)\exp\Big(\tilde{\lambda}M_t - \frac{\tilde{\lambda}^2}{2}(\langle M\rangle_t + [M]_t)\Big) \\
&= \exp(-\lambda x + c\lambda^2\beta)V_t(\tilde{\lambda}) \leq \exp(-\lambda x + 2\lambda^2\beta)V_t(\tilde{\lambda})\,,
\end{aligned}$$

where $\tilde{\lambda} = \lambda + c\lambda^2\alpha$, and the first inequality holds since the argument of the exponent is non-negative under $A_t$. Hence, Lemma 16 entails that

$$P\left(\bigcup_{t=1}^{n} A_t\right) \leq \exp(-\lambda x + 2\lambda^2\beta)\,.$$

Finally, upon choosing $\lambda = \min\{\frac{x}{4\beta}, \frac{1}{3\alpha}\}$, we can conclude that

$$\exp(-\lambda x + 2\lambda^2\beta) \leq \exp\left(-\min\left\{\frac{x^2}{8\beta}, \frac{x}{6\alpha}\right\}\right).$$

□

### D.2 Proof of Lemma 1

**Lemma 1.** *Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ for some constant $\eta > 0$ satisfies*

$$f(x_T) - f^* \leq \frac{2}{T}\sum_{t=\lceil T/2\rceil}^{T}(f(x_t) - f^*) + \sum_{t=\lceil T/2\rceil}^{T}\langle \xi_t, w_t\rangle + \frac{\eta}{\sqrt{2T}}\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2 + \frac{\sqrt{2}}{\eta\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T} z_t\,,$$

*where, for $j < T$, $\alpha_j = \frac{1}{(T-j)(T-j+1)}$, and for any time-step $t \geq \lceil T/2\rceil$,*

$$w_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j(x_t - x_j)\,, \quad z_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j B_\psi(x_j, x_t)\,, \quad \textit{and} \quad \rho_t = \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j\,.$$

*Proof.* For any $k \in [T-1]$, Lemma 6 with $j = T-k$ and $r = T$ implies that

$$\sum_{t=T-k}^{T}(f(x_t) - f(x_{T-k})) \leq \sum_{t=T-k}^{T}\langle \xi_t, x_t - x_{T-k}\rangle + \frac{1}{2}\sum_{t=T-k}^{T}\eta_t\|\hat{g}_t\|_*^2 + \sum_{t=T-k}^{T}\tilde{\eta}_t B_\psi(x_{T-k}, x_t)\,,$$

where $\tilde{\eta}_t = 1/\eta_t - 1/\eta_{t-1}$ and $\tilde{\eta}_1 = 1/\eta_1$. We then proceed as in the proof of Lemma 7.1 in (Harvey et al., 2019). Namely, we define $S_k = \frac{1}{k+1}\sum_{t=T-k}^{T} f(x_t)$, which, combined with the previous inequality, yields that

$$\begin{aligned}
S_{k-1} &= S_k + \frac{S_k - f(x_{T-k})}{k} \\
&\leq S_k + \frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle \xi_t, x_t - x_{T-k}\rangle + \frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\|\hat{g}_t\|_*^2 + \frac{1}{k(k+1)}\sum_{t=T-k}^{T}\tilde{\eta}_t B_\psi(x_{T-k}, x_t)\,.
\end{aligned}$$

Since $S_0 = f(x_T)$, by unrolling the recursion we obtain that

$$\begin{aligned}
f(x_T) \leq \frac{1}{\lfloor T/2\rfloor + 1}\sum_{t=\lceil T/2\rceil}^{T} f(x_t) + \sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle \xi_t, x_t - x_{T-k}\rangle + \sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\|\hat{g}_t\|_*^2 \\
+ \sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\tilde{\eta}_t B_\psi(x_{T-k}, x_t)\,. \quad (5)
\end{aligned}$$

One can rewrite the second term on the right-hand side of the above inequality as follows

$$\sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle\xi_t, x_t - x_{T-k}\rangle = \sum_{t=\lceil T/2\rceil}^{T}\sum_{k=(T-t)\vee 1}^{\lfloor T/2\rfloor}\frac{1}{k(k+1)}\langle\xi_t, x_t - x_{T-k}\rangle$$
$$= \sum_{t=\lceil T/2\rceil}^{T}\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\frac{1}{(T-j)(T-j+1)}\langle\xi_t, x_t - x_j\rangle = \sum_{t=\lceil T/2\rceil}^{T}\langle\xi_t, w_t\rangle\,.$$

Similarly, we also have that

$$\sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\|\hat{g}_t\|_*^2 = \frac{1}{2}\sum_{t=\lceil T/2\rceil}^{T}\eta_t\|\hat{g}_t\|_*^2\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\frac{1}{(T-j)(T-j+1)} = \frac{1}{2}\sum_{t=\lceil T/2\rceil}^{T}\eta_t\rho_t\|\hat{g}_t\|_*^2$$
$$\sum_{k=1}^{\lfloor T/2\rfloor}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\tilde{\eta}_t B_\psi(x_{T-k}, x_t) = \sum_{t=\lceil T/2\rceil}^{T}\tilde{\eta}_t\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\frac{1}{(T-j)(T-j+1)}B_\psi(x_j, x_t) = \sum_{t=\lceil T/2\rceil}^{T}\tilde{\eta}_t z_t\,.$$

After plugging these expressions back into (5), we conclude the proof by using that $\lfloor T/2\rfloor+1 \geq T/2$ and observing that for any time-step $t \geq \lceil T/2\rceil$,

$$\eta_t = \frac{\eta}{\sqrt{t}} \leq \frac{\sqrt{2}\eta}{\sqrt{T}} \qquad \text{and} \qquad \tilde{\eta}_t = \frac{1}{\eta}(\sqrt{t}-\sqrt{t-1}) = \frac{1}{\eta(\sqrt{t}+\sqrt{t-1})} \leq \frac{\sqrt{2}}{\eta\sqrt{T}}\,.$$

□

### D.3 Proof of Lemma 2

Recall that for a time-step $s$ such that $\lceil T/2\rceil \leq s \leq T$, $Q_s = \sum_{t=\lceil T/2\rceil}^{s}\langle\xi_t, w_t\rangle$, and that $z^*$ is short for $\max_{\lceil T/2\rceil\leq s\leq T} z_s$.

**Lemma 2.** *In the same setting as Lemma 1, it holds that*

$$z^* \leq \frac{6\sqrt{2}\eta}{T\sqrt{T}}\sum\nolimits_{t=\lceil T/2\rceil}^{T}(f(x_t)-f^*) + \frac{3\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{3\eta^2}{T}\sum\nolimits_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2\,,$$

*where* $n^* = \min\{n\colon n \in \arg\max_{\lceil T/2\rceil\leq s\leq T} Q_s\}$.

*Proof.* Notice that $z_{\lceil T/2\rceil} = 0$ and $Q_{\lceil T/2\rceil} = 0$; hence, the lemma trivially holds when $T = 1$. Thus, we assume for what follows that $T \geq 2$. Let $j$ and $s$ be two time-steps such that $\lceil T/2\rceil + 1 \leq s \leq T$ and $\lceil T/2\rceil \leq j \leq s$. Then, via Lemma 6, we have that

$$\frac{1}{\eta_{s-1}}B_\psi(x_j, x_s) \leq \sum_{t=j}^{s-1}(f(x_j)-f(x_t)) + \sum_{t=j}^{s-1}\langle\xi_t, x_t - x_j\rangle + \frac{1}{2}\sum_{t=j}^{s-1}\eta_t\|\hat{g}_t\|_*^2 + \sum_{t=j}^{s-1}\tilde{\eta}_t B_\psi(x_j, x_t)\,,$$

where $\tilde{\eta}_t = 1/\eta_t - 1/\eta_{t-1}$ and $\tilde{\eta}_1 = 1/\eta_1$. This, in turn, implies that

$$\frac{1}{\eta_{s-1}}\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)}\alpha_j B_\psi(x_j, x_s) \leq \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)}\alpha_j\sum_{t=j}^{s-1}(f(x_j)-f(x_t)) + \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)}\alpha_j\sum_{t=j}^{s-1}\langle\xi_t, x_t - x_j\rangle$$
$$+ \frac{1}{2}\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)}\alpha_j\sum_{t=j}^{s-1}\eta_t\|\hat{g}_t\|_*^2 + \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)}\alpha_j\sum_{t=j}^{s-1}\tilde{\eta}_t B_\psi(x_j, x_t)\,. \quad (6)$$

For the last three terms, we swap the sums obtaining that

$$\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j \sum_{t=j}^{s-1} \langle \xi_t, x_t - x_j\rangle = \sum_{t=\lceil T/2\rceil}^{s-1} \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j \langle \xi_t, x_t - x_j\rangle = \sum_{t=\lceil T/2\rceil}^{s-1} \langle \xi_t, w_t\rangle$$

$$\frac{1}{2}\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j \sum_{t=j}^{s-1} \eta_t \|\hat{g}_t\|_*^2 = \frac{1}{2}\sum_{t=\lceil T/2\rceil}^{s-1} \eta_t \|\hat{g}_t\|_*^2 \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j = \frac{1}{2}\sum_{t=\lceil T/2\rceil}^{s-1} \eta_t \rho_t \|\hat{g}_t\|_*^2 \leq \frac{\eta}{\sqrt{2T}} \sum_{t=\lceil T/2\rceil}^{s-1} \rho_t \|\hat{g}_t\|_*^2$$

$$\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j \sum_{t=j}^{s-1} \tilde{\eta}_t B_\psi(x_j, x_t) = \sum_{t=\lceil T/2\rceil}^{s-1} \tilde{\eta}_t \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j B_\psi(x_j, x_t) = \sum_{t=\lceil T/2\rceil}^{s-1} \tilde{\eta}_t z_t\,.$$

For the first term, if we define $\Delta_t = f(x_t) - f^*$ for $t \in [T]$, we obtain that

$$\begin{aligned}
\sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j \sum_{t=j}^{s-1} (f(x_j) - f(x_t)) &= \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j \sum_{t=j}^{s-1} (\Delta_j - \Delta_t) \\
&= \sum_{j=\lceil T/2\rceil}^{s-1} \alpha_j \Delta_j (s-j) - \sum_{j=\lceil T/2\rceil}^{s-1} \alpha_j \sum_{t=j}^{s-1} \Delta_t \\
&= \sum_{t=\lceil T/2\rceil}^{s-1} \alpha_t \Delta_t (s-t) - \sum_{t=\lceil T/2\rceil}^{s-1} \Delta_t \sum_{j=\lceil T/2\rceil}^{t} \alpha_j \\
&= \sum_{t=\lceil T/2\rceil}^{s-1} \Delta_t \left( \frac{s-t}{(T-t)(T-t+1)} - \frac{1}{T-t} + \frac{1}{T - \lceil T/2\rceil + 1} \right) \\
&\leq \frac{1}{\lfloor T/2\rfloor + 1} \sum_{t=\lceil T/2\rceil}^{s-1} \Delta_t \leq \frac{2}{T} \sum_{t=\lceil T/2\rceil}^{s-1} \Delta_t\,,
\end{aligned}$$

where in the second equality we used that the inner sum is empty when $j = s$ and that $s \leq T$, the fourth equality follows from Lemma 10 and the definition of $\alpha_t$, and the inequality holds since $(s-t)/(T-t+1) < 1$. Returning back to Equation (6), we have that

$$\begin{aligned}
z_s &= \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j B_\psi(x_j, x_s) \leq \frac{\eta_{\lceil T/2\rceil}}{\eta_{s-1}} \sum_{j=\lceil T/2\rceil}^{s\wedge(T-1)} \alpha_j B_\psi(x_j, x_s) \\
&\leq \eta_{\lceil T/2\rceil} \left( \frac{2}{T} \sum_{t=\lceil T/2\rceil}^{s-1} (f(x_t) - f^*) + \sum_{t=\lceil T/2\rceil}^{s-1} \langle \xi_t, w_t\rangle + \frac{\eta}{\sqrt{2T}} \sum_{t=\lceil T/2\rceil}^{s-1} \rho_t \|\hat{g}_t\|_*^2 + \sum_{t=\lceil T/2\rceil}^{s-1} \tilde{\eta}_t z_t \right) \\
&\leq \frac{2\sqrt{2}\eta}{T\sqrt{T}} \sum_{t=\lceil T/2\rceil}^{s-1} (f(x_t) - f^*) + \frac{\sqrt{2}\eta}{\sqrt{T}} \sum_{t=\lceil T/2\rceil}^{s-1} \langle \xi_t, w_t\rangle + \frac{\eta^2}{T} \sum_{t=\lceil T/2\rceil}^{s-1} \rho_t \|\hat{g}_t\|_*^2 + \frac{\sqrt{2}\eta}{\sqrt{T}} \sum_{t=\lceil T/2\rceil}^{s-1} \tilde{\eta}_t z_t\,.
\end{aligned}$$

Notice that the terms in the first, third and fourth sum on the right-hand side of the last inequality are non-negative. Hence, it holds that

$$z^* \leq \frac{2\sqrt{2}\eta}{T\sqrt{T}} \sum_{t=\lceil T/2\rceil}^{T-1} (f(x_t) - f^*) + \frac{\sqrt{2}\eta}{\sqrt{T}} \max_{\lceil T/2\rceil \leq n \leq T-1} \sum_{t=\lceil T/2\rceil}^{n} \langle \xi_t, w_t\rangle + \frac{\eta^2}{T} \sum_{t=\lceil T/2\rceil}^{T-1} \rho_t \|\hat{g}_t\|_*^2 + \frac{\sqrt{2}\eta}{\sqrt{T}} \sum_{t=\lceil T/2\rceil}^{T-1} \tilde{\eta}_t z_t\,.$$

Next, we will bound the last term by relating it back to $z^*$. Since this term is zero when $T = 2$ (recalling that

$z_{\lceil T/2\rceil}=0$), we focus in the following argument on the case when $T\geq 3$. Observe that

$$\begin{aligned}\frac{\sqrt{2}\eta}{\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T-1}\tilde{\eta}_t &= \frac{\sqrt{2}\eta}{\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T-1}\frac{\sqrt{t}-\sqrt{t-1}}{\eta}\\ &= \frac{\sqrt{2}}{\sqrt{T}}(\sqrt{T-1}-\sqrt{\lceil T/2\rceil-1})\\ &\leq \frac{\sqrt{2}}{\sqrt{T}}(\sqrt{T-1}-\sqrt{T/2-1})\end{aligned}$$

As a function of $T$, the last expression is decreasing in $T\geq 3$, and thus (by plugging in $T=3$) can be bounded by $1/\sqrt{3}$ . Hence,

$$\frac{\sqrt{2}\eta}{\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T-1}\tilde{\eta}_t z_t \leq \frac{1}{\sqrt{3}}z^* \leq \frac{2}{3}z^*\,.$$

Consequently,

$$\begin{aligned}z^* &\leq \frac{6\sqrt{2}\eta}{T\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T-1}(f(x_t)-f^*)+\frac{3\sqrt{2}\eta}{\sqrt{T}}\max_{\lceil T/2\rceil\leq n\leq T-1}\sum_{t=\lceil T/2\rceil}^{n}\langle\xi_t,w_t\rangle+\frac{3\eta^2}{T}\sum_{t=\lceil T/2\rceil}^{T-1}\rho_t\|\hat{g}_t\|_*^2\\ &\leq \frac{6\sqrt{2}\eta}{T\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T}(f(x_t)-f^*)+\frac{3\sqrt{2}\eta}{\sqrt{T}}Q_{n^*}+\frac{3\eta^2}{T}\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2\,.\end{aligned}$$

□

### D.4 Proof of Lemma 3

**Lemma 3.** *In the same setting as Lemma 1, it holds under Assumption 3 that*

$$\langle Q\rangle_T+[Q]_T\leq 4\sigma^2 z^*\log(4T)+2z^*\sum\nolimits_{t=\lceil T/2\rceil}^{T}\rho_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\,.$$

*Proof.* Recall that $w_t=\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j(x_t-x_j)$, $z_t=\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j B_\psi(x_j,x_t)$, and $\rho_t=\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j$ for time-step $t\geq\lceil T/2\rceil$, and observe that

$$\|w_t\|^2=\rho_t^2\left\|\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\frac{\alpha_j}{\rho_t}(x_t-x_j)\right\|^2\leq\rho_t^2\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\frac{\alpha_j}{\rho_t}\|x_t-x_j\|^2\leq 2\rho_t\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)}\alpha_j B_\psi(x_j,x_t)=2\rho_t z_t\,,\tag{7}$$

where the first inequality holds via the convexity of $\|\cdot\|^2$, and the second follows from the fact that $\|x_t-x_j\|^2\leq$

$2B_\psi(x_j, x_t)$ as $\psi$ is 1-strongly convex. Hence,

$$\begin{aligned}
\langle Q\rangle_T + [Q]_T &= \sum_{t=\lceil T/2\rceil}^{T} \Big(\mathbb{E}_t[|\langle \xi_t, w_t\rangle|^2] + |\langle \xi_t, w_t\rangle|^2\Big) \\
&\le \sum_{t=\lceil T/2\rceil}^{T} \Big(\mathbb{E}_t[\|\xi_t\|_*^2\|w_t\|^2] + \|\xi_t\|_*^2\|w_t\|^2\Big) \\
&= \sum_{t=\lceil T/2\rceil}^{T} \|w_t\|^2\Big(\mathbb{E}_t[\|\xi_t\|_*^2] + \|\xi_t\|_*^2\Big) \\
&\le 2z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t\Big(\mathbb{E}_t[\|\xi_t\|_*^2] + \|\xi_t\|_*^2\Big) \\
&= 4z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t\mathbb{E}_t[\|\xi_t\|_*^2] + 2z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t\Big(\|\xi_t\|_*^2 - \mathbb{E}_t[\|\xi_t\|_*^2]\Big) \\
&\le 4\sigma^2 z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t + 2z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t\Big(\|\xi_t\|_*^2 - \mathbb{E}_t[\|\xi_t\|_*^2]\Big) \\
&\le 4\sigma^2 z^* \log(4T) + 2z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t\Big(\|\xi_t\|_*^2 - \mathbb{E}_t[\|\xi_t\|_*^2]\Big)\,,
\end{aligned}$$

where the first inequality follows from the definition of the dual norm, the second equality holds since $w_t$ is $\mathcal{F}_{t-1}$-measurable, the second inequality follows from (7) and the definition of $z^*$, the third inequality follows from Assumption 3, and the last inequality is an application of Lemma 11. □

### D.5 Proof of Theorem 2

**Theorem 2.** *Let $\Xi_1, \Xi_2 : (0,1) \to (0,\infty)$ be two functions such that for any $\delta \in (0,1)$,*

$$P\left(\frac{1}{\sqrt{T}}\sum\nolimits_{t=1}^{T}(f(x_t) - f^*) > \Xi_1(\delta)\right) \le \delta$$

*and*

$$P\left(\sum\nolimits_{t=\lceil T/2\rceil}^{T} \rho_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2) > \Xi_2(\delta)\right) \le \delta\,.$$

*Then, under Assumptions 1–3, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies the following with probability at least $1-\delta$:*

$$f(x_T) - f^* \le \frac{35}{\sqrt{T}}\left(2\Xi_1(\delta/3) + \sqrt{2}\eta G^2\log(4T) + 9\sqrt{2}\eta\Big(\Xi_2(\delta/3) + 2\sigma^2\log(4T)\Big)\log(3/\delta)\right).$$

*Proof.* From Lemma 1, we have that

$$f(x_T) - f^* \le \frac{2}{T}\sum_{t=\lceil T/2\rceil}^{T}(f(x_t) - f^*) + Q_T + \frac{\eta}{\sqrt{2T}}\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2 + \frac{\sqrt{2}}{\eta\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T} z_t\,.$$

Notice that

$$\begin{aligned}
\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2 &= \sum_{t=\lceil T/2\rceil}^{T}\rho_t\|g_t-\xi_t\|_*^2\\
&\le 2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|g_t\|_*^2+\|\xi_t\|_*^2)\\
&= 2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|g_t\|_*^2+\mathbb{E}_t\|\xi_t\|_*^2)+2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\\
&\le 2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(G^2+\sigma^2)+2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\\
&\le 2(G^2+\sigma^2)\log(4T)+2\sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\,,
\end{aligned}$$

where the second inequality follows from Assumptions 2 and 3, and the third inequality follows from Lemma 11. For what follows, define

$$\begin{aligned}
\Lambda_1 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f(x_t)-f^*)\\
\Lambda_2 &= \sum_{t=\lceil T/2\rceil}^{T}\rho_t(\|\xi_t\|_*^2-\mathbb{E}_t\|\xi_t\|_*^2)\,.
\end{aligned}$$

From the assumption in the theorem's statement, we have that for any $\delta\in(0,1)$,

$$P(\Lambda_1>\Xi_1(\delta))\le\delta \qquad\text{and}\qquad P(\Lambda_2>\Xi_2(\delta))\le\delta\,. \tag{8}$$

Additionally, define $\Xi_3=(G^2+\sigma^2)\log(4T)$. Subsequently, it holds that

$$f(x_T)-f^*\le\frac{2}{\sqrt{T}}\Lambda_1+Q_{n^*}+\frac{\sqrt{2}\eta}{\sqrt{T}}(\Lambda_2+\Xi_3)+\frac{\sqrt{2}}{\eta\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T}z_t\,. \tag{9}$$

On the other hand, we have via Lemma 2 that

$$\begin{aligned}
z^*=\max_{\lceil T/2\rceil\le s\le T}z_s &\le \frac{6\sqrt{2}\eta}{T\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T}(f(x_t)-f^*)+\frac{3\sqrt{2}\eta}{\sqrt{T}}Q_{n^*}+\frac{3\eta^2}{T}\sum_{t=\lceil T/2\rceil}^{T}\rho_t\|\hat{g}_t\|_*^2\\
&\le \frac{6\sqrt{2}\eta}{T}\Lambda_1+\frac{3\sqrt{2}\eta}{\sqrt{T}}Q_{n^*}+\frac{6\eta^2}{T}(\Lambda_2+\Xi_3)\,.
\end{aligned} \tag{10}$$

Hence,

$$\begin{aligned}
\frac{\sqrt{2}}{\eta\sqrt{T}}\sum_{t=\lceil T/2\rceil}^{T}z_t &\le \frac{\sqrt{2}T}{\eta\sqrt{T}}\left(\frac{6\sqrt{2}\eta}{T}\Lambda_1+\frac{3\sqrt{2}\eta}{\sqrt{T}}Q_{n^*}+\frac{6\eta^2}{T}(\Lambda_2+\Xi_3)\right)\\
&= \frac{12}{\sqrt{T}}\Lambda_1+6Q_{n^*}+\frac{6\sqrt{2}\eta}{\sqrt{T}}(\Lambda_2+\Xi_3)\,.
\end{aligned}$$

Plugging back into (9) yields that

$$f(x_T)-f^*\le\frac{14}{\sqrt{T}}\Lambda_1+7Q_{n^*}+\frac{7\sqrt{2}\eta}{\sqrt{T}}(\Lambda_2+\Xi_3)\,. \tag{11}$$

Our aim in the sequel is to use the above inequality in conjunction with (8) and Proposition 1 to bound the error in high probability. Towards that end, we start with the following upper bound on the TCV and TQV of $Q_{n^*}$, which is implied by Lemma 3 and the fact that the TCV and TQV are non-decreasing.

$$\langle Q\rangle_{n^*} + [Q]_{n^*} \leq \langle Q\rangle_T + [Q]_T \leq 4\sigma^2 z^* \log(4T) + 2z^* \sum_{t=\lceil T/2\rceil}^{T} \rho_t(\|\xi_t\|_*^2 - \mathbb{E}_t\|\xi_t\|_*^2) = 2z^*(2\tilde{\Xi}_3 + \Lambda_2)\,,$$

where $\tilde{\Xi}_3 \coloneqq \sigma^2 \log(4T)$. Moreover, under the event that $\Lambda_1 \leq \Xi_1(\delta)$ and $\Lambda_2 \leq \Xi_2(\delta)$, we have that

$$\frac{12\sqrt{2}\eta}{T}\Lambda_1 + \frac{6\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{12\eta^2}{T}(\Lambda_2 + \Xi_3) \leq \frac{12\sqrt{2}\eta}{T}\Xi_1(\delta) + \frac{6\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{12\eta^2}{T}(\Xi_2(\delta) + \Xi_3) \tag{12}$$

and

$$\Lambda_2 + 2\tilde{\Xi}_3 \leq \Xi_2(\delta) + 2\tilde{\Xi}_3\,, \tag{13}$$

which implies that under the same event,

$$\begin{aligned}
\langle Q\rangle_{n^*} + [Q]_{n^*} &\leq 2z^*(\Lambda_2 + 2\tilde{\Xi}_3) \\
&\leq \left(\frac{12\sqrt{2}\eta}{T}\Lambda_1 + \frac{6\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{12\eta^2}{T}(\Lambda_2 + \Xi_3)\right)(\Lambda_2 + 2\tilde{\Xi}_3) \\
&\leq \left(\frac{12\sqrt{2}\eta}{T}\Lambda_1 + \frac{6\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{12\eta^2}{T}(\Lambda_2 + \Xi_3)\right)(\Xi_2(\delta) + 2\tilde{\Xi}_3) \\
&\leq \left(\frac{12\sqrt{2}\eta}{T}\Xi_1(\delta) + \frac{6\sqrt{2}\eta}{\sqrt{T}}Q_{n^*} + \frac{12\eta^2}{T}(\Xi_2(\delta) + \Xi_3)\right)(\Xi_2(\delta) + 2\tilde{\Xi}_3)\,,
\end{aligned} \tag{14}$$

where the second inequality follows from (10) and the fact that $\Lambda_2 + 2\tilde{\Xi}_3$ is non-negative,[7] the third inequality follows from (13) and the fact that the first bracketed expression on the left-hand side is non-negative as it is an upper bound for the non-negative quantity $2z^*$, whereas the last inequality follows from (12) and the fact that $\Xi_2(\delta) + 2\tilde{\Xi}_3$ is non-negative. As a last bit of notation, we define

$$\begin{aligned}
R_1(\delta) &= \frac{6\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta) + 2\tilde{\Xi}_3) \\
R_2(\delta) &= \left(\frac{12\sqrt{2}\eta}{T}\Xi_1(\delta) + \frac{12\eta^2}{T}(\Xi_2(\delta) + \Xi_3)\right)(\Xi_2(\delta) + 2\tilde{\Xi}_3) \\
\zeta(\delta) &= \frac{14}{\sqrt{T}}\Xi_1(\delta) + \frac{7\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta) + \Xi_3) + 7\sqrt{8R_2(\delta)\log(\delta)} + 42R_1(\delta)\log(\delta)\,,
\end{aligned}$$

and (for any time-step $s$ such that $\lceil T/2\rceil \leq s \leq T$) the events

$$\begin{aligned}
A_1 &= \{\Lambda_1 \leq \Xi_1(\delta/3)\} \cap \{\Lambda_2 \leq \Xi_2(\delta/3)\} \\
A_2(s) &= \left\{Q_s > \sqrt{8R_2(\delta/3)\log(3/\delta)} + 6R_1(\delta/3)\log(3/\delta)\right\} \\
A_3(s) &= \left\{\langle Q\rangle_s + [Q]_s \leq R_1(\delta/3)Q_s + R_2(\delta/3)\right\}.
\end{aligned}$$

[7] As $\tilde{\Xi}_3$ is an upper bound for $\sum_{t=\lceil T/2\rceil}^{T} \rho_t \mathbb{E}_t[\|\xi_t\|_*^2]$.

Now, notice that

$$\begin{aligned}
P\Big(f(x_T) - f^* > \zeta(\delta/3)\Big) &= P\Big(\{f(x_T) - f^* > \zeta(\delta/3)\} \cap A_1\Big) + P\Big(\{f(x_T) - f^* > \zeta(\delta/3)\} \cap \overline{A_1}\Big) \\
&\leq P\Big(\{f(x_T) - f^* > \zeta(\delta/3)\} \cap A_1\Big) + P(\overline{A_1}) \\
&\overset{(a)}{\leq} P\Big(\{f(x_T) - f^* > \zeta(\delta/3)\} \cap A_1\Big) + 2\delta/3 \\
&\overset{(b)}{\leq} P\left(\left\{\frac{14}{\sqrt{T}}\Lambda_1 + 7Q_{n^*} + \frac{7\sqrt{2}\eta}{\sqrt{T}}(\Lambda_2 + \Xi_3) > \zeta(\delta/3)\right\} \cap A_1\right) + 2\delta/3 \\
&\leq P\left(\left\{\frac{14}{\sqrt{T}}\Xi_1(\delta/3) + 7Q_{n^*} + \frac{7\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3) > \zeta(\delta/3)\right\} \cap A_1\right) + 2\delta/3 \\
&= P\left(A_2(n^*) \cap A_1\right) + 2\delta/3 \\
&\overset{(c)}{\leq} P\Big(A_2(n^*) \cap A_3(n^*)\Big) + 2\delta/3 \\
&\leq P\left(\bigcup_{s=\lceil T/2\rceil}^{T} (A_1(s) \cap A_2(s))\right) + 2\delta/3 \\
&\overset{(d)}{\leq} \delta/3 + 2\delta/3 = \delta\,,
\end{aligned}$$

where $(a)$ follows from (8) and a union bound, $(b)$ follows from (11), $(c)$ follows from (14) and the definitions of $R_1$, $R_2$, and $A_3$, whereas $(d)$ follows from Proposition 1 and the fact that $(Q_t)_{t=\lceil T/2\rceil}^T$ is a (square integrable) martingale adapted to $(\mathcal{F}_t)_{t=\lceil T/2\rceil}^T$ (with $Q_{\lceil T/2\rceil} = 0$).

Hence, with probability at least $1-\delta$,

$$\begin{aligned}
\frac{1}{7}(f(x_T) - f^*) &\leq \frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3) + \sqrt{8R_2(\delta/3)\log(3/\delta)} + 6R_1(\delta/3)\log(3/\delta) \\
&\overset{(a)}{=} \frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3) + \frac{36\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + 2\tilde{\Xi}_3)\log(3/\delta) \\
&\quad + 2\sqrt{2}\sqrt{\frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3)}\sqrt{\frac{6\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + 2\tilde{\Xi}_3)\log(3/\delta)} \\
&\overset{(b)}{\leq} \frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3) + \frac{36\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + 2\tilde{\Xi}_3)\log(3/\delta) \\
&\quad + 4\left(\frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3)\right) + \frac{3\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + 2\tilde{\Xi}_3)\log(3/\delta) \\
&= 5\left(\frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + \Xi_3)\right) + \frac{39\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + 2\tilde{\Xi}_3)\log(3/\delta) \\
&\overset{(c)}{=} 5\left(\frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}(\Xi_2(\delta/3) + (G^2+\sigma^2)\log(4T))\right) \\
&\qquad\qquad + \frac{39\sqrt{2}\eta}{\sqrt{T}}\Big(\Xi_2(\delta/3) + 2\sigma^2\log(4T)\Big)\log(3/\delta) \\
&\overset{(d)}{\leq} 5\left(\frac{2}{\sqrt{T}}\Xi_1(\delta/3) + \frac{\sqrt{2}\eta}{\sqrt{T}}G^2\log(4T)\right) + \frac{44\sqrt{2}\eta}{\sqrt{T}}\Big(\Xi_2(\delta/3) + 2\sigma^2\log(4T)\Big)\log(3/\delta)\,,
\end{aligned}$$

where $(a)$ follows from the definitions of $R_1$ and $R_2$, $(b)$ follows from the elementary fact that $ab \leq a^2/2 + b^2/2$, $(c)$ follows from the definitions of $\Xi_3$ and $\tilde{\Xi}_3$, and $(d)$ follows from the fact that $\log(3/\delta) \geq 1$. We can then conclude that with probability at least $1-\delta$,

$$f(x_T) - f^* \leq \frac{35}{\sqrt{T}}\left(2\Xi_1(\delta/3) + \sqrt{2}\eta G^2\log(4T) + 9\sqrt{2}\eta\Big(\Xi_2(\delta/3) + 2\sigma^2\log(4T)\Big)\log(3/\delta)\right).$$

□

### D.6 Proof of Corollary 3

**Corollary 3.** *For any $\delta \in (0,1)$ and $\eta > 0$, Algorithm 1 with $\eta_t = \frac{\eta}{\sqrt{t}}$ satisfies the following with probability at least $1-\delta$, where $C_1$ and $C_2$ are constant depending solely on, respectively, $\theta$ and $p$.*
*(i) Under Assumptions 1, 2 and 4, $f(x_T) - f^*$ is bounded by*

$$\frac{C_1 \log(eT)}{\sqrt{T}} \left( \frac{B_\psi(x^*, x_1)}{\eta} + \eta \left( G^2 + \phi^2 \log^{2\theta+1}(e/\delta) \right) \right)$$

*(ii) Under Assumptions 1, 2 and 5, $f(x_T) - f^*$ is bounded by*

$$\frac{C_2 \log(eT)}{\sqrt{T}} \left( \frac{B_\psi(x^*, x_1)}{\eta} + \eta \Big( G^2 + \phi^2 (1/\delta)^{2/p} \log(e/\delta) \Big) \right)$$

*Proof.* Starting with case (i), we let $C, C_1, C_2, \ldots$ denote positive constants—depending only on $\theta$—whose values may change between steps. In the notation of Theorem 2, we choose

$$\Xi_1(\delta) = C \log(eT) \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \left( G^2 + \phi^2 \log^{2\theta}(e/\delta) \right) \right)$$

by virtue of Corollary 1. While invoking Lemma 7(ii) with $s = 2$ and $\omega_t = \rho_t$ allows us to choose[8]

$$\Xi_2(\delta) = C_1 \phi^2 \sqrt{\sum_{t=\lceil T/2 \rceil}^{T} \rho_t^2 \log(2/\delta)} + C_2 \phi^2 \max_{\lceil T/2 \rceil \le t \le T} \rho_t \log^{2\theta} \left( \frac{2e \sum_{t=\lceil T/2 \rceil}^{T} \rho_t^2}{\max_{\lceil T/2 \rceil \le t \le T} \rho_t^2 \delta} \right) .$$

Then, using that (via Lemma 11)

$$\sum_{t=\lceil T/2 \rceil}^{T} \rho_t^2 \le 3 \qquad \text{and} \qquad \max_{\lceil T/2 \rceil \le t \le T} \rho_t = \rho_T = \frac{1}{2} ,$$

Theorem 2 yields that

$$\begin{aligned} f(x_T) - f^* &\le \frac{C \log(eT)}{\sqrt{T}} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \left( G^2 + \phi^2 \log^{2\theta}(e/\delta) \right) + \eta G^2 \right. \\ &\qquad \left. + \eta \Big( \phi^2 \sqrt{\log(e/\delta)} + \phi^2 \log^{2\theta}(e/\delta) + \phi^2 \Big) \log(e/\delta) \right) \\ &\le \frac{C \log(eT)}{\sqrt{T}} \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \left( G^2 + \phi^2 \log^{2\theta+1}(e/\delta) \right) \right) , \end{aligned}$$

where upon invoking Theorem 2, we used the fact that Assumption 4 implies Assumption 3 with $\sigma^2 = 2\Gamma(2\theta+1)\phi^2$ thanks to Lemma 12.

For case (ii), we let $C$ denote a positive constant—depending only on $p$—whose value may change between steps. Via Corollary 2, we can choose

$$\Xi_1(\delta) = C \log(eT) \left( \frac{1}{\eta} B_\psi(x^*, x_1) + \eta \Big( G^2 + \phi^2 (1/\delta)^{2/p} \Big) \right)$$

Invoking Lemma 8(ii) with $\omega_t = \rho_t$ allows us to choose

$$\Xi_2(\delta) = 2\phi^2 \sqrt{6 \log(1/\delta)} + 2(2 + (p/6)) \phi^2 (3/\delta)^{2/p} ,$$

[8] This is valid despite the fact that, contrary to Lemma 7(ii), the indices here start from $\lceil T/2 \rceil$.

where we have used that

$$\sum_{t=\lceil T/2\rceil}^{T} \rho_t^{p/2} \le \sum_{t=\lceil T/2\rceil}^{T} \rho_t^2 \le 3\,,$$

which holds via Lemma 11 and the fact that $p > 4$ and $\rho_t \le 1$. Theorem 2 then implies that

$$\begin{aligned} f(x_T) - f^* &\le \frac{C\log(eT)}{\sqrt{T}}\left(\frac{1}{\eta}B_\psi(x^*, x_1) + \eta\Big(G^2 + \phi^2(1/\delta)^{2/p}\Big) + \eta G^2 \right. \\ &\qquad \left. + \eta\Big(\phi^2\sqrt{\log(e/\delta)} + \phi^2(1/\delta)^{2/p} + \phi^2\Big)\log(e/\delta)\right) \\ &\le \frac{C\log(eT)}{\sqrt{T}}\left(\frac{1}{\eta}B_\psi(x^*, x_1) + \eta\Big(G^2 + \phi^2(1/\delta)^{2/p}\log(e/\delta)\Big)\right), \end{aligned}$$

where upon invoking Theorem 2, we used the fact that Assumption 5 implies Assumption 3 with $\sigma^2 = \phi^2$; while in the second step, we used the fact that $\sqrt{\log(e/\delta)} \le \sqrt{(p/4)(e/\delta)^{(4/p)}} = \sqrt{(p/4)}(e/\delta)^{(2/p)}$. □

### D.7 Auxiliary Lemmas

**Lemma 10.** *Let $a$ and $b$ be two positive integers such that $a \le b < T$. Then,*

$$\sum_{j=a}^{b} \frac{1}{(T-j)(T-j+1)} = \frac{1}{T-b} - \frac{1}{T-a+1}\,.$$

*Proof.*

$$\sum_{j=a}^{b} \frac{1}{(T-j)(T-j+1)} = \sum_{j=a}^{b} \frac{1}{(T-j)} - \frac{1}{(T-j+1)} = \frac{1}{T-b} - \frac{1}{T-a+1}\,.$$

□

**Lemma 11.** *For $j < T$, let $\alpha_j = \frac{1}{(T-j)(T-j+1)}$. Then, for $T \ge 1$, we have that*

$$\sum_{t=\lceil T/2\rceil}^{T} \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j \le \log(4T) \qquad \text{and} \qquad \sum_{t=\lceil T/2\rceil}^{T} \left(\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j\right)^2 \le 3\,.$$

*Proof.* By Lemma 10,

$$\sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j \le \frac{1}{T - t\wedge(T-1)}\,.$$

Assuming $T \ge 2$ (as the lemma follows directly otherwise), we have that

$$\begin{aligned} \sum_{t=\lceil T/2\rceil}^{T} \sum_{j=\lceil T/2\rceil}^{t\wedge(T-1)} \alpha_j &\le \sum_{t=\lceil T/2\rceil}^{T} \frac{1}{T - t\wedge(T-1)} \\ &= 2 + \sum_{t=\lceil T/2\rceil}^{T-2} \frac{1}{T-t} \\ &\le 2 + \int_{\lceil T/2\rceil}^{T-1} \frac{1}{T-t}dt \\ &= 2 + \log(\lfloor T/2\rfloor) \le \log(4T)\,. \end{aligned}$$

Similarly,

$$\sum_{t=\lceil T/2 \rceil}^{T} \left( \sum_{j=\lceil T/2 \rceil}^{t \wedge (T-1)} \alpha_j \right)^2 \leq 2 + \int_{\lceil T/2 \rceil}^{T-1} \frac{1}{(T-t)^2} dt \leq 3 \,.$$

□

# E CONCENTRATION INEQUALITIES FOR MARTINGALES WITH HEAVY-TAILED INCREMENTS

We collect in this section relevant concentration results for Martingales with heavy-tailed increments. We treat two families of heavy-tailed random variables: a class of sub-Weibull random variables, and a class of random variables with polynomially decaying tails (implied by a bounded moment assumption).

## E.1 Sub-Weibull Increments

Before stating the main concentration inequality in Proposition 2, we collect some basic results concerning sub-Weibull random variables. The following lemma (adapted from (Madden et al., 2021)) provides an upper bound for the $p$-th absolute moment of a sub-Weibull random variable.

**Lemma 12.** *(Madden et al., 2021, Lemma 22) Let $X$ be a sub-Weibull$(\theta, \phi)$ random variable. Then, for any $p > 0$, it satisfies*

$$\mathbb{E}|X|^p \leq 2\Gamma(\theta p + 1)\phi^p \,.$$

The following lemma shows that centering a random variable preserves the sub-Weibull property up to a constant depending on $\theta$.

**Lemma 13.** *Let $X$ be a sub-Weibull$(\theta, \phi)$ random variable. Then $X - \mathbb{E}X$ is sub-Weibull$(\theta, c_\theta \phi)$, where $c_\theta = 2^{\max\{\theta,1\}+1}\Gamma(\theta+1)/\ln^\theta(2)$.*

*Proof.* If $\theta \leq 1$, define

$$\|X\|_{\psi_{1/\theta}} = \inf\left\{ t > 0 \,:\, \mathbb{E}\Big[\exp\Big( (|X|/t)^{1/\theta} \Big)\Big] \leq 2 \right\}$$

which is an (Orlicz) norm for the space $L_{\psi_{1/\theta}} = \{X \,:\, \|X\|_{\psi_{1/\theta}} < \infty\}$ (Vershynin, 2018, Section 2.7.1). Clearly, $X$ is sub-Weibull$(\theta, \phi)$ if and only if $\|X\|_{\psi_{1/\theta}} \leq \phi$. Starting with the triangle inequality, we proceed in the same manner as in the proof of (Vershynin, 2018, Lemma 2.6.8) to get that

$$\|X - \mathbb{E}X\|_{\psi_{1/\theta}} \leq \|X\|_{\psi_{1/\theta}} + \|\mathbb{E}X\|_{\psi_{1/\theta}} \leq \phi + \frac{|\mathbb{E}X|}{\ln^\theta(2)} \leq \phi + \frac{\mathbb{E}|X|}{\ln^\theta(2)} \leq \left( \frac{2\Gamma(\theta+1)}{\ln^\theta(2)} + 1 \right)\phi \,,$$

where the last inequality is an application of Lemma 12. Hence, the lemma follows for the case when $\theta \leq 1$ after using that $2\Gamma(\theta+1)/\ln^\theta(2) \geq 1$. On the other hand, when $\theta > 1$, $\|\cdot\|_{\psi_{1/\theta}}$ is no longer a norm. Instead, we exploit

the fact that $x^{1/\theta}$ is a sub-additive function in $x$ for $\theta > 1$ and $x \geq 0$. In particular, we have that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\left(\frac{|X-\mathbb{E}X|}{c_\theta\phi}\right)^{1/\theta}\right)\right] &\leq \mathbb{E}\left[\exp\left(\left(\frac{|X|+\mathbb{E}|X|}{c_\theta\phi}\right)^{1/\theta}\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(\frac{|X|}{c_\theta\phi}\right)^{1/\theta} + \left(\frac{\mathbb{E}|X|}{c_\theta\phi}\right)^{1/\theta}\right)\right] \\
&= \exp\left(\left(\frac{\mathbb{E}|X|}{c_\theta\phi}\right)^{1/\theta}\right)\mathbb{E}\left[\left(\exp\left((|X|/\phi)^{1/\theta}\right)\right)^{(1/c_\theta)^{1/\theta}}\right] \\
&\leq \exp\left(\left(\frac{\mathbb{E}|X|}{c_\theta\phi}\right)^{1/\theta}\right)\left(\mathbb{E}\left[\exp\left((|X|/\phi)^{1/\theta}\right)\right]\right)^{(1/c_\theta)^{1/\theta}} \\
&\leq \exp\left(\left(\frac{\mathbb{E}|X|}{c_\theta\phi}\right)^{1/\theta}\right)2^{(1/c_\theta)^{1/\theta}} \\
&\leq \exp\left(\left(\frac{2\Gamma(\theta+1)}{c_\theta}\right)^{1/\theta}\right)2^{(1/c_\theta)^{1/\theta}} \\
&\leq \exp\left(2\left(\frac{2\Gamma(\theta+1)}{c_\theta}\right)^{1/\theta}\right) = 2\,,
\end{aligned}$$

where the third inequality is an application of Jensen's inequality as the fact that $0 < (1/c_\theta)^{1/\theta} < 1$ implies the concavity of $x^{(1/c_\theta)^{1/\theta}}$ for $x \geq 0$, the fourth inequality uses that $X$ is sub-Weibull$(\theta, \phi)$, the fifth inequality follows via Lemma 12, and the last inequality uses the fact that $2\Gamma(\theta+1) \geq 1$. □

The following lemma collects upper bounds for the moment-generating function (MGF) of (centered) sub-Weibull random variables, depending on the value of $\theta$. The MGF of a random variable $X$ is a function of $\lambda \in \mathbb{R}$ given by $\mathbb{E}[\exp(\lambda X)]$. As mentioned before, our focus in this work is on the heavy-tailed regime where $\theta \geq 1$, though we also consider the canonical case of $\theta = 1/2$ for comparison. In the latter case, we have the standard bound on the MGF of a sub-Gaussian random variable (see, e.g., (Vershynin, 2018, Proposition 2.5.2)). When $\theta = 1$, a similar bound (see, e.g., (Vershynin, 2018, Proposition 2.7.1)) holds only for a certain range of $\lambda$. When $\theta > 1$, one cannot bound the MGF in general; thus, we settle for a bound on the MGF of a truncated version of the random variable due to Bakhshizadeh et al. (2023). This last result is reported in (Madden et al., 2021, Lemma 31) for a specific choice of the truncation parameter, which we will slightly modify when applying this lemma.

**Lemma 14.** *Let $X$ be a sub-Weibull$(\theta, \phi)$ random variable with $\mathbb{E}[X] = 0$.*

*(i) (Vershynin, 2018, Proposition 2.5.2) If $\theta = 1/2$,*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(4e\phi^2\lambda^2) \qquad \forall \lambda \in \mathbb{R}\,.$$

*(ii) (Vershynin, 2018, Proposition 2.7.1) If $\theta = 1$,*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(2e^2\phi^2\lambda^2) \qquad \forall \lambda : |\lambda| \leq \frac{1}{2e\phi}\,.$$

*(iii) If $\theta \geq 1$, let $L = \phi h$ for some parameter $h > 0$, and define $\tilde{X} = X\mathbb{I}\{X \leq L\}$. Then,*

$$\mathbb{E}[\exp(\lambda \tilde{X})] \leq \exp(a\phi^2\lambda^2) \qquad \forall \lambda \in \left[0, \frac{1}{2h^{1-\frac{1}{\theta}}\phi}\right],$$

*where*

$$a = (2^{2\theta}+1)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{6}h^{\frac{1}{\theta}-1}\,.$$

*Proof.*
*(iii)* Since $\mathbb{E}[X] = 0$, we have that for any $\lambda \in \left[0, \frac{1}{2h^{1-\frac{1}{\theta}}\phi}\right]$,

$$\begin{aligned}
\log \mathbb{E}[\exp(\lambda \tilde{X})] &\leq \frac{\lambda^2}{2}\left(\mathbb{E}\left[\tilde{X}^2 \mathbb{I}\{\tilde{X} \leq 0\}\right] + \mathbb{E}\left[\tilde{X}^2 \exp(\lambda \tilde{X}) \mathbb{I}\{\tilde{X} > 0\}\right]\right) \\
&\leq \frac{\lambda^2}{2}\left(\mathbb{E}\left[X^2 \mathbb{I}\{X \leq 0\}\right] + 2^{2\theta+1}\Gamma(2\theta+1)\phi^2 + \frac{2^{3\theta}\Gamma(3\theta+1)}{3} L^{\frac{1}{\theta}-1}\phi^{3-\frac{1}{\theta}}\right) \\
&= \frac{\lambda^2}{2}\left(\mathbb{E}\left[X^2 \mathbb{I}\{X \leq 0\}\right] + 2^{2\theta+1}\Gamma(2\theta+1)\phi^2 + \frac{2^{3\theta}\Gamma(3\theta+1)}{3} h^{\frac{1}{\theta}-1}\phi^2\right) \\
&\leq \frac{\lambda^2}{2}\left(2\Gamma(2\theta+1)\phi^2 + 2^{2\theta+1}\Gamma(2\theta+1)\phi^2 + \frac{2^{3\theta}\Gamma(3\theta+1)}{3} h^{\frac{1}{\theta}-1}\phi^2\right),
\end{aligned}$$

where the first inequality follows from Lemma 1 in (Bakhshizadeh et al., 2023), the second inequality follows from Corollary 2 in the same paper,[9] the equality holds by the definition of $L$, and the last inequality is an application of Lemma 12. □

The following proposition provides time-uniform concentration inequalities for martingales with conditionally sub-Weibull increments. Case (i) is a standard sub-Gaussian concentration result included for completeness, whereas Case (ii) considers the heavy-tailed regime where $\theta \geq 1$. The latter generalizes a result in (Madden et al., 2021, Proposition 11), which corresponds to the case when $s = 0$. In our problem, this generalized form allows us in come cases to avoid an extra poly-logarithmic dependence on the time horizon, at the cost of a constant depending on $\theta$. This is thanks to the (possibly) non-uniform union bound employed when $s > 0$, which can take advantage of the non-uniformity of the sequence $(m_i)$.

**Proposition 2.** *Assume that $(X_i)_{i=1}^n$ is a martingale difference sequence adapted to filtration $\mathbb{F} = (\mathcal{F}_i)_{i=0}^n$, where $n$ is a positive integer, and let $S_t = \sum_{i=1}^t X_i$ for $t \in [n]$. Furthermore, assume that for each $i \in [n]$, $X_i$ is sub-Weibull$(\theta, m_i)$ conditioned on $\mathcal{F}_{i-1}$; that is,*

$$\mathbb{E}\left[\exp\left(\left(|X_i|/m_i\right)^{1/\theta}\right) \mid \mathcal{F}_{i-1}\right] \leq 2\,,$$

*for some constant $m_i > 0$, and define $m_* = \max_i m_i$. Then, for any $\delta \in (0,1)$:*

*(i) If $\theta = 1/2$,*

$$P\left(\bigcup_{t=1}^n \left\{S_t \geq 4\sqrt{e\textstyle\sum_{i=1}^n m_i^2 \log(1/\delta)}\right\}\right) \leq \delta\,.$$

*(ii) If $\theta \geq 1$; then for any $s \geq 0$,*

$$P\left(\bigcup_{t=1}^n \left\{S_t \geq \sqrt{C_1 \textstyle\sum_{i=1}^n m_i^2 \log(2/\delta)} + 4m_* \max\left\{\log^{\theta-1}\left(\frac{2e\sum_{j=1}^n m_j^s}{m_*^s \delta}\right), (s\theta - s)^{\theta-1}\right\} \log(2/\delta)\right\}\right) \leq \delta\,,$$

*where $C_1 = 2^{3\theta+1}\Gamma(3\theta+1)$.*

*Proof.*

(i) We have via Lemma 14(i) that for every $i \in [n]$,

$$\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \leq \exp(4em_i^2\lambda^2) \qquad \forall \lambda \in \mathbb{R}\,.$$

Hence, the required result follows from Lemma 17(i) using that $r^2 \leq 4e\sum_{i=1}^n m_i^2$.

[9] In the notation of Bakhshizadeh et al. (2023), we have that $\alpha = \theta$, $c_\alpha = \phi^{-1/\theta}$, $\lambda = \beta I(L)/L$ with $I(L) = (L/\phi)^{1/\theta}$ and $\beta \in [0, 1/2]$. Compared to Corollary 2 in (Bakhshizadeh et al., 2023), the extra factor of 2 in the last two terms on the right-hand side of the inequality is because in our case (similar to Madden et al. (2021)), we start with the assumption that $X$ is sub-Weibull$(\theta, \phi)$, which implies the tail bound $P(|X| \geq t) \leq 2\exp(-I(t))$.

(ii) For $i \in [n]$, let $\ell_i = m_i h_i$, where

$$h_i = \log^\theta\left(\frac{e\sum_{j=1}^n m_j^s}{m_i^s \delta'}\right)$$

for some $\delta' \in (0,1)$. Define $\tilde{X}_i = X_i \mathbb{I}\{X_i \le \ell_i\}$ and $\tilde{S}_t = \sum_{i=1}^t \tilde{X}_i$. Note that for any $x > 0$,

$$P\left(\bigcup_{t=1}^n \{S_t \ge x\}\right) \le P\left(\bigcup_{t=1}^n \{\tilde{S}_t \ge x\}\right) + P\left(\bigcup_{i=1}^n \{X_i > \ell_i\}\right). \tag{15}$$

Starting with the second term, we perform a union bound and proceed in a similar manner to the proof of Proposition 11 in (Madden et al., 2021):

$$\begin{aligned}
P\left(\bigcup_{i=1}^n \{X_i > \ell_i\}\right) &\le \sum_{t=1}^n P\left(X_i > \ell_i\right) \\
&= \sum_{i=1}^n P\left(\exp\left(\left(X_i/m_i\right)^{1/\theta}\right) > \exp\left(h_i^{1/\theta}\right)\right) \\
&\le \sum_{i=1}^n \exp\left(-h_i^{1/\theta}\right)\mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(X_i/m_i\right)^{1/\theta}\right) \mid \mathcal{F}_{i-1}\right]\right] \\
&\le 2\sum_{i=1}^n \exp\left(-h_i^{1/\theta}\right) \\
&= \frac{2}{e}\sum_{i=1}^n \frac{m_i^s \delta'}{\sum_{j=1}^n m_j^s} = \frac{2}{e}\delta' \le \delta'.
\end{aligned} \tag{16}$$

Returning to the first term in (15), notice that for $i \in [n]$, Lemma 14(iii) implies that

$$\mathbb{E}[\exp(\lambda \tilde{X}_i) \mid \mathcal{F}_{i-1}] \le \exp(a m_i^2 \lambda^2) \qquad \forall \lambda \in \left[0, \frac{1}{2h_i^{1-\frac{1}{\theta}} m_i}\right],$$

where[10] $a = (2^{2\theta}+1)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{6}$. In preparation for applying Lemma 17(ii), we study the term

$$\max_{i\in[n]} m_i h_i^{1-\frac{1}{\theta}} = \max_{i\in[n]} m_i \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_i^s \delta'}\right).$$

Assuming that $s > 0$ and $\theta > 1$, let $w = \sum_{j=1}^n m_j^s/\delta'$, and observe that $e^{1/s} w^{1/s} \ge w^{1/s} \ge m_*$. Define $f(z) = z\log^{\theta-1}(ew/z^s)$, and let $\hat{z}_1 = \exp(1-\theta+1/s)w^{1/s}$ and $\hat{z}_2 = e^{1/s}w^{1/s}$. By inspecting its first derivative,

$$f'(z) = \log^{\theta-2}(ew/z^s)(\log(ew/z^s) - s(\theta-1)),$$

we observe that $f$ is increasing in $(0, \hat{z}_1)$ and decreasing in $(\hat{z}_1, \hat{z}_2)$. Hence, if $m_* \le \hat{z}_1$; then,

$$\max_{i\in[n]} m_i \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_i^s \delta'}\right) = m_* \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_*^s \delta'}\right).$$

Otherwise,

$$\max_{i\in[n]} m_i \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_i^s \delta'}\right) \le \hat{z}_1 \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{\hat{z}_1^s \delta'}\right) = \hat{z}_1 (s\theta - s)^{\theta-1} \le m_* (s\theta - s)^{\theta-1}.$$

Combing both cases yields that

$$\max_{i\in[n]} m_i \log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_i^s \delta'}\right) \le m_* \max\left\{\log^{\theta-1}\left(\frac{e\sum_{j=1}^n m_j^s}{m_*^s \delta'}\right), (s\theta - s)^{\theta-1}\right\},$$

[10] We have used the fact that $h_i \ge 1$ and $\theta \ge 1$ to bound the value of $a$ stated in the lemma.

which, trivially, also holds when either $s = 0$ or $\theta = 1$. Subsequently, if we define $u_1 = a\sum_{i=1}^{n} m_i^2$ and $u_2$ as twice the right-hand side of the above inequality, we can apply Lemma 17(ii) with $r^2 \leq u_1$ and $b \leq u_2$ to obtain that

$$P\left(\bigcup_{t=1}^{n}\left\{\tilde{S}_t \geq x\right\}\right) \leq \exp\left(-\min\left\{\frac{x^2}{4u_1}, \frac{x}{2u_2}\right\}\right).$$

Finally, by choosing $x = \sqrt{4u_1 \log(1/\delta')} + 2u_2 \log(1/\delta')$, we can upper bound the right-hand side of the above inequality with $\delta'$. Combining this with (16) and (15), the required result follows after setting $\delta' = \delta/2$ and using that $a \leq 2^{3\theta+1}\Gamma(3\theta+1)$.

□

### E.2 Increments with a Bounded Moment Condition

The following proposition, a weaker version of Corollary 3.2 in (Rio, 2017), is an analogue of Proposition 2 when we only have the assumption that the increments of the martingale have a finite $p$-th absolute moment for some $p > 2$.

**Proposition 3.** *Assume that $(X_i)_{i=1}^n$ is a martingale difference sequence adapted to filtration $\mathbb{F} = (\mathcal{F}_i)_{i=0}^n$, where $n$ is a positive integer, and let $S_t = \sum_{i=1}^{t} X_i$ for $t \in [n]$. Moreover, assume that there exists a constant $p > 2$ such that for each $i \in [n]$, $X_i$ satisfies*

$$\mathbb{E}\left[(|X_i|/m_i)^p \mid \mathcal{F}_{i-1}\right] \leq 1$$

*for some finite constant $m_i > 0$. Then, for any $\delta \in (0, 1)$:*

$$P\left(\bigcup_{t=1}^{n}\left\{S_t > \sqrt{2\textstyle\sum_{i=1}^{n} m_i^2 \log(1/\delta)} + (2 + (p/3))(\textstyle\sum_{i=1}^{n} m_i^p/\delta)^{1/p}\right\}\right) \leq \delta\,.$$

*Proof.* Using that $\max\{0, X_i\} \leq |X_i|$ and $\mathbb{E}\left[X_i^2 \mid \mathcal{F}_{i-1}\right] \leq \left(\mathbb{E}\left[|X_i|^p \mid \mathcal{F}_{i-1}\right]\right)^{2/p} \leq m_i^2$, the result follows from Corollary 3.2 and Remark 3.3 in (Rio, 2017). □

### E.3 Auxiliary Lemmas

For a random variable $X$, define $\|X\|_p := (\mathbb{E}|X|^p)^{1/p}$ for $p > 0$. The following lemma relates $\|X - \mathbb{E}X\|_p$ to $\|X\|_p$ when $p \geq 1$.

**Lemma 15.** *Let $X$ be a random variable satisfying $\|X\|_p < \infty$ for some $p \geq 1$. Then, $\|X - \mathbb{E}X\|_p \leq 2\|X\|_p$.*

*Proof.* We have that

$$\|X - \mathbb{E}X\|_p \leq \|X\|_p + \|\mathbb{E}X\|_p = \|X\|_p + |\mathbb{E}X| \leq \|X\|_p + \mathbb{E}|X| \leq 2\|X\|_p\,,$$

where the first inequality is an application of the triangle's inequality as $\|\cdot\|_p$ is a norm for $p \geq 1$, the second inequality is an application of Jensen's inequality, and the last inequality holds since $\|X\|_p$ is an increasing function in $p$. □

The following lemma, similar in spirit to (Madden et al., 2021, Lemma 26), allows us to reuse a standard argument when proving time-uniform concentration inequalities.

**Lemma 16.** *Fix a positive integer $n$ and assume that $(V_t)_{t=0}^n$ is a non-negative supermartingale adapted to filtration $(\mathcal{F}_t)_{t=0}^n$ with $V_0 = 1$. Let $(A_t)_{t=1}^n$ be a sequence of events adapted to the same filtration, and assume that there exists a constant $\zeta > 0$ such that for any $t \in [n]$, it holds almost surely that $\mathbb{I}\{A_t\} \leq \zeta V_t$. Then,*

$$P\left(\bigcup_{t=1}^{n} A_t\right) \leq \zeta\,.$$

*Proof.* Define the stopping time $\tau = \min\{t \in [n] : \mathbb{I}\{A_t\} = 1\}$, where $\min(\emptyset) = \infty$. Since $(V_t)_{t=0}^n$ is a supermartingale, the stopped process $(V_{t\wedge\tau})_{t=0}^n$ is also a supermartingale (Williams, 1991, Theorem 10.9), where $t \wedge \tau = \min\{t, \tau\}$. This implies in particular that

$$\mathbb{E}[V_{n\wedge\tau}] \leq \mathbb{E}[V_0] = 1\,.$$

Hence,

$$P\left(\bigcup_{t=1}^n A_t\right) = P\left(A_{n\wedge\tau}\right) = \mathbb{E}[\mathbb{I}\{A_{n\wedge\tau}\}] \leq \zeta\mathbb{E}[V_{n\wedge\tau}] \leq \zeta\,,$$

where the first inequality holds since

$$P(\mathbb{I}\{A_{n\wedge\tau}\} > \zeta V_{n\wedge\tau}) \leq P\left(\bigcup_{t=1}^n \{\mathbb{I}\{A_t\} > \zeta V_t\}\right) = 0\,.$$

□

The following lemma provides, via standard tools, concentration inequalities for sums of random variables enjoying sub-Gaussian or sub-exponential type bounds on their (conditional) moment generating functions.

**Lemma 17.** *Assume that $(X_i)_{i=1}^n$ is a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_i)_{i=0}^n$, where $n$ is a positive integer, and let $S_t = \sum_{i=1}^t X_i$ for $t \in [n]$. Moreover, let $(R_i)_{i=0}^n$ be a sequence of random variables adapted to the same filtration, and define $r^2 = \left\|\sum_{i=1}^n R_{i-1}^2\right\|_\infty$.*

*(i) If for all $i \in [n]$, it holds that*

$$\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \leq \exp(R_{i-1}^2\lambda^2) \qquad \forall\lambda \in \mathbb{R}\,;$$

*then, for all $x > 0$,*

$$P\left(\bigcup_{t=1}^n \{S_t \geq x\}\right) \leq \exp\left(-\frac{x^2}{4r^2}\right).$$

*(ii) Let $(B_i)_{i=0}^n$ be an $\mathbb{F}$-adapted sequence of positive random variables, and define $b = \max_i \|B_i\|_\infty$. If for all $i \in [n]$, it holds that*

$$\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \leq \exp(R_{i-1}^2\lambda^2) \qquad \forall\lambda \in \left[0, \frac{1}{B_{i-1}}\right];$$

*then, for all $x > 0$,*

$$P\left(\bigcup_{t=1}^n \{S_t \geq x\}\right) \leq \exp\left(-\min\left\{\frac{x^2}{4r^2}, \frac{x}{2b}\right\}\right).$$

*Proof.* Define the set $\Lambda$ as $\mathbb{R}_{\geq 0}$ in case *(i)* and as $[0, 1/b]$ in case *(ii)*. Then, in either case, for any fixed $\lambda \in \Lambda$, the process $(V_t(\lambda))_{t=0}^n$, where

$$V_t(\lambda) = \prod_{i=1}^t \frac{\exp(\lambda X_i)}{\exp(R_{i-1}^2\lambda^2)}\,, \qquad\qquad V_0(\lambda) = 1$$

is an $\mathbb{F}$-adapted non-negative supermartingale. Moreover, notice that for any $t \in [n]$, it holds almost surely that

$$\begin{aligned}
\mathbb{I}\{S_t \geq x\} &\leq \exp\left(\lambda S_t - \lambda x + \lambda^2 r^2 - \lambda^2 \sum_{i=1}^t R_{i-1}^2\right) \\
&= \exp\left(-\lambda x + \lambda^2 r^2\right) \exp\left(\lambda \sum_{i=1}^t X_i - \lambda^2 \sum_{i=1}^t R_{i-1}^2\right) \\
&= \exp\left(-\lambda x + \lambda^2 r^2\right) V_t(\lambda)\,.
\end{aligned}$$

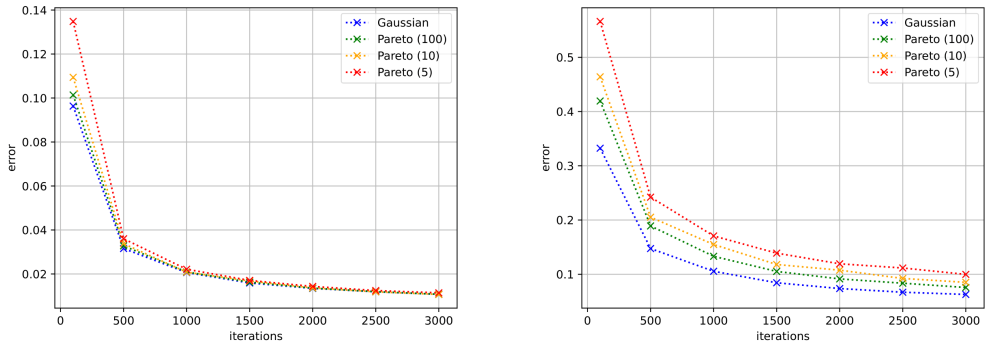
Consequently, Lemma 16 implies that

$$P\left(\bigcup_{t=1}^{n}\{S_t \geq x\}\right) \leq \exp\left(-\lambda x + \lambda^2 r^2\right) .$$

From this, the result in case *(i)* follows by choosing $\lambda = \frac{x}{2r^2}$, while the result in case *(ii)* follows by choosing $\lambda = \min\{\frac{x}{2r^2}, \frac{1}{b}\}$ and using that $\frac{r^2}{b^2} \leq \frac{x}{2b}$ whenever $\frac{x}{2r^2} \geq \frac{1}{b}$. □

# F ADDITIONAL EXPERIMENTS

In this section, we present an additional experiment comparing the performance of the average of the iterates with that of the last iterate under polynomially-tailed noise. As before, we use Algorithm 1 to minimize $f(x) = |x|$ over $\mathbb{R}$ with $\psi(x) = 1/2\|x\|_2^2$ (i.e., classical SGD). For the noise, we consider the Gaussian distribution with variance 1 and three different (symmetric) Pareto distributions with the shape parameter set to 5, 10, and 100 respectively. For a fair comparison, the Pareto distributions are scaled to have unit variance. We use $1/\sqrt{T}$ as a fixed step-size and run the algorithm for seven values of $T$ ranging from 100 to 3k. We report the 99-percentile of the optimization error evaluated over 10k runs for each of the aforementioned noise distributions. The results for the average iterate and the last iterate are reported in plots (a) and (b) respectively. Much like the Weibull case, the average iterate appears more robust to heavy-tailed noise compared to the last iterate. This is again deducible from the fact that the 99-percentile curves of the optimization error of the average iterate converge towards the Gaussian rate (as predicted by the two-regime bounds), while the separation between them seems to persist (at least for longer) for the last iterate.




(a) (b)

Figure 2: The performance of the average iterate and the last iterate are reported in the plots on the left and the right respectively. The plots show the 99-percentile of the error across runs for different choices of the noise distribution.