High Probability Convergence of Stochastic Gradient Methods

Zijian Liu [*1] **Ta Duy Nguyen** [*2] **Thien Hang Nguyen** [*3] **Alina Ene** [2] **Huy L. Nguyen** [3]

Abstract

In this work, we describe a generic approach to show convergence with high probability for both stochastic convex and non-convex optimization with sub-Gaussian noise. In previous works for convex optimization, either the convergence is only in expectation or the bound depends on the diameter of the domain. Instead, we show high probability convergence with bounds depending on the initial distance to the optimal solution. The algorithms use step sizes analogous to the standard settings and are universal to Lipschitz functions, smooth functions, and their linear combinations. The method can be applied to the non-convex case. We demonstrate an $O((1+\sigma^2\log(1/\delta))/T+\sigma/\sqrt{T})$ convergence rate when the number of iterations T is known and an $O((1+\sigma^2\log(T/\delta))/\sqrt{T})$ convergence rate when T is unknown for SGD, where $1-\delta$ is the desired success probability. These bounds improve over existing bounds in the literature. We also revisit AdaGrad-Norm (Ward et al., 2019) and show a new analysis to obtain a high probability bound that does not require the bounded gradient assumption made in previous works. The full version of our paper contains results for the standard per-coordinate AdaGrad.

1. Introduction

Stochastic optimization is a fundamental area with extensive applications in many domains, ranging from machine learning to algorithm design and beyond. The design and analysis of iterative methods for stochastic optimization has been the focus of a long line of work, leading to a rich understanding of the convergence of paradigmatic iterative methods such as stochastic gradient descent, mirror descent, and accelerated methods for both convex and non-convex optimization. However, most of these works only establish convergence guarantees that hold only in expectation. Although very meaningful, these results do not fully capture the convergence behaviors of the algorithms when we perform only a small number of runs of the algorithm, as it is typical in modern machine learning applications where there are significant computational and statistical costs associated with performing multiple runs of the algorithm (Harvey et al., 2019; Madden et al., 2020; Davis et al., 2021). Thus an important direction is to establish convergence guarantees for a single run of the algorithm that hold not only in expectation but also with high probability.

Compared to the guarantees that hold in expectation, high probability guarantees are significantly harder to obtain and they hold in more limited settings with stronger assumptions on the problem settings and the stochastic noise distribution. Most existing works that establish high probability guarantees focus on the setting where the length of the stochastic noise follows a light-tail (sub-Gaussian) distribution (Juditsky et al., 2011; Lan, 2012; 2020; Li & Orabona, 2020; Madden et al., 2020; Kavis et al., 2021). Recent works also study the more challenging heavy-tail setting, notably under a bounded variance (Nazin et al., 2019; Gorbunov et al., 2020; Cutkosky & Mehta, 2021) or bounded p-moment assumption (Cutkosky & Mehta, 2021) on the length of the stochastic noise. Both settings are highly relevant in practice (Zhang et al., 2020).

Despite this important progress, the convergence of cornerstone methods is not fully understood even in the more structured light-tailed noise setting. Specifically, the existing works for both convex and non-convex optimization with light-tailed noise rely on strong assumptions on the optimization domain and the gradients that significantly limit their applicability:

The problem domain is restricted to either the unconstrained domain or a constrained domain with bounded Bregman diameter. The convergence guarantees established depend on the Bregman diameter of the domain instead of the initial distance to the optimum. Even for compact domains, since the diameter can be much larger than the initial distance,

[*]Equal contribution [1]Stern School of Business, New York University [2]Department of Computer Science, Boston University [3]Khoury College of Computer Sciences, Northeastern University. Correspondence to: Zijian Liu <zl3067@nyu.edu>, Ta Duy Nguyen <taduy@bu.edu>, Thien Hang Nguyen <nguyen.thien@northeastern.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

these guarantees are pessimistic and diminish the benefits of good initializations. Thus an important direction remains to establish high probability guarantees for general optimization that scale only with the initial Bregman distance.

The gradients or stochastic gradients are assumed to be bounded even in the smooth setting. These additional assumptions are very restrictive and they significantly limit the applicability of the algorithm, e.g., they do not apply to important settings such as quadratic optimization. Moreover, the stochastic gradient assumption is more restrictive than other commonly studied assumptions, such as the gradients and the stochastic noise being bounded almost surely.

The above assumptions are not merely a technical artifact, and they stem from very important considerations. The high probability convergence guarantees are established via martingale concentration inequalities that impose necessary conditions on how much the martingale sequence can change in each step. However, the natural martingale sequences that arise in optimization depend on quantities such as the distance between the iterates and the optimum and the stochastic gradients, which are not a priori bounded. The aforementioned assumptions ensure that the concentration inequalities can be readily applied due to the relevant stochastic terms being all bounded almost surely. These difficulties are even more pronounced for the analysis of adaptive algorithms in the AdaGrad family that set the step sizes based on the stochastic gradients. The adaptive step sizes introduce correlations between the step sizes and the update directions, and a crucial component is the analysis of the evolution of the adaptive step sizes and the cumulative stochastic noise. If the gradients are bounded, both of these challenges can be overcome by paying error terms proportional to the lengths of the gradients and stochastic gradients. Removing the bounded gradient assumptions requires new technical insights and tools.

In addition to requiring stronger assumptions, due to the technical challenges involved, several of the prior works are only able to establish convergence guarantees that do not match the ideal sub-Gaussian rates. For example, a common approach is to control the relevant quantities across all T iterations of the algorithm via repeated applications of the concentration inequalities, leading to convergence rates that have additional factors that are poly-logarithmic in T. Additionally, achieving noise-adaptive rates that smoothly interpolate between the faster rate in the deterministic setting and the state of the art rate in the stochastic setting is very challenging with existing techniques.

This work aims to contribute to this line of work and overcome the aforementioned challenges. To this end, we introduce a novel generic approach to show convergence with high probability under sub-Gaussian gradient noise. Our approach is very general and flexible, and it can be used both in the convex and non-convex setting. Using our approach, we establish high-probability convergence guarantees for several fundamental settings:

In the *convex setting*, we analyze stochastic mirror descent and stochastic accelerated mirror descent for general optimization domains and Bregman distances, and we analyze the classical algorithms without any changes. These well studied algorithms encompass the main algorithmic frameworks for convex optimization with non-adaptive step sizes (Lan, 2020). Our convergence guarantees scale with only the Bregman distance between the initial point and the optimum, and thus they can leverage good initializations. Our high-probability convergence rates are analogous to known results for convergence in expectation (Juditsky et al., 2011; Lan, 2012). The algorithms are universal for both Lipschitz functions and smooth functions.

In the *non-convex setting*, we analyze the SGD as well as the AdaGrad-Norm algorithm (Ward et al., 2019). Compared to existing works for SGD (Madden et al., 2020; Li & Orabona, 2020), our rates have better dependency on the time horizon and the success probability. For AdaGrad-Norm, our approach allows us to remove the restrictive assumption on the gradients as made in previous work (Kavis et al., 2021). In the full version of our paper[1], using a slightly different technique, we give a high probability convergence of the standard per-coordinate AdaGrad (Duchi et al., 2011). To the best of our knowledge, this is the first result for high probability convergence of AdaGrad.

1.1. Our techniques

Compared to prior works that rely on black-box applications of martingale concentration inequalities such as Freedman's inequality and its extensions (Harvey et al., 2019; Madden et al., 2020), in this work we introduce a "white-box" concentration argument that leverages existing convergence analyses for first-order methods. More precisely, the high-level approach is to define a novel martingale sequence derived from the standard convergence analyses and derive concentration results for this sequence from first principles. By leveraging the structure of the optimization problem, we are able to overcome the aforementioned key difficulties associated with black-box applications of martingale concentration results: these concentration results require certain important conditions on how much the martingale sequence can change, which are generally not a priori satisfied for the natural martingales that arise in optimization. By seamlessly combining the optimization and probability toolkits, we obtain a flexible analysis template that allows us to handle general optimization domains with very large or even unbounded diameter, general objectives that are not globally Lipschitz, and adaptive step sizes.

[1] Available at https://arxiv.org/abs/2302.14843

Our technique is inspired by classical works in concentration inequalities, specifically a type of martingale inequalities where the variance of the martingale difference is bounded by a linear function of the previous value. This technique is first applied to showing high probability convergence by Harvey et al. (2019) in the strongly convex setting. Our proof is inspired by the proof of Theorem 7.3 by Chung & Lu (2006). In each time step with iterate x_t, let $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ be the error in our gradient estimate. Classical proofs of convergence evolve around analyzing the sum of $\langle \xi_t, x^* - x_t \rangle$, which can be viewed as a martingale sequence. Assuming a bounded domain, the concentration of the sum can be shown via classical martingale inequalities. The key new insight is that instead of analyzing this sum, we analyze a related sum where the coefficients decrease over time to account for the fact that we have a looser grip on the distance to the optimal solution as time increases. Nonetheless, the coefficients are kept within a constant factor of each other and the same asymptotic convergence is attained with high probability.

1.2. Related work

Convex optimization: Nemirovski et al. (2009); Lan (2012) establish high probability bounds for stochastic mirror descent and accelerated stochastic mirror descent with sub-Gaussian noise. The rates shown in these works match the best rates known in expectation, but they depend on the Bregman diameter $\max_{x,y \in \mathcal{X}} \mathbf{D}_\psi(x, y)$ of the domain, which can be unbounded. Our work complements the analysis with a novel concentration argument that allows us to establish convergence with respect to the distance $\mathbf{D}_\psi(x^*, x_1)$ from the initial point. Our analysis applies to the general setting considered in (Lan, 2020) and we use the same sub-Gaussian assumption on the noise.

The works by Nazin et al. (2019); Gorbunov et al. (2020) and Parletta et al. (2022) consider the more general setting of bounded variance noise. However, their problem settings are more restricted than ours. Specifically, Nazin et al. (2019) analyze stochastic mirror descent only in the setting where the optimization domain has bounded Bregman diameter. Parletta et al. (2022) analyze modifications of stochastic gradient descent, but only for problems with bounded domains. The work by Gorbunov et al. (2020) for smooth functions and by Gorbunov et al. (2021) for non-smooth functions, analyze stochastic gradient descent and accelerated stochastic gradient descent with gradient clipping, for unconstrained optimization with the ℓ_2 setup. In contrast, our work addresses the sub-Gaussian noise setting but it applies to general optimization, and we analyze the classical stochastic mirror descent and accelerated mirror descent without any modifications and with general Bregman distances and optimization domains.

The algorithm of Davis et al. (2021) is restricted to well-conditioned objectives that are both smooth and strongly convex, and do not apply to general convex optimization. Additionally, compared to classical methods such as SGD and stochastic mirror descent, the proposed algorithm solves an auxiliary optimization problem in each iteration and is thus more computationally expensive. The high-probability convergence of SGD is studied in Kakade & Tewari (2008); Rakhlin et al. (2011); Hazan & Kale (2014); Harvey et al. (2019); Dvurechensky & Gasnikov (2016). These works either assume that the function is strongly convex or the domain is compact. In contrast, our work applies to non-strongly convex optimization with a general domain.

Non-convex optimization: Li & Orabona (2020) demonstrate a high probability bound for an SGD algorithm with momentum while Madden et al. (2020) and Li & Liu (2022) show for the vanilla SGD and generalize to the family of sub-Weibull noise. However, the existing bounds are not optimal, which we improve in our work, using a very different approach. Convergence in high probability of algorithms with adaptive step sizes for non-convex problems has also been studied, for example, by Li & Orabona (2020); Kavis et al. (2021). We note that the algorithm in (Li & Orabona, 2020) is not fully adaptive due to the dependence of the initial step size on the problem parameters, whereas in (Kavis et al., 2021) the gradients or stochastic gradients are required to be uniformly bounded almost surely. Using techniques from Liu et al. (2022) and extending the argument for SGD in Section 4.1, we are able to establish convergence in high probability of the vanilla version of AdaGrad-Norm (Ward et al., 2019; Faw et al., 2022) without any of these additional assumptions. We provide a more detailed comparison with prior work in the subsequent sections.

High probability convergence in the heavy-tail noise regime has also been studied. However, instead of analyzing existing algorithms, most works propose new algorithms which usually require gradient clipping to ensure convergence. Zhang et al. (2020) propose a gradient clipping algorithm that converges in expectation for noise distributions with heavier tails that satisfy the assumption that the p-moments are bounded for $1 < p \leq 2$. Cutkosky & Mehta (2021) propose a more complex clipped SGD algorithm with momentum under the same noise assumption, for which they show a high probability convergence. In another line of works, Zhang & Cutkosky (2022) consider parameter-free algorithms that adapt to the initial distance in the heavy tail regime. In contrast, we focus here on vanilla algorithms that have been successfully employed, including stochastic mirror descent, stochastic gradient descent and AdaGrad-Norm with sub-Gaussian noise, and fill in the missing pieces in the literature. We believe our techniques are general and they may lead to further progress in the heavy tailed setting, and we leave this direction to future work.

2. Preliminaries

We consider the problem $\min_{x\in\mathcal{X}} f(x)$ where $f:\mathbb{R}^d\to\mathbb{R}$ is the objective function and $\mathcal{X}$ is the domain of the problem. In the convex case, we consider the general setting where f is potentially not strongly convex and the domain $\mathcal{X}$ is convex but not necessarily compact. The distance between solutions in $\mathcal{X}$ is measured by a general norm $\|\cdot\|$. Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$. In the non-convex case, we consider the setting where $\mathcal{X}$ is $\mathbb{R}^d$ and $\|\cdot\|$ is the ℓ_2 norm.

In this paper, we use the following assumptions:

(1) Existence of a minimizer: There exists $x^* = \arg\min_{x\in\mathcal{X}} f(x)$.

(2) Unbiased estimator: We assume to have access to a history independent, unbiased gradient estimator $\widehat{\nabla} f(x)$ for any $x\in\mathcal{X}$, that is, $\mathbb{E}\left[\widehat{\nabla} f(x)\mid x\right]=\nabla f(x)$.

(3) Sub-Gaussian noise: $\left\|\widehat{\nabla} f(x)-\nabla f(x)\right\|_*$ is a σ-sub-Gaussian random variable (Definition 2.1).

There are several equivalent definitions of sub-Gaussian random variables up to an absolute constant scaling (see, e.g., Proposition 2.5.2 in (Vershynin, 2018)). For convenience, we use the following property as the definition.

Definition 2.1. A random variable X is σ-sub-Gaussian if

$$\mathbb{E}\left[\exp\left(\lambda^2X^2\right)\right]\le\exp\left(\lambda^2\sigma^2\right)\ \ \forall\lambda \text{ such that } |\lambda|\le\frac{1}{\sigma}.$$

We will also use the following helper lemma whose proof we defer to the Appendix.

Lemma 2.2. *For any $a\ge 0$, $0\le b\le\frac{1}{2\sigma}$ and an σ-sub-Gaussian random variable X,*

$$\mathbb{E}\left[1+b^2X^2+\sum_{i=2}^{\infty}\frac{(aX+b^2X^2)^i}{i!}\right]\le\exp\left(3\left(a^2+b^2\right)\sigma^2\right).$$

When $b=0$, the upper bound improves to $\exp\left(2a^2\sigma^2\right)$.

3. Convex case: Stochastic Mirror Descent and Accelerated Stochastic Mirror Descent

In this section, we analyze the Stochastic Mirror Descent algorithm (Algorithm 1) and Accelerated Stochastic Mirror Descent algorithm (Algorithm 2) for convex optimization. We define the Bregman divergence $\mathbf{D}_\psi(x,y)=\psi(x)-\psi(y)-\langle\nabla\psi(y),x-y\rangle$ where $\psi:\mathbb{R}^d\to\mathbb{R}$ is a 1-strongly convex mirror map with respect to $\|\cdot\|$ on $\mathcal{X}$. We remark that the domain of ψ is defined as $\mathbb{R}^d$ for simplicity, though it is not necessary.

Algorithm 1 Stochastic Mirror Descent Algorithm

Parameters: initial point $x_1\in\mathcal{X}$, step sizes $\{\eta_t\}$, strongly convex mirror map ψ

for $t=1$ to T:

$x_{t+1}=\arg\min_{x\in\mathcal{X}}\left\{\eta_t\left\langle\widehat{\nabla} f(x_t),x\right\rangle+\mathbf{D}_\psi(x,x_t)\right\}$

return $\frac{1}{T}\sum_{t=1}^T x_t$

3.1. Analysis of Stochastic Mirror Descent

The end result of this section is the convergence guarantee of Algorithm 1 for constant step sizes (when the time horizon T is known) and time-varying step sizes (when T is unknown) presented in Theorem 3.1. However, we will emphasize presenting the core idea of our approach, which will serve as the basis for the analysis in subsequent sections. For simplicity, here we consider the non-smooth setting, and assume that f is G-Lipschitz continuous, i.e., we have $\|\nabla f(x)\|_*\le G$ for all $x\in\mathcal{X}$. However, this is not necessary. The analysis for the smooth setting follows via a simple modification to the analysis presented here as well as the analysis for the accelerated setting given in the next section.

Theorem 3.1. *Assume f is G-Lipschitz continuous and satisfies Assumptions (1), (2), (3), with probability at least $1-\delta$, the iterate sequence $(x_t)_{t\ge1}$ output by Algorithm 1 satisfies*

(1) Setting $\eta_t=\sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)T}}$, *then* $\mathbf{D}_\psi(x^*,x_{T+1})\le 4\mathbf{D}_\psi(x^*,x_1)$, *and*

$$\frac{1}{T}\sum_{t=1}^T\left(f(x_t)-f(x^*)\right)\le\frac{4\sqrt{6}}{\sqrt{T}}\sqrt{\mathbf{D}_\psi(x^*,x_1)\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)},$$

(2) Setting $\eta_t=\sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)t}}$, *then* $\mathbf{D}_\psi(x^*,x_{T+1})\le 2(2+\log T)\mathbf{D}_\psi(x^*,x_1)$, *and*

$$\frac{1}{T}\sum_{t=1}^T\left(f(x_t)-f(x^*)\right)\le\frac{2\sqrt{6}}{\sqrt{T}}(2+\log T)\times\sqrt{\mathbf{D}_\psi(x^*,x_1)\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)}.$$

We define $\xi_t:=\widehat{\nabla} f(x_t)-\nabla f(x_t)$ and let $\mathcal{F}_t=\sigma(\xi_1,\dots,\xi_{t-1})$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. The starting point of our analysis is the following inequality that follows from the standard stochastic mirror descent analysis (see, e.g., (Lan, 2020)). We include the proof in the Appendix for completeness.

Lemma 3.2. *(Lan, 2020) For every iteration t, we have*

$$\begin{aligned} A_t &:= \eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - \eta_t^2 G^2 \\ &\quad + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) \\ &\leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \eta_t^2 \left\|\xi_t\right\|_*^2. \end{aligned}$$

We now turn our attention to our main concentration argument. Towards our goal of obtaining a high-probability convergence rate, we analyze the moment generating function for a random variable that is closely related to the left-hand side of the inequality above. We let $\{w_t\}$ be a sequence where $w_t \geq 0$ for all t. We define

$$Z_t = w_t A_t - v_t \mathbf{D}_\psi\left(x^*, x_t\right), \qquad \forall\, 1 \leq t \leq T,$$

where $v_t = 6\sigma^2 \eta_t^2 w_t^2$,

$$\text{and } S_t = \sum_{i=t}^{T} Z_i, \qquad \forall\, 1 \leq t \leq T+1.$$

Before proceeding with the analysis, we provide intuition for our approach. If we consider S_1, we see that it combines the gains in function value gaps with weights given by the sequence $\{w_t\}$ and the losses given by the Bregman divergence terms $\mathbf{D}_\psi\left(x^*, x_t\right)$ with coefficients v_t chosen based on the step size η_t and w_t. The intuition here is that we want to transfer the error from the stochastic error terms on the RHS of Lemma 3.2 into the loss term $v_t \mathbf{D}_\psi\left(x^*, x_t\right)$ then leverage the progression of the Bregman divergence $\mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right)$ to absorb this loss. For the first step, we can do that by setting the coefficient v_t to equalize coefficient of divergence term that will appear from the RHS of Lemma 3.2. For the second step, we can aim at making all the divergence terms telescope, by selecting v_t and w_t such that $w_t + v_t \leq w_{t-1}$ to have a telescoping sum of the terms $w_t \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - w_{t-1}\mathbf{D}_\psi\left(x^*, x_t\right)$. In the end we will obtain a bound for the function value gaps in terms of only the deterministic quantities, namely η_t, w_t, G and the initial distance. In Theorem 3.3, we upper bound the moment generating function of S_1 and derive a set of conditions for the weights $\{w_t\}$ that allow us to absorb the stochastic errors. In Corollary 3.4, we show how to choose the weights $\{w_t\}$ and obtain a convergence rate that matches the standard rates that hold in expectation.

We now give our main concentration argument that bounds the moment generating function of S_t inspired by the proof of Theorem 7.3 in (Chung & Lu, 2006).

Theorem 3.3. *Suppose that $w_t \eta_t^2 \leq \frac{1}{4\sigma^2}$ for every $1 \leq t \leq T$. For every $1 \leq t \leq T+1$, we have*

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \sum_{i=t}^{T} w_i \eta_i^2\right).$$

Proof. We proceed by induction on t. Consider the base case $t = T+1$. We have the inequality holds true trivially. Next, we consider $1 \leq t \leq T$. We have

$$\begin{aligned} &\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] = \mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right]. \end{aligned} \tag{1}$$

We now analyze the inner expectation. Conditioned on $\mathcal{F}_{t+1}$, Z_t is fixed. Using the inductive hypothesis, we obtain

$$\begin{aligned} &\mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \\ &\leq \exp\left(Z_t\right) \exp\left(3\sigma^2 \sum_{i=t+1}^{T} w_i \eta_i^2\right). \end{aligned} \tag{2}$$

Plugging into (1), we obtain

$$\begin{aligned} &\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \\ &\leq \mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right] \exp\left(3\sigma^2 \sum_{i=t+1}^{T} w_i \eta_i^2\right). \end{aligned} \tag{3}$$

By Lemma 3.2

$$\begin{aligned} &\exp\left(Z_t\right) = \exp\Bigg(w_t\big(\eta_t\left(f\left(x_t\right) - f\left(x^*\right)\right) - \eta_t^2 G^2 \\ &+ \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right)\big) - v_t \mathbf{D}_\psi\left(x^*, x_t\right)\Bigg) \\ &\leq \exp\left(w_t \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + w_t \eta_t^2 \left\|\xi_t\right\|_*^2\right) \\ &\quad \times \exp\left(-v_t \mathbf{D}_\psi\left(x^*, x_t\right)\right). \end{aligned}$$

Next, we analyze the first term in the last line of the above inequality in expectation. Let $X_t = \left\langle \xi_t, x^* - x_t \right\rangle$. Using Taylor expansion of e^x, and that $\mathbb{E}\left[X_t \mid \mathcal{F}_t\right] = 0$, we have

$$\begin{aligned} &\mathbb{E}\left[\exp\left(w_t X_t + w_t \eta_t^2 \left\|\xi_t\right\|_*^2\right) \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\Bigg[1 + w_t \eta_t^2 \left\|\xi_t\right\|_*^2 \\ &\quad + \sum_{i=2}^{\infty} \frac{1}{i!}\left(w_t X_t + w_t \eta_t^2 \left\|\xi_t\right\|_*^2\right)^i \mid \mathcal{F}_t\Bigg] \\ &\overset{(a)}{\leq} \mathbb{E}\Bigg[1 + w_t \eta_t^2 \left\|\xi_t\right\|_*^2 \\ &\quad + \sum_{i=2}^{\infty} \frac{1}{i!}\left(w_t \eta_t \left\|x^* - x_t\right\| \left\|\xi_t\right\|_* + w_t \eta_t^2 \left\|\xi_t\right\|_*^2\right)^i \mid \mathcal{F}_t\Bigg] \\ &\overset{(b)}{\leq} \exp\left(3\sigma^2\left(w_t^2 \eta_t^2 \left\|x^* - x_t\right\|^2 + w_t \eta_t^2\right)\right) \\ &\overset{(c)}{\leq} \exp\left(3\sigma^2\left(2 w_t^2 \eta_t^2 \mathbf{D}_\psi\left(x^*, x_t\right) + w_t \eta_t^2\right)\right). \end{aligned} \tag{4}$$

For (a), we use Cauchy-Schwartz and obtain $X_t = \eta_t \left\langle \xi_t, x^* - x_t \right\rangle \leq \eta_t \left\|\xi_t\right\|_* \left\|x^* - x_t\right\|$. For (b), we apply Lemma 2.2 with $X = \left\|\xi_t\right\|_*$, $a = w_t \eta_t \left\|x^* - x_t\right\|$, and

$b^2 = w_t\eta_t^2 \le \frac{1}{4\sigma^2}$. For (c), we use that $\mathbf{D}_\psi(x^*, x_t) \ge \frac{1}{2}\|x^* - x_t\|^2$ from the strong convexity of ψ.

Plugging back into (3) and using that $v_t = 6\sigma^2\eta_t^2 w_t^2$, we obtain the desired inequality

$$\begin{aligned}&\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right]\\ &\le \exp\left(\left(6\sigma^2\eta_t^2 w_t^2 - v_t\right)\mathbf{D}_\psi(x^*, x_t) + 3\sigma^2\sum_{i=t}^{T} w_i\eta_i^2\right)\\ &= \exp\left(3\sigma^2\sum_{i=t}^{T} w_i\eta_i^2\right).\end{aligned}$$

□

Using Markov's inequality, we obtain the following convergence guarantee.

Corollary 3.4. *Suppose the sequence $\{w_t\}$ satisfies the conditions of Theorem 3.3 and that $w_t + 6\sigma^2\eta_t^2 w_t^2 \le w_{t-1}$. For any $\delta > 0$, with probability at least $1-\delta$:*

$$\begin{aligned}&\sum_{t=1}^{T} w_t\eta_t\left(f(x_t) - f(x^*)\right) + w_T\mathbf{D}_\psi(x^*, x_{T+1})\\ &\le w_0\mathbf{D}_\psi(x^*, x_1) + \left(G^2 + 3\sigma^2\right)\sum_{t=1}^{T} w_t\eta_t^2 + \log\left(\frac{1}{\delta}\right).\end{aligned}$$

With the above result in hand, we complete the convergence analysis by showing how to define the sequence $\{w_t\}$ with the desired properties. Below we give the choice of η_t and w_t for fixed step sizes. The choice for time-varying step sizes can be found in Corollary B.1 in the appendix.

Corollary 3.5. *Suppose we run the Stochastic Mirror Descent algorithm with fixed step sizes $\eta_t = \frac{\eta}{\sqrt{T}}$. Let $w_T = \frac{1}{12\sigma^2\eta^2}$ and $w_{t-1} = w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2$ for all $1 \le t \le T$. The sequence $\{w_t\}$ satisfies the conditions required by Corollary 3.4. By Corollary 3.4, for any $\delta > 0$, the following events hold with probability at least $1-\delta$: $\mathbf{D}_\psi(x^*, x_{T+1}) \le 2\mathbf{D}_\psi(x^*, x_1) + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta^2$, and*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) &\le \frac{1}{\sqrt{T}}\frac{2\mathbf{D}_\psi(x^*, x_1)}{\eta}\\ &+ \frac{12}{\sqrt{T}}\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta.\end{aligned}$$

In particular, setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*, x_1)}{6\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)T}}$ we obtain the first case of Theorem 3.1.

Proof. Recall from Corollary 3.4 that the sequence $\{w_t\}$ needs to satisfy the following conditions for all $1 \le t \le T$:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \le w_{t-1}; \quad \text{and} \quad w_t\eta_t^2 \le \frac{1}{4\sigma^2}.$$

Algorithm 2 Accelerated Stochastic Mirror Descent Algorithm (Lan, 2020).

Parameters: initial point $x_0 = y_0 = z_0 \in \mathcal{X}$, step size η, strongly convex mirror map ψ

for $t = 1$ to T:

Set $\alpha_t = \frac{2}{t+1}$

$x_t = (1-\alpha_t)\, y_{t-1} + \alpha_t z_{t-1}$

$z_t = \arg\min_{x\in\mathcal{X}}\left(\eta_t\left\langle\widehat{\nabla} f(x_t), x\right\rangle + \mathbf{D}_\psi(x, z_{t-1})\right)$

$y_t = (1-\alpha_t)\, y_{t-1} + \alpha_t z_t$

return y_T

Let $C = 6\sigma^2\eta^2$. We set $w_T = \frac{1}{C + 6\sigma^2\eta^2} = \frac{1}{2C}$. For $1 \le t \le T$, we set w_t so that the first condition holds with equality

$$w_{t-1} = w_t + 6\sigma^2 w_t^2\eta_t^2 = w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2.$$

We can show by induction that, for every $1 \le t \le T$,

$$w_t \le \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 t}.$$

The base case $t = T$ follows from the definition of w_T. Consider $1 \le t \le T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned}w_{t-1} &= w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2\\ &\le \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 t} + \frac{6\sigma^2\eta^2}{T\left(C + \frac{6}{T}\sigma^2\eta^2 t\right)^2}\\ &\le \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 t}\\ &\quad + \frac{\left(C + \frac{6}{T}\sigma^2\eta^2 t\right) - \left(C + \frac{6}{T}\sigma^2\eta^2(t-1)\right)}{\left(C + \frac{6}{T}\sigma^2\eta^2(t-1)\right)\left(C + \frac{6}{T}\sigma^2\eta^2 t\right)}\\ &= \frac{1}{C + \frac{6}{T}\sigma^2\eta^2(t-1)}\end{aligned}$$

as needed. This fact implies the second condition as follows:

$$w_t\eta_t^2 = w_t\frac{\eta^2}{T} \le \frac{\eta^2}{6\sigma^2\eta^2 t} = \frac{1}{6\sigma^2}.$$

Thus, using Corollary 3.4, $w_T = \frac{1}{2C}$, and $\frac{1}{2C} \le w_t \le \frac{1}{C}$ for all $0 \le t \le T$, we obtain the desired inequalities. □

3.2. Analysis of Accelerated Stochastic Mirror Descent

In this section, we extend the analysis detailed in the previous section to analyze the Accelerated Stochastic Mirror Descent Algorithm (Algorithm 2). We assume that f satisfies the following condition: for all $x, y \in \mathcal{X}$

$$\begin{aligned}f(y) \le f(x) &+ \langle\nabla f(x), y - x\rangle\\ &+ G\|y - x\| + \frac{L}{2}\|y - x\|^2.\end{aligned} \tag{5}$$

Note that L-smooth functions, G-Lipschitz functions, and their sums all satisfy the above condition. Here, we obtain the following guarantees in Theorem 3.6.

Theorem 3.6. *Assume f satisfies Assumptions (1), (2), (3) and condition (5). Then, with probability at least $1-\delta$, the output y_T of the Accelerated Stochastic Mirror Descent algorithm (Algorithm 2) satisfies*

(1) Setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*,z_0)}t}{\sqrt{6}\sqrt{G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)}T^{3/2}}\right\}$, *then* $\mathbf{D}_\psi(x^*,z_T) \le 4\mathbf{D}_\psi(x^*,z_0)$ *and*

$$f(y_T) - f(x^*) \le \frac{16L\mathbf{D}_\psi(x^*,z_0)}{T^2} + \frac{8\sqrt{6}}{\sqrt{T}}\sqrt{\mathbf{D}_\psi(x^*,z_0)\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)}.$$

(2) Setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*,z_0)}}{\sqrt{6}\sqrt{G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)}t^{1/2}}\right\}$, *then* $\mathbf{D}_\psi(x^*,z_T) \le 2(2+\log T)\mathbf{D}_\psi(x^*,z_0)$ *and*

$$f(y_T) - f(x^*) \le \frac{16L\mathbf{D}_\psi(x^*,z_0)}{T^2} + \frac{4\sqrt{6}(2+\log T)}{\sqrt{T}} \times \sqrt{\mathbf{D}_\psi(x^*,z_0)\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)}.$$

We will only highlight the application of the previous analysis here. Define $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$. We start with the inequalities shown in the standard analysis, e.g, from (Lan, 2020) (proof in the Appendix).

Lemma 3.7. *(Lan, 2020) For every iteration t, we have*

$$\begin{aligned} B_t &:= \frac{\eta_t}{\alpha_t}\left(f(y_t) - f(x^*)\right) \\ &\quad - \frac{\eta_t(1-\alpha_t)}{\alpha_t}\left(f(y_{t-1}) - f(x^*)\right) \\ &\quad - \frac{\eta_t^2}{1-L\alpha_t\eta_t}G^2 \\ &\quad + \mathbf{D}_\psi(x^*,z_t) - \mathbf{D}_\psi(x^*,z_{t-1}) \\ &\le \eta_t\langle \xi_t, x^* - z_{t-1}\rangle + \frac{\eta_t^2}{1-L\alpha_t\eta_t}\|\xi_t\|_*^2. \end{aligned}$$

We now turn our attention to our main concentration argument. Similar to the previous section, we define

$$Z_t = w_t B_t - v_t \mathbf{D}_\psi(x^*, z_{t-1}), \qquad \forall\, 1 \le t \le T,$$

where $v_t = 6\sigma^2 w_t^2 \eta_t^2$,

$$\text{and } S_t = \sum_{i=t}^{T} Z_i, \qquad \forall\, 1 \le t \le T+1.$$

Notice that we are following the same steps as in the previous section. By transferring the error terms in the RHS of Lemma 3.7 into the Bregman divergence terms $\mathbf{D}_\psi(x^*, z_{t-1})$, we can absorb them by setting the coefficients appropriately. In the same manner, we can show the following Theorem:

Theorem 3.8. *Suppose that $\frac{w_t\eta_t^2}{1-L\alpha_t\eta_t} \le \frac{1}{4\sigma^2}$ for every $0 \le t \le T$. Then, for every $1 \le t \le T+1$, we have*

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \le \exp\left(3\sigma^2\sum_{i=t}^{T} w_i \frac{\eta_i^2}{1-L\alpha_i\eta_i}\right).$$

Corollary 3.9. *Suppose the sequence $\{w_t\}$ satisfies the conditions of Theorem 3.8. For any $\delta > 0$, the following event holds with probability at least $1-\delta$:*

$$\begin{aligned} &\sum_{t=1}^{T} w_t\left(\frac{\eta_t}{\alpha_t}\left(f(y_t) - f(x^*)\right) - \frac{\eta_t(1-\alpha_t)}{\alpha_t}\left(f(y_{t-1}) - f(x^*)\right)\right) \\ &\quad + w_T\mathbf{D}_\psi(x^*,z_T) \\ &\le w_0\mathbf{D}_\psi(x^*,z_0) \\ &\quad + \left(G^2+3\sigma^2\right)\sum_{t=1}^{T} w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right). \end{aligned}$$

With the above result in hand, we can complete the convergence analysis by showing how to define the sequence $\{w_t\}$ with the desired properties. For the algorithm with known T, we set $\alpha_t = \frac{2}{t+1}$, $\eta_t = \eta t$ for $\eta \le \frac{1}{4L}$, $w_T = \frac{1}{3\sigma^2\eta^2T(T+1)(2T+1)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \le t \le T$. For the algorithm with unknown T, we set $\alpha_t = \frac{2}{t+1}$, $\eta_t = \min\{\frac{t}{4L}, \frac{\eta}{\sqrt{t}}\}$, $w_T = \frac{1}{12\sigma^2\left(\sum_{t=1}^{T}\eta_t^2\right)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \le t \le T$. In the Appendix, we show that these choices have the desired properties (Corollaries B.2 and B.3).

4. Nonconvex case: Stochastic Gradient Descent and AdaGrad-Norm

In this section, we analyze the Stochastic Gradient Descent (SGD) algorithm (Algorithm 3) and the adaptive version, commonly known as AdaGrad-Norm (Algorithm 4) for nonconvex optimization, where we look to find an approximate stationary point of f. Here, we assume that the optimization problem has domain $\mathcal{X} = \mathbb{R}^d$, and that f is an L-smooth function, i.e., the gradients of f is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathbb{R}^d.$$

This implies the following inequality on f at any $x, y \in \mathbb{R}^d$:

$$f(y) - f(x) \le \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2. \qquad (6)$$

Algorithm 3 Stochastic Gradient Descent (SGD)

Parameters: initial point x_1, step sizes $\{\eta_t\}$
for $t = 1$ to T do
  $x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t)$

4.1. Analysis of Stochastic Gradient Descent

In this section, we provide a high probability analysis of SGD (Algorithm 3) that is tighter than previous works. Our main result is presented in Theorem 4.1.

Theorem 4.1. *Assume that f is L-smooth and satisfies Assumptions (1), (2), (3), and let $\Delta_1 := f(x_1) - f(x^*)$. Then, with probability at least $1-\delta$, the iterate sequence $(x_t)_{t\geq 1}$ output by Algorithm 3 satisfies*

(1) Setting $\eta_t = \min\left\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 LT}}\right\}$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L}{T} + 5\sigma\sqrt{\frac{\Delta_1 L}{T}} + \frac{12\sigma^2 \log\frac{1}{\delta}}{T}.$$

(2) Setting $\eta_t = \frac{1}{L\sqrt{t}}$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L + 3\sigma^2(1+\log T) + 12\sigma^2\log\frac{1}{\delta}}{\sqrt{T}}.$$

Comparison with prior work: When the time horizon T is known to the algorithm, by choosing the step size η in part (1) of Theorem 4.1, the bound is adaptive to noise, i.e, when $\sigma = 0$ we recover $O(\frac{1}{T})$ convergence rate of the (deterministic) gradient descent algorithm. Notice that the bound in this case does not have a $\log T$ term incurred. When T is unknown, the extra $\log T$ appears as a result of setting a time-varying step size $\eta_t = \frac{1}{L\sqrt{t}}$. This $\log T$ appears as an additive term to the $\log\frac{1}{\delta}$ term, as opposed to being multiplicative, i.e, $\log T \log\frac{1}{\delta}$ as in previous works (Li & Orabona, 2020; Madden et al., 2020; Li & Liu, 2022).

To proceed to the analysis, we define for $t \geq 1$

$$\Delta_t := f(x_t) - f(x^*); \quad \xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t).$$

We let $\mathcal{F}_t := \sigma(\xi_1, \ldots, \xi_{t-1})$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. The following lemma serves as the fundamental step of our analysis, the proof of which can be found in the appendix.

Lemma 4.2. *For $t \geq 1$, we have*

$$C_t := \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla f(x_t)\|^2 + \Delta_{t+1} - \Delta_t \leq \left(L\eta_t^2 - \eta_t\right)\langle \nabla f(x_t), \xi_t\rangle + \frac{L\eta_t^2}{2}\|\xi_t\|^2. \quad (7)$$

Now we can follow the similar concentration argument from the convex setting. The difference now is the error term in the RHS of (7) can be transferred into the gradient term $\|\nabla f(x_t)\|^2$ instead of a function value gap term. This actually makes things easier since this term can be readily absorbed by the gradient term in C_t, and we do not have to carefully impose an additional condition on w_t to make a telescoping sum. For $w_t \geq 0$, we define

$$Z_t = w_t C_t - v_t\|\nabla f(x_t)\|^2, \qquad \forall\, 1 \leq t \leq T,$$
$$\text{where } v_t = 3\sigma^2 w_t^2\eta_t^2(\eta_t L - 1)^2,$$
$$\text{and } S_t = \sum_{i=t}^{T} Z_i, \qquad \forall\, 1 \leq t \leq T+1.$$

Using the same technique as in the previous section, we can prove the following key inequality.

Theorem 4.3. *Suppose that η_t and w_t satisfy $0 \leq w_t\eta_t^2 L \leq \frac{1}{2\sigma^2}$ for all $1 \leq t \leq T$. Then*

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2\sum_{s=t}^{T}\frac{w_t\eta_t^2 L}{2}\right). \quad (8)$$

Markov's inequality gives us the following guarantee.

Corollary 4.4. For all $1 \leq t \leq T$, if $\eta_t L \leq 1$ and $0 \leq w_t\eta_t^2 L \leq \frac{1}{2\sigma^2}$, then

$$\sum_{t=1}^{T}\left[w_t\eta_t\left(1 - \frac{\eta_t L}{2}\right) - v_t\right]\|\nabla f(x_t)\|^2 + w_T\Delta_{T+1} \leq w_1\Delta_1 + \sum_{t=2}^{T}(w_t - w_{t-1})\Delta_t + 3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2} + \log\frac{1}{\delta}. \quad (9)$$

Equipped with Lemmas 4.2 and 4.3, we are ready to prove Theorem 4.1 by specifying the choice of w_t that satisfy the condition of Lemma 4.3. In the first case, we choose $\eta_t = \eta$, $w_t = w = \frac{1}{6\sigma^2\eta}$ where $\eta = \min\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 LT}}\}$. In the second case, we set $\eta_t = \frac{\eta}{\sqrt{t}}$ and $w_t = w = \frac{1}{6\sigma^2\eta}$, where $\eta = \frac{1}{L}$. We show the full proof in the Appendix.

4.2. Analysis of AdaGrad-Norm

In this section, we show that AdaGrad-Norm (Algorithm 4) converges with high probability under minimal assumptions. Our main result is presented in Theorem 4.5.

Algorithm 4 AdaGrad-Norm

Parameters: x_1, η, b_0
for $t = 1$ to T do
  $b_t = \sqrt{b_0^2 + \sum_{i=1}^{t}\left\|\widehat{\nabla} f(x_i)\right\|^2}$
  $x_{t+1} = x_t - \frac{\eta}{b_t}\widehat{\nabla} f(x_i)$

Theorem 4.5. *Assume f is L-smooth and satisfies Assumptions (1), (2), (3). With probability at least $1-3\delta$, the iterate sequence $(x_t)_{t\geq 1}$ output by Algorithm 4 satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq \frac{4}{T}\sqrt{4\sigma^2 T + r(\delta)}\left(\eta L \log\frac{4\sigma^2 T + r(\delta)}{b_0^2} + g(\delta)\right) + \frac{1}{T}\left(32\eta L\log\frac{64\eta L}{b_0} + 16 g(\delta)\right)^2,$$

where $r(\delta) = 2b_0^2 + 4\sigma^2\log\frac{1}{\delta} = O(1+\sigma^2\log\frac{1}{\delta})$ *and* $g(\delta) = \frac{\Delta_1}{\eta} + \frac{\|\nabla f(x_1)\|^2}{4b_0^2} + \frac{8\sigma^2}{b_0}\left(1+\frac{\eta L}{b_0}\right)\log\frac{1}{\delta} + \frac{\eta L\sigma^2}{b_0^2}\left(1+\log\frac{T}{\delta}\right) = O(1+\sigma^2\log\frac{T}{\delta})$.

Comparison with prior work: (Ward et al., 2019; Faw et al., 2022) show the convergence of this algorithm with polynomial dependency on $\frac{1}{\delta}$ where $1-\delta$ is the success probability. The latter relaxes several assumptions made in the former, including the boundedness of the gradients and noise variance. When assuming a sub-Gaussian noise, (Kavis et al., 2021) show a convergence in high probability, but still assume that the gradients are bounded. We remove this assumption and establish the convergence of Algorithm 4 in the following theorem. We highlight that the bound in 4.5 is adaptive to noise. When $\sigma = 0$, we obtain the $O(\frac{1}{T})$ convergence of the deterministic AdaGrad-Norm.

We next give an overview of the technique. We will start from Lemma 4.6 (proof in the Appendix) and proceed to bound each term in the RHS of (10). In contrast to the techniques used in (Kavis et al., 2021), in which they multiply both sides of (23) by b_t to separate b_t from the term $\langle\nabla f(x_t),\xi_t\rangle$, we rely on the insight from (Liu et al., 2022) and multiply by $\frac{b_t}{2b_t-b_0}$. This factor is but a small deviation from a constant, which helps us obtain a coefficient for $\langle\nabla f(x_t),\xi_t\rangle$ that depends on b_{t-1}. This makes the term $\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}$ a sub-Gaussian random variable. To bound $\sum_{t=1}^{T}\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}$, we follow an argument similar to the proof of Lemma 4.3. Finally, by bounding $\sum_{t=1}^{T}\|\xi_t\|^2$ via a simple concentration argument, we can obtain a relationship between b_T and $\sum_{t=1}^{T}\|\nabla f(x_t)\|^2$. Combining this with Lemma 4.6, we arrive at Theorem 4.5 via a self-bounding argument as used in (Li & Orabona, 2019).

Lemma 4.6. *For $t\geq 1$, let $\xi_t = \widehat{\nabla} f(x_t) - \nabla f(x_t)$, and $M_t = \max_{i\leq t}\|\xi_i\|^2$ then we have*

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2b_t - b_0} \leq \frac{\eta L M_T}{b_0^2} + \frac{\Delta_1}{\eta} - \frac{b_T\Delta_{T+1}}{\eta(2b_T - b_0)} + \frac{\eta L}{2}\log\frac{b_T^2}{b_0^2} - \sum_{t=1}^{T}\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}. \quad (10)$$

Now, notice that $\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}$ follows a sub-Gaussian distribution with mean 0, we can obtain a bound for $\sum_{t=1}^{T} -\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}$ in the next lemma. The choice of the coefficient w is crucial but will be specified later.

Lemma 4.7. *For any $w>0$, with probability at least $1-\delta$*

$$\sum_{t=1}^{T} -\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0} \leq \frac{4w\eta^2L^2\sigma^2}{b_0^2}\log\frac{b_T^2}{b_0^2} + \sum_{t=2}^{T}\frac{4w\sigma^2\|\nabla f(x_{t-1})\|^2}{(2b_{t-1}-b_0)^2} + \frac{2w\sigma^2\|\nabla f(x_1)\|^2}{b_0^2} + \frac{1}{w}\log\frac{1}{\delta}. \quad (11)$$

Returning to Lemma 4.6, by choosing an appropriate coefficient w in Lemma 4.7, we can use a fraction of the LHS of (10) to cancel out the term $\sum_{t=2}^{T}\frac{4w\sigma^2\|\nabla f(x_{t-1})\|^2}{(2b_{t-1}-b_0)^2}$ in (11). It is also known that with probability at least $1-\delta$, $M_T \leq \sigma^2\left(1+\log\frac{T}{\delta}\right)$ (Li & Orabona, 2020; Liu et al., 2022). Further, we have a relationship between $\sum_{t=1}^{T}\|\nabla f(x_t)\|^2$ and b_T:

$$b_T \leq \sqrt{b_0^2 + 2\sum_{t=1}^{T}\|\xi_t\|^2 + \sum_{t=1}^{T}2\|\nabla f(x_t)\|^2}. \quad (12)$$

The term $\sum_{t=1}^{T}\|\xi_t\|^2$ can be bounded by $\sigma^2 T + \sigma^2\log\frac{1}{\delta}$ with high probability as in Lemma C.1.

Finally for the LHS of 4.6, we have $\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2b_t-b_0} \geq \frac{1}{b_T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2$. Now we can solve a system combining the two relationships between $\sum_{t=1}^{T}\|\nabla f(x_t)\|^2$ and b_T to obtain the desired bound.

5. Conclusion

In this work, we present a generic approach to prove high probability convergence of stochastic gradient methods under sub-Gaussian noise. In the convex case, we show high probability bounds for (accelerated) SMD that depend on the distance from the initial solution to the optimal solution and do not require the bounded domain or bounded Bregman divergence assumption. In the non-convex case, we apply the same approach and obtain a high probability bound for SGD that improves over existing works. We also show that the boundedness of the gradients can be removed when showing high probability convergence of AdaGrad-Norm.

Acknowledgement

TDN and AE were supported in part by NSF CAREER grant CCF-1750333, NSF grant III-1908510, and an Alfred P. Sloan Research Fellowship. THN and HN were supported by NSF CAREER grant CCF-1750716.

References

Chung, F. and Lu, L. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1): 79–127, 2006.

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

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. From low probability to high confidence in stochastic convex optimization. *The Journal of Machine Learning Research*, 22(1):2237–2274, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Dvurechensky, P. and Gasnikov, A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. *arXiv preprint arXiv:2202.05791*, 2022.

Gorbunov, E., Danilova, M., and Gasnikov, A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Kakade, S. M. and Tewari, A. On the generalization ability of online strongly convex programming algorithms. *Advances in Neural Information Processing Systems*, 21, 2008.

Kavis, A., Levy, K. Y., and Cevher, V. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *International Conference on Learning Representations*, 2021.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Lan, G. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

Li, S. and Liu, Y. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pp. 12931–12963. PMLR, 2022.

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Liu, Z., Nguyen, T. D., Ene, A., and Nguyen, H. L. On the convergence of adagrad on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. *arXiv preprint arXiv:2209.14827*, 2022.

Madden, L., Dall'Anese, E., and Becker, S. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

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

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

Parletta, D. A., Paudice, A., Pontil, M., and Salzo, S. High probability bounds for stochastic subgradient schemes with heavy tailed noise. *arXiv preprint arXiv:2208.08567*, 2022.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

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

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

Zhang, J. and Cutkosky, A. Parameter-free regret in high probability with heavy tails. *arXiv preprint arXiv:2210.14355*, 2022.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

A. Proof of Lemma 2.2

Proof. Consider two cases either $a \geq 1/(2\sigma)$ or $a \leq 1/(2\sigma)$. First suppose $a \geq 1/(2\sigma)$. We use the inequality $uv \leq \frac{u^2}{4} + v^2$ here to first obtain

$$\left(aX + b^2X^2\right)^i \leq \left|aX + b^2X^2\right|^i \leq \left(a\left|X\right| + b^2X^2\right)^i \leq \left(\frac{1}{4\sigma^2}X^2 + a^2\sigma^2 + b^2X^2\right)^i.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}\left[1 + b^2X^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(aX + b^2X^2\right)^i\right] &\leq \mathbb{E}\left[1 + b^2X^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\frac{1}{4\sigma^2}X^2 + a^2\sigma^2 + b^2X^2\right)^i\right] \\
&= \mathbb{E}\left[b^2X^2 + \exp\left(\left(\frac{1}{4\sigma^2} + b^2\right)X^2 + a^2\sigma^2\right) - \left(\frac{1}{4\sigma^2} + b^2\right)X^2 - a^2\sigma^2\right] \\
&= \mathbb{E}\left[\exp\left(\left(\frac{1}{4\sigma^2} + b^2\right)X^2 + a^2\sigma^2\right) - \frac{1}{4\sigma^2}X^2 - a^2\sigma^2\right] \\
&\leq \exp\left(\left(\frac{1}{4\sigma^2} + b^2\right)\sigma^2 + a^2\sigma^2\right) \\
&\leq \exp\left(2a^2\sigma^2 + b^2\sigma^2\right) \\
&\leq \exp\left(3\left(a^2 + b^2\right)\sigma^2\right).
\end{aligned}$$

Next, let $c = \max(a, b) \leq 1/(2\sigma)$. We have

$$\begin{aligned}
\mathbb{E}\left[1 + b^2X^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(aX + b^2X^2\right)^i\right] &= \mathbb{E}\left[\exp\left(aX + b^2X^2\right) - aX\right] \\
&\leq \mathbb{E}\left[\left(aX + \exp\left(a^2X^2\right)\right)\exp\left(b^2X^2\right) - aX\right] \\
&= \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + aX\left(\exp\left(b^2X^2\right) - 1\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + c\left|X\right|\left(\exp\left(c^2X^2\right) - 1\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + \exp\left(2c^2X^2\right) - 1\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2 + 2c^2\right)X^2\right)\right] \\
&\leq \exp\left(\left(a^2 + b^2 + 2c^2\right)\sigma^2\right) \\
&\leq \exp\left(3\left(a^2 + b^2\right)\sigma^2\right).
\end{aligned}$$

In the first inequality, we use the inequality $e^x - x \leq e^{x^2}\forall x$. In the third inequality, we use $x\left(e^{x^2} - 1\right) \leq e^{2x^2} - 1\ \forall x$. This inequality can be proved with the Taylor expansion.

$$\begin{aligned}
x\left(e^{x^2} - 1\right) &= \sum_{i=1}^{\infty}\frac{1}{i!}x^{2i+1} \leq \sum_{i=1}^{\infty}\frac{1}{i!}\frac{x^{2i} + x^{2i+2}}{2} \\
&= \frac{x^2}{2} + \sum_{i=2}^{\infty}\left(\frac{1+i}{2i!}\right)x^{2i} \leq \frac{x^2}{2} + \sum_{i=2}^{\infty}\left(\frac{2^i}{i!}\right)x^{2i} \\
&\leq e^{2x^2} - 1.
\end{aligned}$$

The case when $b = 0$ simply follows from the above proof. □

B. Missing Proofs from Section 3

B.1. Stochastic Mirror Descent

Proof of Lemma 3.2. By the optimality condition, we have

$$\left\langle \eta_t \widehat{\nabla} f(x_t) + \nabla_x \mathbf{D}_\psi \left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle \geq 0$$

and thus

$$\left\langle \eta_t \widehat{\nabla} f(x_t), x_{t+1} - x^* \right\rangle \leq \left\langle \nabla_x \mathbf{D}_\psi \left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle.$$

Note that

$$\begin{aligned}\left\langle \nabla_x \mathbf{D}_\psi \left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle &= \left\langle \nabla \psi \left(x_{t+1}\right) - \nabla \psi \left(x_t\right), x^* - x_{t+1} \right\rangle \\ &= \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x_{t+1}, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right)\end{aligned}$$

and thus

$$\begin{aligned}\eta_t \left\langle \widehat{\nabla} f(x_t), x_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \mathbf{D}_\psi \left(x_{t+1}, x_t\right) \\ &\leq \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2,\end{aligned}$$

where we have used that $\mathbf{D}_\psi \left(x_{t+1}, x_t\right) \geq \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2$ by the strong convexity of ψ.

By convexity,

$$f\left(x_t\right) - f\left(x^*\right) \leq \left\langle \nabla f\left(x_t\right), x_t - x^* \right\rangle = \left\langle \xi_t, x^* - x_t \right\rangle + \left\langle \widehat{\nabla} f\left(x_t\right), x_t - x^* \right\rangle.$$

Combining the two inequalities, we obtain

$$\begin{aligned}&\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \mathbf{D}_\psi \left(x^*, x_t\right) \\ &\leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \eta_t \left\langle \widehat{\nabla} f(x_t), x_t - x_{t+1} \right\rangle - \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2 \\ &\leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \frac{\eta_t^2}{2} \left\|\widehat{\nabla} f(x_t)\right\|_*^2.\end{aligned}$$

Using the triangle inequality and the bounded gradient assumption $\left\|\nabla f(x)\right\|_* \leq G$, we obtain

$$\left\|\widehat{\nabla} f(x_t)\right\|_*^2 = \left\|\xi_t + \nabla f(x_t)\right\|_*^2 \leq 2 \left\|\xi_t\right\|_*^2 + 2 \left\|\nabla f(x_t)\right\|_*^2 \leq 2 \left(\left\|\xi_t\right\|_*^2 + G^2\right).$$

Thus

$$\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \mathbf{D}_\psi \left(x^*, x_t\right) \leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \eta_t^2 \left(\left\|\xi_t\right\|_*^2 + G^2\right)$$

as needed. □

Proof of Corollary 3.4. Let

$$K = 3\sigma^2 \sum_{t=1}^T w_t \eta_t^2 + \log\left(\frac{1}{\delta}\right).$$

By Theorem 3.3 and Markov's inequality, we have

$$\begin{aligned}\Pr\left[S_1 \geq K\right] &\leq \Pr\left[\exp\left(S_1\right) \geq \exp\left(K\right)\right] \\ &\leq \exp\left(-K\right) \mathbb{E}\left[\exp\left(S_1\right)\right] \\ &\leq \exp\left(-K\right) \exp\left(3\sigma^2 \sum_{t=1}^T w_t \eta_t^2\right) \\ &= \delta.\end{aligned}$$

Note that since $v_t + w_t \leq w_{t-1}$

$$\begin{aligned}
S_1 &= \sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} w_t\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2 \sum_{t=1}^{T} w_t\eta_t^2 + \sum_{t=1}^{T} \left(w_t \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \left(v_t + w_t\right)\mathbf{D}_\psi\left(x^*, x_t\right)\right) \\
&\geq \sum_{t=1}^{T} w_t\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2 \sum_{t=1}^{T} w_t\eta_t^2 + \sum_{t=1}^{T} \left(w_t \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - w_{t-1}\mathbf{D}_\psi\left(x^*, x_t\right)\right) \\
&= \sum_{t=1}^{T} w_t\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2 \sum_{t=1}^{T} w_t\eta_t^2 + w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) - w_0 \mathbf{D}_\psi\left(x^*, x_1\right).
\end{aligned}$$

Therefore, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T} w_t\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq w_0 \mathbf{D}_\psi\left(x^*, x_1\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_{t+1}\eta_t^2 + \log\left(\frac{1}{\delta}\right).$$

□

Next we extend the analysis to the setting where the T is not known and we use the step sizes $\eta_t = \frac{\eta}{\sqrt{t}}$ to complete the proof of Theorem 3.1.

Corollary B.1. *Suppose we run the Stochastic Mirror Descent algorithm with time-varying step sizes $\eta_t = \frac{\eta}{\sqrt{t}}$. Let $w_T = \frac{1}{12\sigma^2\eta^2\left(\sum_{t=1}^{T} \frac{1}{t}\right)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \leq t \leq T$. The sequence $\{w_t\}$ satisfies the conditions required by Corollary 3.4. By Corollary 3.4, for any $\delta > 0$, the following events hold with probability at least $1 - \delta$: $\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 2\mathbf{D}_\psi\left(x^*, x_1\right) + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta^2(1 + \log T)$, and*

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \left(f\left(x_t\right) - f\left(x^*\right)\right) &\leq \frac{1}{\sqrt{T}} \frac{2\mathbf{D}_\psi\left(x^*, x_1\right)}{\eta} \\
&\quad + \frac{12}{\sqrt{T}}\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta(1 + \log T).
\end{aligned}$$

In particular, setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*, x_1)}{6\left(G^2 + \sigma^2\left(1 + \ln\left(\frac{1}{\delta}\right)\right)\right)t}}$ we obtain the second case of Theorem 3.1.

Proof of Corollary B.1. Recall from Corollary 3.4 that the sequence $\{w_t\}$ needs to satisfy the following conditions for all $1 \leq t \leq T$:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1} \quad \text{and} \quad w_t\eta_t^2 \leq \frac{1}{4\sigma^2}.$$

Let $M_t = 6\sigma^2 \sum_{i=1}^{t} \eta_i^2$ and $C = M_T = 6\sigma^2\eta^2\left(\sum_{t=1}^{T} \frac{1}{t}\right)$. We set $w_T = \frac{1}{C + M_T}$. For $1 \leq t \leq T$, we set w_t so that the first condition holds with equality

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2.$$

We can show by induction that, for every $1 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C + M_t}.$$

The base case $t = T$ follows from the definition of w_T. Consider $1 \leq t \leq T$. Using the definition of w_t and the inductive

hypothesis, we obtain

$$\begin{aligned} w_{t-1} &= w_t + 6\sigma^2\eta_t^2 w_t^2 \\ &\leq \frac{1}{C+M_t} + \frac{6\sigma^2\eta_t^2}{(C+M_t)^2} \\ &\leq \frac{1}{C+M_t} + \frac{(C+M_t)-(C+M_{t-1})}{(C+M_t)(C+M_{t-1})} \\ &= \frac{1}{C+M_{t-1}} \end{aligned}$$

as needed.

Using this fact, we now show that $\{w_t\}$ satisfies the second condition. For every $1 \leq t \leq T$, we have

$$w_t\eta_t^2 \leq \frac{\eta_t^2}{C} \leq \frac{\eta_t^2}{6\sigma^2\eta_t^2} = \frac{1}{6\sigma^2}$$

as needed.

Thus, by Corollary 3.4, with probability $\geq 1-\delta$, we have

$$\sum_{t=1}^{T} w_t\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq w_0 \mathbf{D}_\psi\left(x^*, x_1\right) + \left(G^2+3\sigma^2\right)\sum_{t=1}^{T} w_t\eta_t^2 + \log\left(\frac{1}{\delta}\right).$$

Note that $w_T = \frac{1}{2C}$ and $\frac{1}{2C} \leq w_t \leq \frac{1}{C}$ for all $1 \leq t \leq T$. Thus, we obtain

$$\frac{1}{2C}\eta_T \sum_{t=1}^{T} \left(f\left(x_t\right) - f\left(x^*\right)\right) + \frac{1}{2C}\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq \frac{1}{C}\mathbf{D}_\psi\left(x^*, x_1\right) + \left(G^2+3\sigma^2\right)\frac{1}{C}\sum_{t=1}^{T}\eta_t^2 + \log\left(\frac{1}{\delta}\right).$$

Plugging in $\eta_t = \frac{\eta}{\sqrt{t}}$ and simplifying, we obtain

$$\begin{aligned} \frac{\eta}{\sqrt{T}}\sum_{t=1}^{T}\left(f\left(x_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, x_{T+1}\right) &\leq 2\mathbf{D}_\psi\left(x^*, x_1\right) + \left(2G^2+6\sigma^2\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right) + 2C\log\left(\frac{1}{\delta}\right) \\ &= 2\mathbf{D}_\psi\left(x^*, x_1\right) + \left(2G^2 + 6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right). \end{aligned}$$

Thus, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right) - f\left(x^*\right)\right) \leq \frac{1}{\sqrt{T}}\left(\frac{2\mathbf{D}_\psi\left(x^*, x_1\right)}{\eta} + \left(2G^2 + 6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta\left(\sum_{t=1}^{T}\frac{1}{t}\right)\right),$$

and

$$\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 2\mathbf{D}_\psi\left(x^*, x_1\right) + \left(2G^2 + 6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right).$$

□

B.2. Accelerated Stochastic Mirror Descent

Proof of Lemma 3.7. Starting with smoothness, we obtain

$$
\begin{aligned}
f(y_t) &\leq f(x_t) + \langle \nabla f(x_t), y_t - x_t \rangle + G\|y_t - x_t\| + \frac{L}{2}\|y_t - x_t\|^2 \ \forall x \in \mathcal{X} \\
&= f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle + \langle \nabla f(x_t), y_t - y_{t-1} \rangle + G\|y_t - x_t\| + \frac{L}{2}\|y_t - x_t\|^2 \\
&= (1-\alpha_t)\underbrace{\left(f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle\right)}_{\text{convexity}} + \alpha_t \underbrace{\left(f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle\right)}_{\text{convexity}} \\
&+ \alpha_t \langle \nabla f(x_t), z_t - y_{t-1} \rangle + G\|y_t - x_t\| + \frac{L}{2}\|y_t - x_t\|^2 \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x_t) + \alpha_t \langle \nabla f(x_t), z_t - x_t \rangle + G \underbrace{\|y_t - x_t\|}_{=\alpha_t\|z_t - z_{t-1}\|} + \frac{L}{2}\underbrace{\|y_t - x_t\|^2}_{=\alpha_t^2\|z_t - z_{t-1}\|^2} \\
&= (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x_t) + \alpha_t \langle \nabla f(x_t), z_t - x_t \rangle + G\alpha_t\|z_t - z_{t-1}\| + \frac{L}{2}\alpha_t^2\|z_t - z_{t-1}\|^2 .
\end{aligned}
$$

By the optimality condition for z_t,

$$
\eta_t \left\langle \widehat{\nabla} f(x_t), z_t - x^* \right\rangle \leq \langle \nabla_x \mathbf{D}_\psi(z_t, z_{t-1}), x^* - z_t \rangle = \mathbf{D}_\psi(x^*, z_{t-1}) - \mathbf{D}_\psi(z_t, z_{t-1}) - \mathbf{D}_\psi(x^*, z_t) .
$$

Rearranging, we obtain

$$
\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1}) + \mathbf{D}_\psi(z_t, z_{t-1}) \leq \eta_t \left\langle \widehat{\nabla} f(x_t), x^* - z_t \right\rangle = \eta_t \langle \nabla f(x_t) + \xi_t, x^* - z_t \rangle .
$$

By combining the two inequalities, we obtain

$$
\begin{aligned}
&f(y_t) + \frac{\alpha_t}{\eta_t}\left(\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1}) + \mathbf{D}_\psi(z_t, z_{t-1})\right) \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t \underbrace{\left(f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle\right)}_{\text{convexity}} \\
&+ G\alpha_t\|z_t - z_{t-1}\| + \frac{L}{2}\alpha_t^2\|z_t - z_{t-1}\|^2 + \alpha_t \langle \xi_t, x^* - z_t \rangle \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x^*) + G\alpha_t\|z_t - z_{t-1}\| + \frac{L}{2}\alpha_t^2\|z_t - z_{t-1}\|^2 + \alpha_t \langle \xi_t, x^* - z_t \rangle .
\end{aligned}
$$

Subtracting $f(x^*)$ from both sides, rearranging, and using that $\mathbf{D}_\psi(z_t, z_{t-1}) \geq \frac{1}{2}\|z_t - z_{t-1}\|^2$, we obtain

$$
\begin{aligned}
&f(y_t) - f(x^*) + \frac{\alpha_t}{\eta_t}\left(\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1})\right) \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_t \rangle + G\alpha_t\|z_t - z_{t-1}\| - \alpha_t\frac{1 - L\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&= (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \alpha_t \langle \xi_t, z_t - z_{t-1} \rangle \\
&+ G\alpha_t\|z_t - z_{t-1}\| - \alpha_t\frac{1 - L\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \alpha_t\|z_t - z_{t-1}\|\left(\|\xi_t\|_* + G\right) - \alpha_t\frac{1 - L\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \frac{\alpha_t\eta_t}{2(1 - L\alpha_t\eta_t)}\left(\|\xi_t\|_* + G\right)^2 .
\end{aligned}
$$

Finally, we divide by $\frac{\alpha_t}{\eta_t}$, and obtain

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*,z_t\right)-\mathbf{D}_\psi\left(x^*,z_{t-1}\right)\\
&\leq \frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)+\eta_t\left\langle\xi_t,x^*-z_{t-1}\right\rangle+\frac{\eta_t^2}{2\left(1-L\alpha_t\eta_t\right)}\left(\left\|\xi_t\right\|_*+G\right)^2\\
&\leq \frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)+\eta_t\left\langle\xi_t,x^*-z_{t-1}\right\rangle+\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left(\left\|\xi_t\right\|_*^2+G^2\right).
\end{aligned}$$

□

Proof of Theorem 3.8. We proceed by induction on t. Consider the base case $t=T+1$, the inequality trivially holds. Next, we consider $t\leq T$. We have

$$\mathbb{E}\left[\exp\left(S_t\right)\mid\mathcal{F}_t\right]=\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right)\mid\mathcal{F}_t\right]=\mathbb{E}\left[\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right)\mid\mathcal{F}_{t+1}\right]\mid\mathcal{F}_t\right]. \tag{13}$$

We now analyze the inner expectation. Conditioned on $\mathcal{F}_{t+1}$, Z_t is fixed. Using the inductive hypothesis, we obtain

$$\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right)\mid\mathcal{F}_{t+1}\right]\leq\exp\left(Z_t\right)\exp\left(3\sigma^2\sum_{i=t+1}^{T}w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right). \tag{14}$$

Let $X_t=\eta_t\left\langle\xi_t,x^*-z_{t-1}\right\rangle$. By Lemma 3.7, we have

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)-\frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)-\frac{\eta_t^2}{1-L\alpha_t\eta_t}G^2\\
&\quad+\mathbf{D}_\psi\left(x^*,z_t\right)-\mathbf{D}_\psi\left(x^*,z_{t-1}\right)\\
&\leq X_t+\frac{\eta_t^2}{\left(1-L\alpha_t\eta_t\right)}\left\|\xi_t\right\|_*^2,
\end{aligned}$$

and thus

$$Z_t\leq w_tX_t+w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2-v_t\mathbf{D}_\psi\left(x^*,z_{t-1}\right).$$

Plugging into (14), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right)\mid\mathcal{F}_{t+1}\right]\\
&\leq\exp\left(w_tX_t-v_t\mathbf{D}_\psi\left(x^*,z_{t-1}\right)+w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2+3\sigma^2\sum_{i=t+1}^{T}w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right).
\end{aligned}$$

Plugging into (13), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(S_t\right)\mid\mathcal{F}_t\right]\\
&\leq\exp\left(-v_t\mathbf{D}_\psi\left(x^*,z_{t-1}\right)+3\sigma^2\sum_{i=t+1}^{T}w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right)\mathbb{E}\left[\exp\left(w_tX_t+w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right)\mid\mathcal{F}_t\right].
\end{aligned} \tag{15}$$

Next, we analyze the expectation on the RHS of the above inequality. We have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(w_t X_t + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{\infty}\frac{1}{i!}\left(w_t X_t + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right)^i \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[1 + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(w_t X_t + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right)^i \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[1 + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(w_t\eta_t\left\|x^*-z_{t-1}\right\|\left\|\xi_t\right\|_* + w_t \frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right)^i \mid \mathcal{F}_t\right] \\
&\leq \exp\left(3\left(w_t^2\eta_t^2\left\|x^*-z_{t-1}\right\|^2 + w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\right)\sigma^2\right) \\
&\leq \exp\left(3\left(2w_t^2\eta_t^2\mathbf{D}_\psi\left(x^*, z_{t-1}\right) + w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\right)\sigma^2\right). \qquad (16)
\end{aligned}$$

On the first line we used the Taylor expansion of e^x, and on the second line we used that $\mathbb{E}\left[X_t \mid \mathcal{F}_t\right] = 0$. On the third line, we used Cauchy-Schwartz and obtained

$$X_t = \eta_t\left\langle\xi_t, x^* - z_{t-1}\right\rangle \leq \eta_t\left\|\xi_t\right\|_*\left\|x^*-z_{t-1}\right\|.$$

On the fourth line, we applied Lemma 2.2 with $X = \left\|\xi_t\right\|_*$, $a = w_t\eta_t\left\|x^*-z_{t-1}\right\|$, and $b^2 = w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t} \leq \frac{1}{4\sigma^2}$. On the fifth line, we used that $\mathbf{D}_\psi\left(x^*, z_{t-1}\right) \geq \frac{1}{2}\left\|x^*-z_{t-1}\right\|^2$, which follows from the strong convexity of ψ.

Plugging in (16) into (15) and using that $v_t = 6\sigma^2 w_t^2\eta_t^2$, we obtain

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2\sum_{i=t}^{T} w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right)$$

as needed. □

Proof of Corollary 3.9. Let

$$K = 3\sigma^2\sum_{t=1}^{T} w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

By Theorem 3.8 and Markov's inequality, we have

$$\begin{aligned}
\Pr\left[S_1 \geq K\right] &\leq \Pr\left[\exp\left(S_1\right) \geq \exp\left(K\right)\right] \\
&\leq \exp\left(-K\right)\mathbb{E}\left[\exp\left(S_1\right)\right] \\
&\leq \exp\left(-K\right)\exp\left(3\sigma^2\sum_{t=1}^{T} w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\right) \\
&= \delta.
\end{aligned}$$

Note that since $v_t + w_t \leq w_{t-1}$

$$\begin{aligned}
S_1 &= \sum_{t=1}^{T} Z_t \\
&= \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + \sum_{t=1}^{T} w_t \mathbf{D}_\psi\left(x^*, z_t\right) - \left(v_t + w_t\right) \mathbf{D}_\psi\left(x^*, z_{t-1}\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} \\
&\geq \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + \sum_{t=1}^{T} w_t \mathbf{D}_\psi\left(x^*, z_t\right) - w_{t-1} \mathbf{D}_\psi\left(x^*, z_{t-1}\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} \\
&= \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + w_T \mathbf{D}_\psi\left(x^*, z_T\right) - w_0 \mathbf{D}_\psi\left(x^*, z_0\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t}.
\end{aligned}$$

Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right) \\
&\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

□

Corollary B.2. *Suppose we run the Accelerated Stochastic Mirror Descent algorithm with the standard choices $\alpha_t = \frac{2}{t+1}$ and $\eta_t = \eta t$ with $\eta \leq \frac{1}{4L}$. Let $w_T = \frac{1}{3\sigma^2\eta^2 T(T+1)(2T+1)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \leq t \leq T$. The sequence $\{w_t\}_{0 \leq t \leq T}$ satisfies the conditions required by Corollary 3.9. By Corollary 3.9, with probability at least $1-\delta$, $\mathbf{D}_\psi\left(x^*, z_T\right) \leq 2\mathbf{D}_\psi\left(x^*, z_0\right) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3$ and*

$$f\left(y_T\right) - f\left(x^*\right) \leq \frac{4\mathbf{D}_\psi\left(x^*, z_0\right)}{\eta T^2} + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta T.$$

In particular, setting $\eta = \min\left\{\frac{1}{4L}, \frac{\sqrt{\mathbf{D}_\psi\left(x^*, z_0\right)}}{\sqrt{6}\sqrt{G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)}T^{3/2}}\right\}$, we obtain the first case of Theorem 3.6.

Proof of Corollary B.2. Recall from Corollary 3.9 that the sequence $\{w_t\}$ needs to satisfy the following conditions:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1}, \quad \forall 1 \leq t \leq T, \tag{17}$$

$$\frac{w_t \eta_t^2}{1 - L\alpha_t\eta_t} \leq \frac{1}{4\sigma^2}, \quad \forall 0 \leq t \leq T. \tag{18}$$

We will set $\{w_t\}$ so that it satisfies the following additional condition, which will allow us to telescope the sum on the RHS of Corollary 3.9:

$$w_{t-1}\frac{\eta_{t-1}}{\alpha_{t-1}} \geq w_t \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}, \quad \forall 1 \leq t \leq T. \tag{19}$$

Given w_T, we set w_{t-1} for every $1 \leq t \leq T$ so that the first condition (17) holds with equality:

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2 = w_t + 6\sigma^2\eta^2 t^2 w_t^2.$$

Let $C = \sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)$. We set

$$w_T = \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^T i^2} = \frac{1}{C + \sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)} = \frac{1}{2\sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)}.$$

Given this choice for w_T, we now verify that, for all $0 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^t i^2} = \frac{1}{C + \sigma^2\eta^2 t\left(t+1\right)\left(2t+1\right)}.$$

We proceed by induction on t. The base case $t = T$ follows from the definition of w_T. Consider $t \leq T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned}
w_{t-1} &= w_t + 6\sigma^2\eta^2 t^2 w_t^2 \\
&\leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^t i^2} + \frac{6\sigma^2\eta^2 t^2}{\left(C + 6\sigma^2\eta^2\sum_{i=1}^t i^2\right)^2} \\
&\leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^t i^2} + \frac{\left(C + 6\sigma^2\eta^2\sum_{i=1}^t i^2\right) - \left(C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2\right)}{\left(C + 6\sigma^2\eta^2\sum_{i=1}^t i^2\right)\left(C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2\right)} \\
&= \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2}
\end{aligned}$$

as needed. Let us now verify that the second condition (18) also holds. Using that $\frac{2t}{t+1} \leq 2$, $L\eta \leq \frac{1}{4}$, and $T \geq 2$, we obtain

$$\begin{aligned}
\frac{w_t\eta_t^2}{1 - L\alpha_t\eta_t} = \frac{w_t\eta^2 t^2}{1 - L\eta\frac{2t}{t+1}} &\leq 2w_t\eta^2 t^2 \leq \frac{2\eta^2 t^2}{C + 6\sigma^2\eta^2 t^2} \\
&= \frac{t^2}{\sigma^2 T\left(T+1\right)\left(2T+1\right) + 3\sigma^2 t^2} \\
&\leq \frac{1}{\sigma^2\left(2T+1\right) + 3\sigma^2} \leq \frac{1}{4\sigma^2}
\end{aligned}$$

as needed.

Let us now verify that the third condition (19) also holds. Since $\eta_t = \eta t$ and $\alpha_t = \frac{2}{t+1}$, we have $\frac{\eta_{t-1}}{\alpha_{t-1}} = \frac{\eta_t(1-\alpha_t)}{\alpha_t} = \frac{\eta t(t-1)}{2}$. Since $w_t \leq w_{t-1}$, it follows that condition (19) holds.

We now turn our attention to the convergence. By Corollary 3.9, with probability $\geq 1 - \delta$, we have

$$\begin{aligned}
&\sum_{t=1}^T w_t\left(\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right) - f\left(x^*\right)\right) - \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_{t-1}\right) - f\left(x^*\right)\right)\right) + w_T\mathbf{D}_\psi\left(x^*, z_T\right) \\
&\leq w_0\mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right)\sum_{t=1}^T w_t\frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

Grouping terms on the LHS and using that $\alpha_1 = 1$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T-1}\left(w_t\frac{\eta_t}{\alpha_t} - w_{t+1}\frac{\eta_{t+1}\left(1-\alpha_{t+1}\right)}{\alpha_{t+1}}\right)\left(f\left(y_t\right) - f\left(x^*\right)\right) + w_T\frac{\eta_T}{\alpha_T}\left(f\left(y_T\right) - f\left(x^*\right)\right) + w_T\mathbf{D}_\psi\left(x^*, z_T\right) \\
&\leq w_0\mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right)\sum_{t=1}^T w_t\frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

Since $\{w_t\}$ satisfies condition (19), the coefficient of $f(y_t) - f(x^*)$ is non-negative and thus we can drop the above sum. We obtain

$$w_T \frac{\eta_T}{\alpha_T}\left(f(y_T) - f(x^*)\right) + w_T \mathbf{D}_\psi(x^*, z_T) \leq w_0 \mathbf{D}_\psi(x^*, z_0) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

Using that $w_T = \frac{1}{2C}$ and $w_t \leq \frac{1}{C}$ for all $0 \leq t \leq T-1$, we obtain

$$\begin{aligned}
&\frac{1}{2C}\frac{\eta_T}{\alpha_T}\left(f(y_T) - f(x^*)\right) + \frac{1}{2C}\mathbf{D}_\psi(x^*, z_T) \\
&\leq \frac{1}{C}\mathbf{D}_\psi(x^*, z_0) + \frac{1}{C}\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{\eta_T}{\alpha_T}\left(f(y_T) - f(x^*)\right) + \mathbf{D}_\psi(x^*, z_T) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + 2\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + 2C\log\left(\frac{1}{\delta}\right) \\
&= 2\mathbf{D}_\psi(x^*, z_0) + 2\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + 2\sigma^2\log\left(\frac{1}{\delta}\right)\eta^2 T(T+1)(2T+1).
\end{aligned}$$

Using that $L\eta \leq \frac{1}{4}$ and $\frac{2t}{t+1} \leq 2$, we obtain

$$\sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} = \sum_{t=1}^{T} \frac{\eta^2 t^2}{1 - L\eta\frac{2t}{t+1}} \leq \sum_{t=1}^{T} 2\eta^2 t^2 = \frac{1}{3}\eta^2 T(T+1)(2T+1).$$

Plugging in and using that $\eta_T = \eta T$ and $\alpha_T = \frac{2}{T+1}$, we obtain

$$\begin{aligned}
&\eta\frac{T(T+1)}{2}\left(f(y_T) - f(x^*)\right) + \mathbf{D}_\psi(x^*, z_T) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + \left(\frac{2}{3}G^2 + 2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T(T+1)(2T+1) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + 2\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T(T+1)(2T+1).
\end{aligned}$$

We can further simplify the bound by lower bounding $T(T+1) \geq T^2$ and upper bounding $T(T+1)(2T+1) \leq 6T^3$. We obtain

$$\eta T^2\left(f(y_T) - f(x^*)\right) + 2\mathbf{D}_\psi(x^*, z_T) \leq 4\mathbf{D}_\psi(x^*, z_0) + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3.$$

Thus we obtain

$$f(y_T) - f(x^*) \leq \frac{4\mathbf{D}_\psi(x^*, z_0)}{\eta T^2} + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta T,$$

and

$$\mathbf{D}_\psi(x^*, z_T) \leq 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3.$$

□

Corollary B.3. *Suppose we run the Accelerated Stochastic Mirror Descent algorithm with the standard choices* $\alpha_t = \frac{2}{t+1}$ *and* $\eta_t = \min\left\{\frac{t}{4L}, \frac{\eta}{\sqrt{t}}\right\}$. *Let* $w_T = \frac{1}{12\sigma^2\sum_{i=1}^T \eta_t^2}$ *and* $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ *for all* $1 \le t \le T$. *The sequence* $\{w_t\}_{0\le t\le T}$ *satisfies the conditions required by Corollary 3.9. By Corollary 3.9, with probability at least* $1-\delta$, $\mathbf{D}_\psi(x^*, z_T) \le 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1+\log T)$ *and*

$$f(y_T) - f(x^*) \le \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) + \frac{2}{T^{1/2}\eta}\left(2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1+\log T)\right).$$

In particular, setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*,z_0)}}{\sqrt{6}\sqrt{G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)}t^{1/2}}\right\}$, *we obtain the second case of Theorem 3.6.*

Proof of Corollary B.3. Recall from Corollary 3.9 that the sequence $\{w_t\}$ needs to satisfy the following conditions:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \le w_{t-1}, \quad \forall 1 \le t \le T, \tag{20}$$

$$\frac{w_t\eta_t^2}{1 - L\alpha_t\eta_t} \le \frac{1}{4\sigma^2}, \quad \forall 0 \le t \le T. \tag{21}$$

We will set $\{w_t\}$ so that it satisfies the following additional condition, which will allow us to telescope the sum on the RHS of Corollary 3.9:

$$w_{t-1}\frac{\eta_{t-1}}{\alpha_{t-1}} \ge w_t\frac{\eta_t(1-\alpha_t)}{\alpha_t}, \quad \forall 1 \le t \le T-1. \tag{22}$$

Given w_T, we set w_{t-1} for every $1 \le t \le T$ so that the first condition (20) holds with equality:

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2 = w_t + 6\sigma^2\eta^2 t^2 w_t^2.$$

Let $C = 6\sigma^2\sum_{i=1}^T \eta_t^2$. We set

$$w_T = \frac{1}{12\sigma^2\sum_{i=1}^T \eta_t^2} = \frac{1}{2C}.$$

Given this choice for w_T, we now verify that, for all $0 \le t \le T$, we have

$$w_t \le \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2}.$$

We proceed by induction on t. The base case $t = T$ follows from the definition of w_T. Consider $t \le T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned}
w_{t-1} &= w_t + 6\sigma^2\eta_t^2 w_t^2 \\
&\le \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2} + \frac{6\sigma^2\eta_t^2}{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right)^2} \\
&\le \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2} + \frac{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right) - \left(C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2\right)}{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right)\left(C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2\right)} \\
&= \frac{1}{C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2}
\end{aligned}$$

as needed.

Let us now verify that the second condition (21) also holds. Using that $L\eta_t \le \frac{t}{4}$, and $T \ge 2$, we obtain

$$\frac{w_t\eta_t^2}{1 - L\alpha_t\eta_t} \le \frac{w_t\eta_t^2}{1 - \frac{t}{4}\frac{2}{t+1}} \le 2w_t\eta_t^2 \le \frac{2\eta_t^2}{6\sigma^2\sum_{i=1}^T \eta_t^2 + 6\sigma^2\sum_{i=1}^t \eta_i^2} \le \frac{2\eta_t^2}{12\sigma^2\eta_t^2} \le \frac{1}{4\sigma^2}$$

as needed.

Let us now verify that the third condition (22) also holds. Since $\alpha_t = \frac{2}{t+1}$, we have

$$\frac{\eta_{t-1}}{\alpha_{t-1}} = \frac{\eta_{t-1} t}{2},$$
$$\frac{\eta_t (1-\alpha_t)}{\alpha_t} = \frac{\eta_t (t-1)}{2}.$$

If $\eta_{t-1} = \frac{t-1}{4L}$ then we have $\eta_t \leq \frac{t}{4L}$ and $\frac{\eta_t(1-\alpha_t)}{\alpha_t} \leq \frac{\eta_{t-1}}{\alpha_{t-1}} = \frac{t(t-1)}{8L}$. If $\eta_{t-1} = \frac{\eta}{\sqrt{t-1}}$ then $\eta_t = \frac{\eta}{\sqrt{t}}$,we also have $\frac{\eta_t(1-\alpha_t)}{\alpha_t} \leq \frac{\eta_{t-1}}{\alpha_{t-1}}$. Since $w_t \leq w_{t-1}$, it follows that condition (22) holds.

We now turn our attention to the convergence. By Corollary 3.9, with probability $\geq 1-\delta$, we have

$$\sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f(y_t) - f(x^*) \right) - \frac{\eta_t (1-\alpha_t)}{\alpha_t} \left(f(y_{t-1}) - f(x^*) \right) \right) + w_T \mathbf{D}_\psi (x^*, z_T)$$
$$\leq w_0 \mathbf{D}_\psi (x^*, z_0) + \left(G^2 + 3\sigma^2 \right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t \eta_t} + \log\left(\frac{1}{\delta}\right).$$

Grouping terms on the LHS and using that $\alpha_1 = 1$, we obtain

$$\sum_{t=1}^{T-1} \left(w_t \frac{\eta_t}{\alpha_t} - w_{t+1} \frac{\eta_{t+1}(1-\alpha_{t+1})}{\alpha_{t+1}} \right) \left(f(y_t) - f(x^*) \right) + w_T \frac{\eta_T}{\alpha_T} \left(f(y_T) - f(x^*) \right) + w_T \mathbf{D}_\psi (x^*, z_T)$$
$$\leq w_0 \mathbf{D}_\psi (x^*, z_0) + \left(G^2 + 3\sigma^2 \right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t \eta_t} + \log\left(\frac{1}{\delta}\right).$$

Since $\{w_t\}$ satisfies condition (22), the coefficient of $f(y_t) - f(x^*)$ is non-negative and thus we can drop the above sum. We obtain

$$w_T \frac{\eta_T}{\alpha_T} \left(f(y_T) - f(x^*) \right) + w_T \mathbf{D}_\psi (x^*, z_T) \leq w_0 \mathbf{D}_\psi (x^*, z_0) + \left(G^2 + 3\sigma^2 \right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t \eta_t} + \log\left(\frac{1}{\delta}\right).$$

Using that $w_T = \frac{1}{2C}$ and $w_t \leq \frac{1}{C}$ for all $0 \leq t \leq T-1$, we obtain

$$\frac{1}{2C} \frac{\eta_T}{\alpha_T} \left(f(y_T) - f(x^*) \right) + \frac{1}{2C} \mathbf{D}_\psi (x^*, z_T)$$
$$\leq \frac{1}{C} \mathbf{D}_\psi (x^*, z_0) + \frac{1}{C} \left(G^2 + 3\sigma^2 \right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t \eta_t} + \log\left(\frac{1}{\delta}\right).$$

Thus

$$\frac{\eta_T}{\alpha_T} \left(f(y_T) - f(x^*) \right) + \mathbf{D}_\psi (x^*, z_T)$$
$$\leq 2\mathbf{D}_\psi (x^*, z_0) + 2 \left(G^2 + 3\sigma^2 \right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t \eta_t} + 2C \log\left(\frac{1}{\delta}\right).$$

Using that $L\eta_t \leq \frac{t}{4}$, we obtain

$$\sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t \eta_t} = \sum_{t=1}^{T} \frac{\eta_t^2}{1 - \frac{t}{4}\frac{2}{t+1}} \leq \sum_{t=1}^{T} 2\eta_t^2 = \frac{C}{3\sigma^2}.$$

Plugging in and using that $\eta_T = \eta T$ and $\alpha_T = \frac{2}{T+1}$, we obtain

$$\frac{\eta_T (T+1)}{2} \left(f(y_T) - f(x^*) \right) + \mathbf{D}_\psi (x^*, z_T)$$
$$\leq 2\mathbf{D}_\psi (x^*, z_0) + \left(2G^2 + 6 \left(1 + \log\left(\frac{1}{\delta}\right) \right) \sigma^2 \right) \frac{C}{3\sigma^2}.$$

If $\frac{T}{4L} \leq \frac{\eta}{\sqrt{T}}$ which means $T^{3/2} \leq 4L\eta$ then $\eta_T = \frac{T}{4L}$ we have

$$C = 6\sigma^2 \sum_{i=1}^{T} \eta_t^2 = \frac{6\sigma^2}{16L^2} \sum_{i=1}^{T} t^2 \leq \frac{3\sigma^2 T^3}{8L^2} \leq 6\sigma^2\eta^2.$$

Hence

$$\begin{aligned}
&\frac{\eta_T\left(T+1\right)}{2}\left(f\left(y_T\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*, z_T\right) \\
&\leq 2\mathbf{D}_\psi\left(x^*, z_0\right)+\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\frac{3T^3}{4L^2},
\end{aligned}$$

which entails

$$\begin{aligned}
f\left(y_T\right)-f\left(x^*\right) &\leq \frac{16L}{T^2}\mathbf{D}_\psi\left(x^*, z_0\right)+\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\frac{6T}{L} \\
&= \frac{16L}{T^2}\mathbf{D}_\psi\left(x^*, z_0\right)+\frac{6}{\sqrt{T}}\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\frac{T^{3/2}}{L} \\
&\leq \frac{16L}{T^2}\mathbf{D}_\psi\left(x^*, z_0\right)+\frac{24}{\sqrt{T}}\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta,
\end{aligned}$$

and

$$\mathbf{D}_\psi\left(x^*, z_T\right) \leq 2\mathbf{D}_\psi\left(x^*, z_0\right)+12\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2.$$

If $\frac{\eta}{\sqrt{T}} \leq \frac{T}{4L}$ then $\eta_T = \frac{\eta}{\sqrt{T}}$. Let T_0 be the largest t such that $\frac{\eta}{\sqrt{t}} \geq \frac{t}{4L}$, we have $T_0^3 \leq 16L^2\eta^2$

$$\begin{aligned}
C &= 6\sigma^2 \sum_{i=1}^{T} \eta_t^2 \\
&= 6\sigma^2 \sum_{i=1}^{T_0} \eta_t^2 + 6\sigma^2 \sum_{i=T_0+1}^{T} \eta_t^2 \\
&= \frac{6\sigma^2}{16L^2} \sum_{i=1}^{T_0} t^2 + 6\sigma^2\eta^2 \sum_{i=T_0+1}^{T} \frac{1}{t} \\
&\leq \frac{6\sigma^2}{16L^2} T_0^3 + 6\sigma^2\eta^2 \sum_{i=T_0+1}^{T} \frac{1}{t} \\
&\leq 6\sigma^2\eta^2 \sum_{i=1}^{T} \frac{1}{t} \leq 6\sigma^2\eta^2(1+\log T).
\end{aligned}$$

Hence

$$f\left(y_T\right)-f\left(x^*\right) \leq \frac{2}{T^{1/2}\eta}\left(2\mathbf{D}_\psi\left(x^*, z_0\right)+12\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1+\log T)\right),$$

and

$$\mathbf{D}_\psi\left(x^*, z_T\right) \leq 2\mathbf{D}_\psi\left(x^*, z_0\right)+12\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1+\log T).$$

Overall we have

$$f\left(y_T\right)-f\left(x^*\right) \leq \frac{16L}{T^2}\mathbf{D}_\psi\left(x^*, z_0\right)+\frac{2}{T^{1/2}\eta}\left(2\mathbf{D}_\psi\left(x^*, z_0\right)+12\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1+\log T)\right).$$

□

C. Missing Proofs from Section 4

C.1. Stochastic Gradient Descent

Proof of Lemma 4.2. We start from the smoothness of f

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= -\eta_t \left\langle \nabla f(x_t), \widehat{\nabla} f(x_t) \right\rangle + \frac{L\eta_t^2}{2} \left\| \widehat{\nabla} f(x_t) \right\|^2. \end{aligned}$$

By writing $\widehat{\nabla} f(x_t) = \xi_t + \nabla f(x_t)$ we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\eta_t \langle \nabla f(x_t), \xi_t + \nabla f(x_t) \rangle + \frac{L\eta_t^2}{2} \|\xi_t + \nabla f(x_t)\|^2 \\ &= -\eta_t \|\nabla f(x_t)\|^2 - \eta_t \langle \nabla f(x_t), \xi_t \rangle \\ &\quad + \frac{L\eta_t^2}{2} \|\xi_t\|^2 + \frac{L\eta_t^2}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \langle \nabla f(x_t), \xi_t \rangle. \end{aligned}$$

We obtain the inequality (7) by rearranging the terms. □

Proof of Theorem 4.3. We prove by induction. The base case $t = T + 1$ trivially holds. Consider $1 \leq t \leq T$, we have

$$\begin{aligned} \mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\exp\left(Z_t\right) \mathbb{E}\left[\exp\left(S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_k\right]. \end{aligned}$$

From the induction hypothesis we have $\mathbb{E}\left[\exp\left(S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \leq \exp\left(3\sigma^2 \sum_{i=t+1}^{T} \frac{w_i \eta_i^2 L}{2}\right)$, hence

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \sum_{i=t+1}^{T} \frac{w_i \eta_i^2 L}{2}\right) \mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right].$$

We have then

$$\begin{aligned} \mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(w_t \left(\eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2 + \Delta_{t+1} - \Delta_t\right) - v_t \|\nabla f(x_T)\|^2\right) \mid \mathcal{F}_t\right] \\ &\leq \mathbb{E}\left[\exp\left(w_t \left(\eta_t(\eta_t L - 1) \langle \nabla f(x_t), \xi_t \rangle + \frac{\eta_t^2 L}{2} \|\xi_t\|^2\right) - v_t \|\nabla f(x_t)\|^2\right) \mid \mathcal{F}_t\right] \\ &= \exp\left(-v_t \|\nabla f(x_t)\|^2\right) \mathbb{E}\left[\exp\left(w_t \left(\eta_t(\eta_t L - 1) \langle \nabla f(x_t), \xi_t \rangle + \frac{\eta_t^2 L}{2} \|\xi_t\|^2\right)\right) \mid \mathcal{F}_t\right] \\ &\leq \exp\left(-v_t \|\nabla f(x_t)\|^2\right) \exp\left(3\sigma^2 \left(w_t^2 \eta_t^2 (\eta_t L - 1)^2 \|\nabla f(x_t)\|^2 + \frac{w_t \eta_t^2 L}{2}\right)\right) \\ &= \exp\left(3\sigma^2 \frac{w_t \eta_t^2 L}{2}\right). \end{aligned}$$

where the second line is due to (7) in Lemma 4.2 and the second to last line is due to Lemma 2.2.Therefore

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \sum_{i=t}^{T} \frac{w_i \eta_i^2 L}{2}\right)$$

which we what we need to show. □

Proof of Corollary 4.4. In Lemma 4.3, Let $t = 1$ we obtain

$$\mathbb{E}\left[\exp\left(S_1\right)\right] \leq \exp\left(3\sigma^2 \sum_{t=1}^{T} \frac{w_t \eta_t^2 L}{2}\right)$$

hence by Markov's inequality we have

$$\Pr\left[S_1 \geq \left(3\sigma^2 \sum_{t=1}^{T} \frac{w_t \eta_t^2 L}{2}\right) + \log\frac{1}{\delta}\right] \leq \delta.$$

In other words, with probability $\geq 1-\delta$ (once the condition in Lemma 4.3 is satisfied)

$$\sum_{t=1}^{T}\left[w_t\eta_t\left(1-\frac{\eta_t L}{2}\right)-v_t\right]\|\nabla f(x_t)\|^2 + w_t\left(\Delta_{t+1}-\Delta_t\right)$$
$$\leq 3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}+\log\frac{1}{\delta}.$$

This gives

$$\sum_{t=1}^{T}\left[w_t\eta_t\left(1-\frac{\eta_t L}{2}\right)-v_t\right]\|\nabla f(x_t)\|^2 + w_T\Delta_{T+1} \leq w_1\Delta_1 + \left(\sum_{t=2}^{T}(w_t-w_{t-1})\Delta_t + 3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}\right)+\log\frac{1}{\delta}$$

as needed. □

Proof of Theorem 4.1 . **First case.**

Starting from this inequality, we will specify the choice of η_t and w_t to obtain the bound. Consider $\eta_t = \eta$ with $\eta L \leq 1$, $w_t = w = \frac{1}{6\sigma^2\eta}$. Note that $w_t\eta_t^2 L = \frac{\eta L}{6\sigma^2} \leq \frac{1}{2\sigma^2}$ satisfies the condition of Lemma 4.3, we have

$$\begin{aligned}\text{LHS of (9)} &= w\Delta_{T+1} + \sum_{t=1}^{T}\left[w\eta\left(1-\frac{\eta L}{2}\right) - 3\sigma^2 w^2\eta^2(\eta L-1)^2\right]\|\nabla f(x_t)\|^2\\ &= w\Delta_{T+1} + w\eta\sum_{t=1}^{T}\left[1-\frac{\eta L}{2}-\frac{1}{2}(\eta L-1)^2\right]\|\nabla f(x_t)\|^2\\ &\geq w\Delta_{T+1} + \frac{w\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\end{aligned}$$

where the last inequality is due to $1-\frac{\eta L}{2}-\frac{(1-\eta L)^2}{2} \geq \frac{1}{2}$ when $0 \leq \eta L \leq 1$. Besides,

$$\text{RHS of (9)} = w\Delta_1 + \frac{3\sigma^2}{2}w\eta^2 LT + \log\frac{1}{\delta}.$$

Hence with probability $\geq 1-\delta$

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + \frac{2\Delta_{T+1}}{\eta} &\leq \frac{2\Delta_1}{\eta} + 3\sigma^2\eta LT + \frac{2}{w\eta}\log\frac{1}{\delta}\\ &= \frac{2\Delta_1}{\eta} + 3\sigma^2\eta LT + 12\sigma^2\log\frac{1}{\delta}.\end{aligned}$$

Finally by choosing $\eta = \min\left\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 LT}}\right\}$ and noticing $\Delta_{T+1} \geq 0$, we obtain the desired inequality.

Second case.

Consider $\eta_t = \frac{\eta}{\sqrt{t}}$ with $\eta L \leq 1$, $w_t = w = \frac{1}{6\sigma^2\eta}$. Again, we have $w_t\eta_t^2 L = \frac{\eta L}{6\sigma^2 t} \leq \frac{1}{2\sigma^2}$, then

$$\begin{aligned}
\text{LHS of (9)} =& \sum_{t=1}^{T}\left[\frac{w\eta}{\sqrt{t}}\left(1-\frac{\eta L}{2\sqrt{t}}\right) - \frac{3\sigma^2 w^2\eta^2}{t}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1}\\
=& \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - \frac{3\sigma^2 w\eta}{\sqrt{t}}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1}\\
\geq& \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - 3\sigma^2 w\eta\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1}\\
=& \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - \frac{1}{2}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1}\\
\geq& \sum_{t=1}^{T}\frac{w\eta}{2\sqrt{t}}\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \geq \frac{w\eta}{2\sqrt{T}}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + w\Delta_{T+1},
\end{aligned}$$

where the second inequality is due to $1-\frac{\eta L}{2\sqrt{t}} - \frac{1}{2}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2 \geq \frac{1}{2}$ when $0 \leq \frac{\eta L}{\sqrt{t}} \leq 1$. Besides,

$$\begin{aligned}
\text{RHS of (9)} =& w\Delta_1 + \frac{3\sigma^2}{2}w\eta^2 L\sum_{t=1}^{T}\frac{1}{t} + \log\frac{1}{\delta}\\
\leq& w\Delta_1 + \frac{3\sigma^2}{2}w\eta^2 L(1+\log T) + \log\frac{1}{\delta}.
\end{aligned}$$

Therefore, with probability $\geq 1-\delta$

$$\begin{aligned}
&\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + \frac{2\sqrt{T}\Delta_{T+1}}{\eta}\\
\leq& \sqrt{T}\left(\frac{2\Delta_1}{\eta} + 3\sigma^2\eta L\left(1+\log T\right) + \frac{2}{w\eta}\log\frac{1}{\delta}\right)\\
=& \sqrt{T}\left(\frac{2\Delta_1}{\eta} + 3\sigma^2\eta L\left(1+\log T\right) + 12\sigma^2\log\frac{1}{\delta}\right).
\end{aligned}$$

Choose $\eta = \frac{1}{L}$, and notice $\Delta_{T+1} \geq 0$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L + 3\sigma^2\left(1+\log T\right) + 12\sigma^2\log\frac{1}{\delta}}{\sqrt{T}}.$$

□

C.2. AdaGrad-Norm

Proof of Lemma 4.6. Starting from the smoothness of f

$$\begin{aligned}
f(x_{t+1}) - f(x_t) \leq& \langle\nabla f(x_t), x_{t+1}-x_t\rangle + \frac{L}{2}\|x_{t+1}-x_t\|^2\\
=& -\frac{\eta}{b_t}\left\langle\nabla f(x_t), \widehat{\nabla} f(x_t)\right\rangle + \frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2\\
=& -\frac{\eta}{b_t}\|\nabla f(x_t)\|^2 - \frac{\eta}{b_t}\langle\nabla f(x_t), \xi_t\rangle + \frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2. \qquad (23)
\end{aligned}$$

Multiplying both sides by $\frac{b_t}{\eta(2b_t-b_0)}$ and rearranging, we obtain

$$\begin{aligned}\frac{\|\nabla f(x_t)\|^2}{2b_t-b_0} &\leq \frac{-\langle\nabla f(x_t),\xi_t\rangle}{2b_t-b_0}+\frac{b_t\left(\Delta_t-\Delta_{t+1}\right)}{\eta\left(2b_t-b_0\right)}+\frac{\eta L}{2b_t\left(2b_t-b_0\right)}\left\|\widehat{\nabla} f(x_t)\right\|^2\\ &=\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)\langle\nabla f(x_t),\xi_t\rangle-\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}\\ &\quad+\frac{b_t\left(\Delta_t-\Delta_{t+1}\right)}{\eta\left(2b_t-b_0\right)}+\frac{\eta L}{2b_t\left(2b_t-b_0\right)}\left\|\widehat{\nabla} f(x_t)\right\|^2. \end{aligned}\tag{24}$$

Note that by the smoothness of f we also have $\|\nabla f(x_t)\|^2 \leq 2L\Delta_t$. Combining with Cauchy-Schwatz inequality we have

$$\begin{aligned}&\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)\langle\nabla f(x_t),\xi_t\rangle\\ &\leq\left(\frac{b_{t-1}}{2b_{t-1}-b_0}-\frac{b_t}{2b_t-b_0}\right)\frac{\|\nabla f(x_t)\|^2}{2\eta L}+\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)^2\frac{\eta L}{\frac{b_{t-1}}{2b_{t-1}-b_0}-\frac{b_t}{2b_t-b_0}}\frac{\|\xi_t\|^2}{2}\\ &=\left(\frac{b_{t-1}}{2b_{t-1}-b_0}-\frac{b_t}{2b_t-b_0}\right)\frac{\Delta_t}{\eta}+\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)\frac{\eta L}{b_0}\|\xi_t\|^2.\end{aligned}$$

Plugging into (24) we obtain

$$\begin{aligned}\frac{\|\nabla f(x_t)\|^2}{2b_t-b_0} &\leq\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)\frac{\eta L}{b_0}\|\xi_t\|^2-\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}\\ &\quad+\frac{b_{t-1}\Delta_t}{\eta\left(2b_{t-1}-b_0\right)}-\frac{b_t\Delta_{t+1}}{\eta\left(2b_t-b_0\right)}+\frac{\eta L}{2b_t\left(2b_t-b_0\right)}\left\|\widehat{\nabla} f(x_t)\right\|^2.\end{aligned}$$

Sum up from 1 to T

$$\begin{aligned}\sum_{t=1}^T\frac{\|\nabla f(x_t)\|^2}{2b_t-b_0} &\leq\sum_{t=1}^T\left(\frac{1}{2b_{t-1}-b_0}-\frac{1}{2b_t-b_0}\right)\frac{\eta LM_T}{b_0}-\sum_{t=1}^T\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}\\ &\quad+\sum_{t=1}^T\left(\frac{b_{t-1}\Delta_t}{\eta\left(2b_{t-1}-b_0\right)}-\frac{b_t\Delta_{t+1}}{\eta\left(2b_t-b_0\right)}\right)+\sum_{t=1}^T\frac{\eta L}{2b_t\left(2b_t-b_0\right)}\left\|\widehat{\nabla} f(x_t)\right\|^2\\ &\leq\frac{\eta LM_T}{b_0^2}+\frac{\Delta_1}{\eta}-\frac{b_T\Delta_{T+1}}{\eta\left(2b_T-b_0\right)}+\frac{\eta L}{2}\sum_{t=1}^T\frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t^2}-\sum_{t=1}^T\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}\\ &\leq\frac{\eta LM_T}{b_0^2}+\frac{\Delta_1}{\eta}-\frac{b_T\Delta_{T+1}}{\eta\left(2b_T-b_0\right)}+\frac{\eta L}{2}\log\frac{b_T^2}{b_0^2}-\sum_{t=1}^T\frac{\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}.\end{aligned}$$

□

Proof of Lemma 4.7. For $1\leq t\leq T$, given $|\lambda|\leq\frac{1}{\sigma}$, we have

$$\mathbb{E}\left[\exp\left(\lambda^2\left\langle\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|},\xi_t\right\rangle^2\right)\mid\mathcal{F}_t\right]\leq\mathbb{E}\left[\exp\left(\lambda^2\|\xi_t\|^2\right)\mid\mathcal{F}_t\right]\leq\exp\left(\lambda^2\sigma^2\right).$$

Thus $-\left\langle\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|},\xi_t\right\rangle$ is a centered σ-sub-Gaussian RV given $\mathcal{F}_t$, and we can apply Lemma 2.2 for $a=\frac{w\|\nabla f(x_t)\|}{2b_{t-1}-b_0}$ and $b=0$, for some constant $w>0$ to get

$$\begin{aligned}\mathbb{E}\left[\exp\left(-\frac{w\langle\nabla f(x_t),\xi_t\rangle}{2b_{t-1}-b_0}\right)\mid\mathcal{F}_t\right] &=\mathbb{E}\left[\exp\left(-\frac{w\|\nabla f(x_t)\|\left\langle\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|},\xi_t\right\rangle}{2b_{t-1}-b_0}\right)\mid\mathcal{F}_t\right]\\ &\leq\exp\left(\frac{2w^2\|\nabla f(x_t)\|^2\sigma^2}{\left(2b_{t-1}-b_0\right)^2}\right).\end{aligned}$$

By a simple induction argument we obtain

$$\mathbb{E}\left[\exp\left(\sum_{t=1}^{T}-\frac{w\left\langle\nabla f(x_t),\xi_t\right\rangle}{2b_{t-1}-b_0}-\frac{2w^2\left\|\nabla f(x_t)\right\|^2\sigma^2}{\left(2b_{t-1}-b_0\right)^2}\right)\right]\leq 1.$$

Hence, by Markov's inequality

$$\Pr\left[\sum_{t=1}^{T}-\frac{w\left\langle\nabla f(x_t),\xi_t\right\rangle}{2b_{t-1}-b_0}-\frac{2w^2\left\|\nabla f(x_t)\right\|^2\sigma^2}{\left(2b_{t-1}-b_0\right)^2}\geq\log\frac{1}{\delta}\right]\leq\delta$$

which implies with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T}-\frac{w\left\langle\nabla f(x_t),\xi_t\right\rangle}{2b_{t-1}-b_0}\leq\sum_{t=1}^{T}\frac{2w^2\sigma^2\left\|\nabla f(x_t)\right\|^2}{\left(2b_{t-1}-b_0\right)^2}+\log\frac{1}{\delta}.$$

However, we now have a mismatch between the index of the numerator and denominator of the first term in the RHS. To resolve this, observing that $\left\|\nabla f(x_t)\right\|^2\leq 2\left\|\nabla f(x_t)-\nabla f(x_{t-1})\right\|^2+2\left\|\nabla f(x_{t-1})\right\|^2$ and using the smoothness of f for the first term, ie $\left\|\nabla f(x_t)-\nabla f(x_{t-1})\right\|\leq L\left\|x_t-x_{t-1}\right\|=\frac{L\eta}{b_{t-1}}\left\|\widehat{\nabla}f(x_{t-1})\right\|$, we have

$$\begin{aligned}\sum_{t=1}^{T}\frac{2w^2\sigma^2\left\|\nabla f(x_t)\right\|^2}{\left(2b_{t-1}-b_0\right)^2}\leq\sum_{t=2}^{T}\frac{4w^2\eta^2L^2\sigma^2\left\|\widehat{\nabla}f(x_{t-1})\right\|^2}{b_{t-1}^2\left(2b_{t-1}-b_0\right)^2}+\sum_{t=2}^{T}\frac{4w^2\sigma^2\left\|\nabla f(x_{t-1})\right\|^2}{\left(2b_{t-1}-b_0\right)^2}\\+\frac{2w^2\sigma^2\left\|\nabla f(x_1)\right\|^2}{b_0^2}.\end{aligned}$$

Finally, since $2b_{t-1}-b_0\geq b_0$ and $\sum_{t=2}^{T}\frac{\left\|\widehat{\nabla}f(x_{t-1})\right\|^2}{b_{t-1}^2}\leq\log\frac{b_T^2}{b_0^2}$, the proof is completed. □

Lemma C.1. *With probability at least* $1-\delta$

$$\sum_{t=1}^{T}\left\|\xi_t\right\|^2\leq\sigma^2T+\sigma^2\log\frac{1}{\delta}.$$

Proof of Lemma C.1. It is not hard to verify that

$$\mathbb{E}\left[\exp\left(\frac{\sum_{t=1}^{T}\left\|\xi_t\right\|^2}{\sigma^2}\right)\right]\leq\exp\left(\sum_{t=1}^{T}1\right)=\exp(T).$$

Thus by Markov's inequality

$$\Pr\left[\sum_{t=1}^{T}\left\|\xi_t\right\|^2\geq\sigma^2T+\sigma^2\log\frac{1}{\delta}\right]=\Pr\left[\exp\left(\frac{\sum_{t=1}^{T}\left\|\xi_t\right\|^2}{\sigma^2}\right)\geq\frac{\exp(T)}{\delta}\right]\leq\delta.$$

Therefore with probability at least $1-\delta$

$$\sum_{t=1}^{T}\left\|\xi_t\right\|^2\leq\sigma^2T+\sigma^2\log\frac{1}{\delta}.$$

□

Proof of Theorem 4.5. From Lemma 4.6 and 4.7, we have with probability at least $1-\delta$

$$\begin{aligned}\sum_{t=1}^{T}\frac{\left\|\nabla f(x_t)\right\|^2}{2b_t-b_0}\leq\frac{\eta LM_T}{b_0^2}+\frac{\Delta_1}{\eta}+\frac{2w\sigma^2\left\|\nabla f(x_1)\right\|^2}{b_0^2}+\left(\frac{\eta L}{2}+\frac{4w\eta^2L^2\sigma^2}{b_0^2}\right)\log\frac{b_T^2}{b_0^2}\\+\sum_{t=2}^{T}\frac{4w\sigma^2\left\|\nabla f(x_{t-1})\right\|^2}{\left(2b_{t-1}-b_0\right)^2}+\frac{1}{w}\log\frac{1}{\delta}.\end{aligned}$$

Here we choose $w = \frac{b_0}{8\sigma^2}\min\left\{1; \frac{b_0}{\eta L}\right\}$ and simplify the result to get

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2b_t - b_0} - \sum_{t=2}^{T}\frac{4w\sigma^2\|\nabla f(x_{t-1})\|^2}{(2b_{t-1}-b_0)^2} \leq \frac{\eta L M_T}{b_0^2} + \frac{\Delta_1}{\eta} + \frac{\|\nabla f(x_1)\|^2}{4b_0^2} + \eta L\log\frac{b_T^2}{b_0^2} + \frac{8\sigma^2}{b_0}\left(1+\frac{\eta L}{b_0}\right)\log\frac{1}{\delta}.$$

Note that by the choice of w, in the LHS of the above,

$$\frac{4w\sigma^2}{(2b_{t-1}-b_0)^2} \leq \frac{b_0}{2(2b_{t-1}-b_0)^2} \leq \frac{1}{2(2b_{t-1}-b_0)}.$$

Hence, we have

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2(2b_t - b_0)} \leq \sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2b_t - b_0} - \sum_{t=2}^{T}\frac{4w\sigma^2\|\nabla f(x_{t-1})\|^2}{(2b_{t-1}-b_0)^2}.$$

which implies

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{4b_T} \leq \frac{\eta L M_T}{b_0^2} + \frac{\Delta_1}{\eta} + \frac{\|\nabla f(x_1)\|^2}{4b_0^2} + \eta L\log\frac{b_T^2}{b_0^2} + \frac{8\sigma^2}{b_0}\left(1+\frac{\eta L}{b_0}\right)\log\frac{1}{\delta}.$$

It is known that with probability at least $1-\delta$, $M_T \leq \sigma^2\left(1+\log\frac{T}{\delta}\right)$ (Li & Orabona, 2020; Liu et al., 2022). By the union bound, we have with probability at least $1-2\delta$

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{4b_T} \leq \eta L\log\frac{b_T^2}{b_0^2} + \frac{\Delta_1}{\eta} + \frac{\|\nabla f(x_1)\|^2}{4b_0^2} + \frac{8\sigma^2}{b_0}\left(1+\frac{\eta L}{b_0}\right)\log\frac{1}{\delta} + \frac{\eta L\sigma^2}{b_0^2}\left(1+\log\frac{T}{\delta}\right)\\
\Rightarrow &\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq 4b_T\left(\eta L\log\frac{b_T^2}{b_0^2} + \underbrace{\frac{\Delta_1}{\eta} + \frac{\|\nabla f(x_1)\|^2}{4b_0^2} + \frac{8\sigma^2}{b_0}\left(1+\frac{\eta L}{b_0}\right)\log\frac{1}{\delta} + \frac{\eta L\sigma^2}{b_0^2}\left(1+\log\frac{T}{\delta}\right)}_{g(\delta)=O(1+\sigma^2\log\frac{T}{\delta})}\right).
\end{aligned}$$

Note that

$$b_T = \sqrt{b_0^2 + \sum_{t=1}^{T}\left\|\widehat{\nabla} f(x_t)\right\|^2} \leq \sqrt{b_0^2 + 2\sum_{t=1}^{T}\|\xi_t\|^2 + \sum_{t=1}^{T}2\|\nabla f(x_t)\|^2}.$$

Now we consider the following two cases

Case 1. $\sum_{t=1}^{T}2\|\nabla f(x_t)\|^2 \leq b_0^2 + 2\sum_{t=1}^{T}\|\xi_t\|^2$, then we have

$$\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq 4\sqrt{2b_0^2 + 4\sum_{t=1}^{T}\|\xi_t\|^2}\left(\eta L\log\frac{2b_0^2 + 4\sum_{t=1}^{T}\|\xi_t\|^2}{b_0^2} + g(\delta)\right).$$

Case 2. $\sum_{t=1}^{T} 2\left\|\nabla f(x_t)\right\|^2 > b_0^2 + 2\sum_{t=1}^{T}\left\|\xi_t\right\|^2$, then we have

$$\begin{aligned}
\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2 &\leq 8\sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2}\left(\eta L\log\frac{4\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2}{b_0^2}+g(\delta)\right)\\
\Rightarrow \sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2} &\leq 16\eta L\log\frac{2\sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2}}{b_0}+8g(\delta)\\
&\leq 16\eta L\log\frac{\sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2}}{32\eta L}+16\eta L\log\frac{64\eta L}{b_0}+8g(\delta)\\
&\leq \frac{\sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2}}{2}+16\eta L\log\frac{64\eta L}{b_0}+8g(\delta)\\
\Rightarrow \sqrt{\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2} &\leq 32\eta L\log\frac{64\eta L}{b_0}+16g(\delta).
\end{aligned}$$

Combining the two cases, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2 \leq 4\sqrt{2b_0^2+4\sum_{t=1}^{T}\left\|\xi_t\right\|^2}&\left(\eta L\log\frac{2b_0^2+4\sum_{t=1}^{T}\left\|\xi_t\right\|^2}{b_0^2}+g(\delta)\right)\\
&+\left(32\eta L\log\frac{64\eta L}{b_0}+16g(\delta)\right)^2.
\end{aligned}$$

The final step is to use Lemma C.1 and union bound to get, with probability at least $1-3\delta$

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2 \leq \frac{4}{T}\sqrt{2b_0^2+4\sigma^2T+4\sigma^2\log\frac{1}{\delta}}&\left(\eta L\log\frac{2b_0^2+4\sigma^2T+4\sigma^2\log\frac{1}{\delta}}{b_0^2}+g(\delta)\right)\\
&+\frac{1}{T}\left(32\eta L\log\frac{64\eta L}{b_0}+16g(\delta)\right)^2.
\end{aligned}$$

□