Dual Averaging Converges for Nonconvex Smooth Stochastic Optimization

Tuo Liu KAUST

El Mehdi Saad UM6P College of Computing

Wojciech Kotłowski Poznan University of Technology

Francesco Orabona KAUST

Abstract

Dual averaging and gradient descent with their stochastic variants stand as the two canonical recipes for first-order optimization: Every modern variant can be viewed as a descendant of one or the other. In the convex regime, these algorithms have been deeply studied, and we know that the two classes are essentially equivalent in terms of theoretical guarantees. Instead, in the nonconvex setting, the situation is drastically different: While it is provable that SGD can minimize the gradient norm of non-convex smooth functions, no finite-time complexity guarantee for Stochastic Dual Averaging (SDA) was known in the same setting. In this paper, we close this gap by showing that a tuned SDA exhibits a rate of convergence of $\mathcal{O}(1/T + \sigma \log T/\sqrt{T})$, similar to that of SGD under the same assumptions. To our best knowledge, this is the first complete convergence theory for dual averaging on non-convex smooth stochastic problems without restrictive assumptions, closing a long-standing open problem in the field. Beyond the base algorithm, we also discuss ADA-DA, a variant that marries SDA with AdaGrad's auto-scaling, which achieves the same rate without requiring knowledge of the noise variance.

1 INTRODUCTION

Stochastic first-order methods have become the computational workhorse for large-scale optimization problems. Among these methods, *Stochastic Dual Averaging* (SDA) offers a principled alternative to classical Stochastic Gradient Descent (SGD) by updating the primal iterate through a running average of past gradients.

First of all, one might wonder whether studying dual averaging (DA) and its variants in the unconstrained setting is worthwhile. Indeed, a widespread misunderstanding is that DA and gradient descent (GD) only differ in the constrained setting. In fact, the two methods already diverge in the unconstrained setting, because they incorporate new subgradients in fundamentally different ways. The origins of DA trace back to Nesterov's seminal work (Nesterov, 2009), where the author observed that the step-size conditions required for GD,

$$\eta_t > 0, \quad \eta_t \to 0, \quad \sum_{t=1}^{\infty} \eta_t = \infty,$$

necessarily make "new subgradients enter the model with decreasing weights" (Nesterov, 2009, page 4), contradicting the principle of recursive methods where fresher information should outweigh stale data. More formally, when initialized at $\boldsymbol{x}_1 = \mathbf{0}$, GD and DA generate the iterates[1]

$$\boldsymbol{x}_{t+1}^{\mathrm{GD}} = -\sum_{i=1}^{t} \eta_i \boldsymbol{g}_i, \quad \boldsymbol{x}_{t+1}^{\mathrm{DA}} = -\eta_t \sum_{i=1}^{t} \boldsymbol{g}_i,$$

where $\boldsymbol{g}_i$ is a subgradient of the objective function at $\boldsymbol{x}_i$. The effective weight assigned to the latest gradient is therefore $\eta_t / \left(\sum_{i=1}^{t} \eta_i\right)$ for GD, but $1/t$ for DA. Because the sequence (η_t) is decreasing, the factor $1/t$ is larger; as a result, DA places comparatively

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

[1] Nesterov (2009) also proposed the possibility in DA to weigh each single gradient by an additional parameter λ_i, in addition to η_t. While potentially useful in the deterministic setting, this idea has been abandoned in the stochastic setting (see, e.g., Xiao, 2010), hence we do not consider it here.

greater emphasis on the latest information. This property might be particularly valuable in non-convex landscapes, where earlier gradient directions may oppose the current descent direction, even in low-noise regimes (e.g., with the Rosenbrock function). Empirical evidence in Jelassi and Defazio (2020) corroborates this intuition, showing that SDA-based optimizers have been shown to achieve competitive performance in deep learning benchmarks (Jelassi and Defazio, 2020), outperforming their SGD counterparts. Beyond deep learning, SDA-type methods have proven highly effective in various other settings, such as inducing sparsity with ℓ_1-regularization (He et al., 2020; Xiao, 2009; McMahan, 2010) and improving convergence in federated composite optimization (Yuan et al., 2021). Moreover, from a theoretical standpoint, the structure of a dual accumulator allows SDA to naturally form the basis of parameter-free optimizers that eliminate the expensive and often unreliable process of learning rate tuning——a major practical advantage over SGD (Orabona and Tommasi, 2017).

However, although a decade of work has thoroughly characterized SDA in convex settings, its behavior on objectives that are smooth and possibly non-convex remains poorly understood, leaving open even the question of whether SDA can provably minimize the expected squared norm of the gradients, with the literature explicitly noting this gap (He et al., 2020). Here, we close this gap by deriving strong theoretical guarantees for SDA on smooth, non-convex objectives.

Our Contribution. We provide the first iterate-level guarantees for SDA on *smooth, possibly non-convex* objectives. For deterministic step sizes of order $1/\sqrt{t}$, our bounds recover the optimal $\mathcal{O}(T^{-1/2})$ rate previously established for SGD (Ghadimi and Lan, 2012), under the noisy-strong-growth (affine noise) assumption. Notably, this rate matches the information-theoretic lower bound for first-order methods on smooth non-convex objectives (Arjevani et al., 2023), showing that SDA is rate-optimal in this setting. In addition, we derive a high-probability bound under the assumption of sub-Gaussian noise. We then address the question of designing an adaptive variant. When neither the smoothness constant nor the noise level is known, and the learner employs AdaGrad-style rates (McMahan and Streeter, 2010; Duchi et al., 2011), we prove convergence bounds that depend solely on the maximal iterate norm. Whenever the iterates remain bounded, the standard optimal rates follow immediately. These results are obtained through novel proof methods for SDA, which may be of independent interest.

Organization of the Paper. The remainder of the paper is organized as follows. Section 2 surveys related work on dual averaging and its variants. Section 3 formalizes the optimization framework and introduces our notation. Section 4 establishes convergence rates of SDA with deterministic step sizes of order $1/\sqrt{t}$. In Section 5, we provide high-probability bounds. Finally, in Section 6, we extend the analysis to the parameter-free setting, deriving bounds for AdaGrad-style adaptive step sizes.

2 RELATED WORK

Dual Averaging was introduced by Nesterov (2009) for the deterministic case. The extension of DA to the stochastic case was done in Xiao (2010), focusing on the case of composite losses. These algorithms are also known in the online convex optimization setting (Gordon, 1999; Zinkevich, 2003; Orabona, 2019), a generalization of the stochastic setting, as Follow-The-Regularized-Leader with linear losses (Gordon, 1999; Shalev-Shwartz and Singer, 2006a,b; Abernethy et al., 2008; Hazan and Kale, 2008). All these analyses are for convex functions.

In the non-convex setting, the literature on dual averaging is very sparse. Clearly, when the learning rate is constant, SGD and SDA coincide in the unconstrained setting. In this specific case, the seminal paper of Ghadimi and Lan (2012) showed that SGD with the optimal constant learning rate achieves a $\mathcal{O}(1/T + \sigma/\sqrt{T})$ convergence in expectation for the squared norm of the gradient of a random iterate. Under sub-Gaussian noise, they also proposed a two-stage procedure to select an iterate with small gradient with high probability. Li and Orabona (2019) proved that SGD on non-convex smooth functions adapts to the noise level when using a delayed version of the AdaGrad-norm step sizes (Streeter and McMahan, 2010), achieving the same rate above without knowledge of σ. Later, Ward et al. (2019) established a similar guarantee for the original AdaGrad-norm step sizes, but under the stronger assumption of bounded gradients. To the best of our knowledge, no results of this form are known for the SDA version of AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011).

Suggala and Netrapalli (2020) showed that follow-the-perturbed-leader, a variant of the follow-the-regularized-leader, where the regularization is achieved through noise with a minimization oracle, can be used for online learning with non-convex losses. They also showed that non-randomized algorithms cannot achieve vanishing average regret in the same setting. There are analyses for modified versions of DA, where

the gradients are rescaled by a "learning rate" (see, e.g., Orabona and Pál, 2021). However, these variations essentially imitate the behavior of SGD in having a decreasing weight for the stochastic gradients, defeating the core motivation of DA of giving equal weights to all gradients in the iterates.

Jelassi and Defazio (2020) provided empirical evidence that DA performs well for the stochastic optimization of smooth non-convex functions. They also provided a convergence rate, but their bound requires the stochastic subgradients to have decreasing weights and that the norm of the iterates is bounded by some constant R. Moreover, it fails to recover fast rates $\mathcal{O}(1/T)$ in the low noise regime $\sigma = 0$.

More recently, Chen and Hazan (2024) proposed the open problem that stochastic smooth optimization can be directly reduced to online learning. If true, it would imply that the standard regret analysis of DA would immediately give us a convergence guarantee for stochastic smooth optimization. However, as far as we know, this is still an open problem.

3 PROBLEM SETUP

We consider the unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the objective function. Let $\widehat{\nabla} f(\boldsymbol{x})$ denote the noisy gradient at the point $\boldsymbol{x}$. For the iterates of stochastic dual averaging (SDA), we set $\boldsymbol{g}_t := \widehat{\nabla} f(\boldsymbol{x}_t)$ and denote by $\boldsymbol{\xi}_t$ the corresponding noise. Our analysis focuses on the convergence of SDA:

$$\boldsymbol{x}_{t+1} = -\eta_t \sum_{i=1}^{t} \boldsymbol{g}_i, \quad \boldsymbol{\xi}_t := \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t), \quad t \geq 1, \tag{1}$$

where η_t is a sequence of step sizes and $\widehat{\nabla} f(\boldsymbol{x}_t)$ is a gradient oracle with noise $\boldsymbol{\xi}_t$. We also choose $\boldsymbol{x}_1 = \mathbf{0}$ for simplicity.[2] Note that when the step sizes are constant, we recover the SGD algorithm. If $\boldsymbol{\xi}_t = \mathbf{0}$ almost surely, we recover the classic dual averaging algorithm. The distance between solutions in $\mathbb{R}^d$ is measured by the ℓ_2 norm. In this paper, depending on the specific theorem, we use a subset of the following assumptions.

Assumption 1 (L-smoothness)**.** *The function f is differentiable and there exists a constant $L > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \ . \tag{A1}$$

Assumption 2 (Lower boundedness)**.** *f admits a finite lower bound, i.e.,*

$$f^* := \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) > -\infty \ . \tag{A2}$$

Assumption 3 (Unbiased estimator)**.** *We have access to a history of independent, unbiased gradient estimator $\widehat{\nabla} f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$, i.e.,*

$$\mathbb{E}\left[\widehat{\nabla} f(\boldsymbol{x}) \mid \boldsymbol{x}\right] = \nabla f(\boldsymbol{x}) \ . \tag{A3}$$

In most classical analyses of stochastic first-order methods one typically assumes access to an oracle with uniformly bounded variance (BV) (Harold et al., 1997), i.e., $\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}) - \widehat{\nabla} f(\boldsymbol{x})\right\|^2\right] \leq \sigma^2$ for every $\boldsymbol{x}$, yet it is provably too restrictive, even for the basic mini-batch least-squares problem, where gradient noise grows with the iterate norm so no finite global σ^2 is admissible (Jain et al., 2018; Zhang and Zhou, 2019). Hence, we use the following assumption.

Assumption 4 (Noisy-strong-growth condition)**.** *The oracle $\widehat{\nabla} f(\cdot)$ satisfies the (ρ, σ)-noisy-strong-growth-condition, i.e., for all $\boldsymbol{x} \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2\right] \leq \rho \|\nabla f(\boldsymbol{x})\|^2 + \sigma^2 \ . \tag{A4}$$

This assumption on the noise is similar to the one used in Schmidt and Le Roux (2013); Bottou (2010); Bottou et al. (2018). It *strictly generalizes* the standard BV assumption (for which only $\rho = 0$ is admitted). By allowing the stochastic error to scale with the signal, the admissible σ^2 can be chosen to reflect only the residual noise at the optimum; in practice, this constant is often orders of magnitude smaller than the global bound required when $\rho = 0$, leading to larger stable step sizes and tighter convergence guarantees. Subsequent work (Mishkin, 2020; Vaswani et al., 2019; Solodkin et al., 2024) showed that many over-parameterized (interpolating) models satisfy a strong-growth condition, i.e., $\sigma = 0$, and leveraged the condition to match accelerated deterministic rates without variance-reduction techniques.

Assumption 5 (Sub-Gaussian noise)**.** *We assume that $\left\|\widehat{\nabla} f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|$ is a σ-sub-Gaussian random variable. 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*

[2] This is without loss of generality: for an arbitrary initial point $\boldsymbol{x}_1$, a coordinate shift $\Delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}_1$ reduces the problem to $\min_{\Delta\boldsymbol{x}} f(\Delta\boldsymbol{x} + \boldsymbol{x}_1)$ initialized at $\Delta\boldsymbol{x}_1 = \mathbf{0}$, and all convergence bounds carry over with Δ replaced by $f(\boldsymbol{x}_1) - f^*$.

(2018)). For convenience, we adopt the following definition. For all λ such that $|\lambda| \leq 1/\sigma$,

$$\mathbb{E}\left[\exp\left(\lambda^2 \left\|\widehat{\nabla} f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2\right) \mid \boldsymbol{x}\right] \leq \exp\left(\lambda^2 \sigma^2\right) . \tag{A5}$$

Our results depend on the standard Assumptions (A1)–(A4). For high probability guarantees, we use Assumption (A5) instead of Assumption (A4).

4 BOUNDS IN EXPECTATION UNDER THE NOISY-STRONG-GROWTH CONDITION

In this section, we study the convergence of SDA under the (ρ, σ)-noisy-strong-growth condition, when using decreasing step sizes of the form

$$\eta_t = \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})}, \tag{2}$$

where $\alpha = \min\{\sigma/(L(1+\rho)), 1\}$. We show in Theorem 4.3, that this schedule leads to a convergence rate for the average squared gradient norm of

$$\mathcal{O}\left(\frac{1}{T} + \frac{\sigma \log T}{\sqrt{T}}\right),$$

where $\mathcal{O}(\cdot)$ hides polynomial dependence on the problem parameters L, σ, and ρ. Thus, we recover the standard performance guarantees of SGD in the smooth non-convex setting, up to a logarithmic factor in T in the slow-rate term.[3] Moreover, in regimes where the noise level is small relative to the gradient norm, the procedure achieves a better rate of $\mathcal{O}(1/T)$, comparable to the fast rates observed in noiseless or low-variance scenarios. The core idea of the analysis lies in reinterpreting SDA updates as an instance of SGD over a sequence of functions $\{f_t\}_{t \leq T}$ defined below.

Main Challenges. Traditional analysis of dual averaging in the convex setting goes through without pain, for the existence of linear surrogates of the loss $\tilde{\ell}_t(\boldsymbol{x}) := \langle \sum_{i=1}^t \boldsymbol{g}_i, \boldsymbol{x}\rangle + \psi_t(\boldsymbol{x})/\eta_t$, for some sequence of regularizers $\psi_t(\cdot)$. Bregman telescoping yields $\mathcal{O}(\sqrt{T})$ regret and hence $\mathcal{O}(1/\sqrt{T})$ optimization error in expectation after averaging (Xiao, 2010). In the non-convex smooth case, however, we immediately lose this property. In general, especially with decreasing step sizes, every new iterate $\boldsymbol{x}_{t+1}$ depends on the entire history of past stochastic gradients with non-vanishing weights. So viewing stochastic dual averaging (SDA) as an explicit primal descent step (1), we run into unavoidable coupling terms like $\mathbb{E}\left[\langle \sum_{i=1}^{t-1} \boldsymbol{g}_i, \nabla f(\boldsymbol{x}_t)\rangle\right]$, which classical analysis techniques cannot handle since a proper *descent lemma* is missing. Recent literature confirms that new techniques are needed. Liu et al. (2023) analyzed non-convex SDA by replacing the convergence guarantee with the surrogate $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2$ and obtained only an $\mathcal{O}(1/t)$ decay. Cutkosky (2019) proposed an online-to-batch framework that bypasses the coupling by inserting an auxiliary projection, but at the price of departing from the original SDA update that we care about.

Dual Averaging as SGD on Implicitly Regularized Functions. Our starting point is the observation by Jelassi and Defazio (2020) that the iterates of SDA are the same as the ones of SGD on a modified sequence of functions.

Propositon 1 ((Jelassi and Defazio, 2020))**.** *Let $\boldsymbol{x}_1 = \boldsymbol{0}$, $\eta_0 = \eta_1$ and $\{f_t\}_{t=1}^T$ be a sequence of functions defined as*

$$f_t(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\gamma_t}{2}\|\boldsymbol{x}\|_2^2, \quad t \geq 1,$$

where $\gamma_t := \eta_t^{-1} - \eta_{t-1}^{-1}$. Then, the SDA update (1) is equivalent to SGD with learning rate η_t on the function $f_t(\boldsymbol{x})$ for all $t \geq 1$.

This result is an immediate consequence of the update of SDA:

$$\begin{aligned}\boldsymbol{x}_{t+1} &= -\eta_t\left(\boldsymbol{g}_t + \sum_{i=1}^{t-1}\boldsymbol{g}_i\right) \\ &= \boldsymbol{x}_t - \eta_t\left(\boldsymbol{g}_t + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\boldsymbol{x}_t\right) .\end{aligned}$$

Remark 4.1. *Despite the equivalence result, SDA and SGD generate* different iterate trajectories *because they weight past gradients differently (as discussed in Section 1). From a technical perspective, the standard smooth-SGD convergence analysis is insufficient here: the equivalence transforms the problem into a* sequence *of time-varying objectives $\{f_t\}$, and controlling $\{\nabla f_t(\boldsymbol{x}_t)\}$ does not automatically yield control over $\{\nabla f(\boldsymbol{x}_t)\}$.*

Given that $f(\cdot)$ is smooth, we have $f_t(\cdot)$ is L_t-smooth with $L_t := L + \gamma_t$. Therefore, leveraging Proposition 1 and the L_t-smoothness of every surrogate f_t, we have

[3] The logarithmic term is due to the time-varying learning rates, and one would suffer a similar term in SGD with time-varying learning rates.

for $t \geq 1$,

$$\sum_{t=1}^{T} \eta_t \left(1 - \frac{\eta_t L_t}{2}\right) \mathbb{E}\left[\|\nabla f_t(\boldsymbol{x}_t)\|^2\right] + \sum_{t=1}^{T} \mathbb{E}\left[f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right] \leq \Delta + \sum_{t=1}^{T} \frac{\eta_t^2 L_t}{2} \mathbb{E}\left[\|\boldsymbol{\xi}_t\|^2\right], \tag{3}$$

where $\Delta := f(\mathbf{0}) - f^*$, and by convention we let $\gamma_0 := 0$ and hence $f_0 := f$. Equation (3) generalizes the classical SGD descent inequality from the stationary case $f_t \equiv f$ to an online setting in which the learner faces a time-varying sequence of smooth losses; so far, the derivation remains agnostic to how these losses are generated by the SDA updates. In our context, however, the surrogates possess the special property

$$f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}_t) = f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t) + \frac{\gamma_{t-1} - \gamma_t}{2} \|\boldsymbol{x}_t\|^2,$$

which shows that up to a telescopic term, the term $\sum_{t=1}^{T} \mathbb{E}\left[f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right]$ in (3) carries a positive component $\left((\gamma_{t-1} - \gamma_t)/2\right) \mathbb{E}\left[\|\boldsymbol{x}_t\|^2\right]$, allowing us to relate the obtained bound on the gradients of f_t to one on the gradients of objective function f. To achieve this, a particular selection of learning rates, detailed in the lemma below, is required.

Lemma 4.2. *Consider the step sizes*

$$\eta_t = \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})},$$

where $\alpha = \min\left\{\sigma/\left(L(1+\rho)\right), 1\right\}$. *For any* $t \geq 2$, *we have,*

$$\eta_t L_t \leq 2,$$
$$\frac{1}{3}\eta_t - \frac{1}{6}\eta_t^2 L_t - \frac{1}{2}\rho\eta_t^2 L_t \geq \frac{\eta_t}{9},$$
$$\frac{\gamma_{t-1} - \gamma_t}{2} - \frac{1}{2}\gamma_t^2\eta_t + \frac{1}{4}\gamma_t^2\eta_t^2 L_t \geq 0 \ .$$

Theorem 4.3. *Suppose* (A1)–(A4) *holds, and let* $\boldsymbol{x}_1 = \mathbf{0}$, *then SDA iterates with step sizes*

$$\eta_t = \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})}, \tag{4}$$

where $\alpha = \min\left\{\sigma/\left(L(1+\rho)\right), 1\right\}$, *satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \left(11\Delta + C(\sigma, T)\right) \cdot \frac{L(1+\rho)^2 + \sigma\sqrt{T}}{T},$$

where $\Delta := f(\mathbf{0}) - f^*$ *and*

$$C(\sigma, T) := \frac{81}{4}\left(\frac{\sigma^2}{L(1+\rho)} + L(1+\rho)\right)\log(1+\alpha\sqrt{T})$$

is a polylog term in T.

Proof sketch. A detailed proof is presented in the Appendix. We consider the equivalence argument introduced in Proposition 1. Let $\widehat{\nabla} f_t(\boldsymbol{x}) := \widehat{\nabla} f(\boldsymbol{x}) + \gamma_t \boldsymbol{x}$. Because $f_t(\cdot)$ is L_t-smooth, the standard descent inequality gives

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

Adding and subtracting $f_{t-1}(\boldsymbol{x}_t)$ inside the left-hand side makes the term $f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t)$ telescope once we sum over t. Taking expectations produces the online descent inequality (3); the proof then specializes the bound by plugging in the noise condition and the step size schedule. Using the noise bound $\mathbb{E}_{t-1}\left[\|\boldsymbol{\xi}_t\|^2\right] \leq \rho\|\nabla f(\boldsymbol{x}_t)\|^2 + \sigma^2$, we arrive at

$$\sum_{t=1}^{T} \mathbb{E}[A_t] \leq f(\mathbf{0}) - f^* + \frac{\sigma^2}{2}\sum_{t=1}^{T} \eta_t^2 L_t,$$

where

$$A_t = \eta_t\left(1 - \frac{\eta_t L_t}{2}\right)\|\nabla f_t(\boldsymbol{x}_t)\|^2 + \frac{\gamma_{t-1} - \gamma_t}{2}\|\boldsymbol{x}_t\|^2 - \frac{\rho}{2}\eta_t^2 L_t \|\nabla f(\boldsymbol{x}_t)\|^2 \ .$$

Choosing the step sizes as in (4), Lemma 4.2 ensures that

$$A_t \geq \frac{\eta_t}{9}\|\nabla f(\boldsymbol{x}_t)\|^2 \ .$$

Substituting this lower bound and evaluating the resulting sums yields

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{11\Delta}{\eta_T} + \frac{81\sigma^2\log(1+\alpha\sqrt{T})}{4\eta_T L(1+\rho)\alpha^2},$$

which coincides with the statement of Theorem 4.3. □

Our bound scales as $\mathcal{O}\left(1/T + \sigma\log(T)/\sqrt{T}\right)$. When the noise level satisfies $\sigma > 0$, this matches the min-max optimal rate for stochastic gradient descent established by Ghadimi and Lan (2012), up to a $\log T$ factor. Note that the learning rates we used *do not* require a priori knowledge of the time horizon T, whereas the optimal step sizes in Ghadimi and Lan (2012) are

tuned using this information. In the low noise regime ($\sigma = 0$)—which, under the noisy-strong-growth condition, corresponds to gradient noise being on the same order as the true gradient—our analysis shows that the accelerated $1/T$ rate is attainable, mirroring the best-known guarantees for standard SGD.

Comparison with Jelassi and Defazio (2020). We want to stress that our result is not just a simple derivation from Proposition 1. Indeed, even if Jelassi and Defazio (2020) started from the same proposition, they did not prove a guarantee for the standard Dual Averaging (DA). In fact, i) Jelassi and Defazio (2020, Theorem 3.1) only applies to a variant they call Modernized Dual Averaging (MDA) algorithm; ii) they also require their scaling parameters (λ_t, β_t) to grow like $\sqrt{t+1}$. Although the MDA algorithm encompasses DA as a special case when $\beta_t \equiv 1$ and $\lambda_t = \eta_t$, the specific choice of parameters in their Theorem 3.1 (the only one for which they guarantee convergence) does not, as *the gradients do not enter into the iterate with equal weights*.

Even further, their guarantee is stated under the *a priori* assumption that the iterates remain in a fixed Euclidean ball, i.e., $\sup_{t\geq 1} \|\boldsymbol{x}_t\| \leq R$ for some known radius R. Instead, our bound in Theorem 4.3 holds *without* any explicit control on the iterates and is therefore *iterate-independent* and, as a result, proves that SDA converges at the optimal rates. Removing this constraint is not trivial, and it requires a different and novel proof strategy. Furthermore, even in settings where the iterates are bounded, the result of Jelassi and Defazio (2020) fails to recover the fast $\mathcal{O}(1/T)$ rate attainable in low-noise regimes, whereas our analysis does.

5 HIGH PROBABILITY CONVERGENCE OF SDA

The bound we have just established controls *the average* performance over a hypothetical ensemble of infinitely many independent runs. Such guarantees in-expectation are of theoretical value, but they do not tell us what happens in typical machine learning applications, where practitioners only perform a few number of runs rather than a whole ensemble. Therefore, to move closer to this practical setting, a stronger result is desired, to ensure that each run will, with probability at least $1-\delta$, stay within the required error tolerance. To achieve a high-probability bound amounts to controlling the *entire tail* of the error distribution, which apparently requires stronger assumptions on the stochastic noise distribution. Empirical evidences (Zhang et al., 2020; Simsekli et al., 2019) has shown that the noise distribution for standard vision tasks can be well-approximated by a sub-Gaussian distribution. In this section, we adopt the noisy-strong-growth condition (4) with $\rho = 0$ and further add Assumption (A5) to our assumption set, and show in Theorem 5.4 a high probability convergence guarantee of SDA.

We introduce the following notations for $t \geq 1$

$$\widehat{\nabla} f_t(\boldsymbol{x}) := \widehat{\nabla} f(\boldsymbol{x}) + \gamma_t \boldsymbol{x}, \quad \Delta_\tau f_t := f_t(\boldsymbol{x}_\tau) - f^* \ .$$

We let $\mathcal{F}_t := \sigma(\xi_1, \ldots, \xi_{t-1})$ denote the natural filtration. Note that $\boldsymbol{x}_t$ is $\mathcal{F}_t$-measurable. The following lemma serves as a fundamental step of our analysis.

Lemma 5.1. *For $t \geq 1$, we have*

$$\begin{aligned} C_t &:= \eta_t \left(1 - \frac{\eta_t L_t}{2}\right) \|\nabla f_t(\boldsymbol{x}_t)\|^2 + \Delta_{t+1} f_t - \Delta_t f_t \\ &\leq \eta_t (\eta_t L_t - 1) \langle \nabla f_t(\boldsymbol{x}_t), \boldsymbol{\xi}_t \rangle + \frac{\eta_t^2 L_t}{2} \|\boldsymbol{\xi}_t\|^2 \ . \quad (5) \end{aligned}$$

We can concentrate the LHS of (5) using a similar argument from Liu et al. (2023). For all $1 \leq t \leq T$ and $w_t \geq 0$, let

$$Z_t := w_t C_t - v_t \|\nabla f_t(\boldsymbol{x}_t)\|^2, \quad S_t := \sum_{i=t}^{T} Z_i,$$

where $v_t = 3\sigma^2 w_t^2 \eta_t^2 (\eta_t L_t - 1)^2$. Crucially, the exponential moment argument in the proof of Theorem 4.3 in Liu et al. (2023) holds for any $\mathcal{F}_t$-measurable sequence $\{\boldsymbol{a}_t\}$ (i.e., depending only on the noise realizations $\xi_1, \ldots, \xi_{t-1}$), not just the specific choice $\boldsymbol{a}_t = \nabla f(\boldsymbol{x}_t)$. Using this observation, we can prove the following key inequality.

Theorem 5.2. *Suppose for all $1 \leq t \leq T$, w_t and η_t satisfy*

$$0 \leq w_t \eta_t^2 L_t \leq \frac{1}{2\sigma^2}, \quad (6)$$

then

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

Markov's inequality gives us the following guarantee in probability.

Corollary 5.2.1. *If condition (6) holds for all $1 \leq t \leq T$, then with probability at least $1-\delta$, we have*

$$\begin{aligned} &\sum_{t=1}^{T} \left(w_t \eta_t \left(1 - \frac{\eta_t L_t}{2}\right) - v_t\right) \|\nabla f_t(\boldsymbol{x}_t)\|^2 \\ &\qquad + \sum_{t=1}^{T} w_t (\Delta_{t+1} f_t - \Delta_t f_t) \\ &\qquad \leq 3\sigma^2 \sum_{t=1}^{T} \frac{w_t \eta_t^2 L_t}{2} + \log \frac{1}{\delta} \ . \quad (7) \end{aligned}$$

Equipped with Lemma 5.1 and Theorem 5.2, we are ready to prove the central lemma for SDA by specifying the choice of w_t that satisfy condition (6).

Lemma 5.3. *Suppose the learning rates $\{\eta_t\}$ is non-increasing with $\eta_t L_t \le 1$, then for any $\delta > 0$, and define the auxiliary function $f_0(\boldsymbol{x}) := f(\boldsymbol{x}) + \gamma_0/2 \|\boldsymbol{x}\|^2$, where γ_0 is an arbitrary constant. Then, the following event holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} \eta_t \|\nabla f_t(\boldsymbol{x}_t)\|^2 + 2\left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1\right) + \sum_{t=1}^{T} \left(f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right) \le 3\sigma^2 \sum_{t=1}^{T} \eta_t^2 L_t + 12\sigma^2 \eta_1 \log\frac{1}{\delta} \,. \tag{8}$$

Theorem 5.4. *Assume* (A1)–(A3)*, and* (A5) *are satisfied. Let $\boldsymbol{x}_1 = \boldsymbol{0}$. Setting $\eta_t = 1/\left(L + \sigma\sqrt{t}\right)$, then with probability at least $1-\delta$, the iterate sequence $\{\boldsymbol{x}_t\}_{t\ge 1}$ output by SDA satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2 \le (4\Delta + G(\sigma,\delta,T)) \cdot \frac{L + \sigma\sqrt{T}}{T},$$

where $\Delta := f(\boldsymbol{0}) - f^*$ and

$$G(\sigma,\delta,T) := 12\sigma \log\left(1 + \frac{\sigma\sqrt{T}}{L}\right) + \frac{24\sigma^2}{L+\sigma}\log\frac{1}{\delta}$$

is a polylog term in T.

Again, in the case where the time horizon T is unknown to the algorithm, by choosing the step size η in Theorem 5.4, the bound is adaptive to noise, i.e, when $\sigma = 0$ we recover $\mathcal{O}(1/T)$ convergence rate of the (deterministic) gradient descent algorithm.

6 ADAPTIVE DUAL AVERAGING

In the previous section, we showed that it is possible to match the rate of SGD with SDA on non-convex smooth functions. However, the optimal learning rate in (2) depends on the unknown variance of the noise, σ^2. Here, we show that, as is the case for SGD, we can design an adaptive version of SDA, using AdaGrad-norm step sizes (Streeter and McMahan, 2010; McMahan and Streeter, 2010; Duchi et al., 2011). The use of such inverse-root step sizes with dual averaging has already been explored empirically in prior work, both for convex optimization and deep learning tasks (Duchi et al., 2011; Defazio and Jelassi, 2022; Huang and Lee, 2024). The adaptive step sizes are given by

$$\eta_t = \frac{\eta}{\sqrt{\gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2}},$$

with parameters $\eta, \gamma > 0$. Using this choice, we can bound in Theorem 6.1 the average squared gradient as

$$\mathcal{O}\left(\sigma\sqrt{\frac{\mathbb{E}\left[B_T^2\right]}{T}} + \frac{\mathbb{E}\left[B_T^2\right]}{T}\right),$$

with

$$B_T := \max_{t\le T} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2, \tag{9}$$

where $\boldsymbol{x}^*$ is an arbitrary stationary point. Hence, when the iterates remain bounded we recover the optimal rate. Similar bounds with a dependence on the maximal iterate norm, were presented in earlier works for SGD with adaptive step sizes (Duchi et al., 2011).

Theorem 6.1. *Assume* (A1)–(A4) *are satisfied. Let $\boldsymbol{x}_1 = \boldsymbol{0}$. Consider SDA iterates using step sizes*

$$\eta_t = \frac{\eta}{\sqrt{\gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2}} \,.$$

Let $\boldsymbol{x}^*$ be an arbitrary stationary point, and define B_T as in (9)*. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \le \left(\frac{2\Delta}{\eta} + H(T)\right)\cdot\frac{1}{T} + \frac{J(\sigma,T)}{\sqrt{T}},$$

where $\Delta = f(\boldsymbol{0}) - f^*$ and

$$H(T) := 2\sqrt{2}L^2\left(\sqrt{2}+\rho\right)\mathbb{E}\left[\left(\frac{13B_T}{2\eta} + 2\eta\right)^2\right] + \frac{\gamma}{2},$$

$$J(\sigma,T) := 2\sqrt{2}L\sigma\left(\mathbb{E}\left[\left(\frac{13B_T}{2\eta} + 2\eta\right)^2\right]\right)^{1/2}.$$

Proof sketch. Full details are deferred to the Appendix; here we record the main chain of inequalities in continuous prose. Let

$$\boldsymbol{\theta}_t := \sum_{i=1}^{t} \boldsymbol{g}_i, \quad S_t := \gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2, \quad t \ge 1 \,.$$

Note that $\eta_t = \eta/\sqrt{S_t}$ for $t \ge 1$. Because SDA keeps the dual accumulator $\boldsymbol{\theta}_t$, its recursion can be written as

$$\begin{aligned}\boldsymbol{x}_{t+1} &= -\eta_t \boldsymbol{\theta}_t \\ &= \boldsymbol{x}_t - \eta_t\left(\boldsymbol{g}_t + (1 - \eta_{t-1}/\eta_t)\boldsymbol{\theta}_{t-1}\right) \\ &= \boldsymbol{x}_t - \eta_t \boldsymbol{g}'_t \,.\end{aligned}$$

Now smoothness of $f(\cdot)$ yields the descent estimate

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) \le -\eta_t \left\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{g}'_t \right\rangle + \frac{L}{2}\eta_t^2 \|\boldsymbol{g}'_t\|^2 \,.$$

Denote $\Delta_t := f(\boldsymbol{x}_t) - f^*$ for $t \ge 1$. Summing over t and expanding $\boldsymbol{g}'_t$ produces a telescoping term $\Delta_1/\eta +$

$\sum_{t=1}^{T}(1/\eta_t - 1/\eta_{t-1})\Delta_t$ plus the variance contribution $(L/2)\sum_{t=1}^{T}\eta_t \left\|\boldsymbol{g}_t'\right\|^2$, which reads

$$\begin{aligned}
&\sum_{t=1}^{T}\langle\nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{g}_t\rangle \\
&\quad\le \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\left(\langle\nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{x}_t\rangle - L\left\|\boldsymbol{x}_t\right\|^2\right)\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right) \\
&\qquad - \left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\Delta_t + L\eta\sum_{t=1}^{T}\frac{\left\|\boldsymbol{g}_t\right\|^2}{\sqrt{S_t}}\ .
\end{aligned}$$

Since $\sum_{t=1}^{T}\left\|\boldsymbol{g}_t\right\|^2/\sqrt{S_t} \le 2\sqrt{S_T}$, the last term on the RHS collapses to $2L\eta\sqrt{S_T}$. Next, the strong-growth condition together with the smoothness bound $\Delta_t \le (L/2)\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|^2$ and the inequality $\|\nabla f(\boldsymbol{x}_t)\| \le L\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|$ allow every term involving $\boldsymbol{x}_t$ or $\boldsymbol{x}_t - \boldsymbol{x}^*$ to be controlled by A single radius B_T. A direct rearrangement then shows

$$\begin{aligned}
&\sum_{t=1}^{T}\langle\nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{g}_t\rangle \\
&\quad\le \frac{\Delta_1}{\eta} + 2\left(\frac{13LB_T}{2\eta}+2L\eta\right)^2 + \frac{1}{4}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2 \\
&\qquad + \sqrt{2}\left(\frac{13LB_T}{2\eta}+2L\eta\right)\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2} + \frac{\gamma}{4}\ .
\end{aligned}$$

Finally, taking expectations and applying Cauchy-Schwarz to the noise term changes $\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2}$ to $\sigma\sqrt{T}$, completing the derivation of the bound claimed in the theorem. □

Remark 6.2. *Observe that the upper bound is minimized when $\gamma = 0$. Indeed, we can safely use $\gamma = 0$ by noticing that we can avoid divisions by zero by not updating the iterates every time the stochastic gradient is zero.*

Although Theorem 6.1 delivers the desired convergence rate whenever the iterates remain bounded, it still falls short of the *iterate-independent* upper-bounds guarantees obtained in Section 4 for deterministic step sizes. The main difficulty stems from the SDA update, given by $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\eta_t\boldsymbol{g}_t - (1-\eta_t/\eta_{t-1})\,\boldsymbol{x}_t$, which includes an extra term in $\boldsymbol{x}_t$ absent from the SGD recursion. With deterministic step sizes, the ratio between factors η_t and $|\eta_t/\eta_{t-1}-1|$ is of order $1/\sqrt{t}$, so this term can be controlled; under the AdaGrad rule no such uniform decay holds. As a result, the update direction becomes a delayed, biased combination of past gradients, and in non-convex landscapes those past gradients could be uninformative for the current update. This difficulty does not arise in plain SGD and is akin to challenges in the analysis of momentum-type methods such as Adam. Classical techniques such as the descent lemma therefore do not suffice, and developing tight, iterate-independent convergence guarantees for adaptive SDA remains an open research avenue.

7 CONCLUSION, LIMITATIONS AND FUTURE WORK

We have revisited *stochastic dual averaging* through the lens of smooth non-convex optimization and provided the first *iterate-level* guarantees that match the $\mathcal{O}(T^{-1/2})$ complexity enjoyed by carefully tuned SGD, without restrictive assumptions. Concretely, we first proved sharp convergence bounds for deterministic step sizes under the noisy-strong-growth condition; next, we extended these guarantees to high-probability statements in the presence of sub-Gaussian noise; and finally, we established adaptive rates for AdaGrad-style learning rates that apply whenever the iterates remain bounded.

Limitations and Future Work. Given that our analysis is the first one of its kind for SDA, there are naturally some open questions and possible extensions. First, the high probability guarantees rely on sub-Gaussian gradient noise. It might be interesting to extend it to heavy-tailed distributions. Second, our analysis focuses on the unconstrained setting; extending it to constrained non-convex optimization requires a different stationarity measure (e.g., the Frank–Wolfe gap) and is a natural direction for future work. Third, the adaptive Ada-DA variant only converges provably when the iterates are bounded, obtaining truly iterate-independent, adaptive bounds remains an open problem.

A promising next step is to derive *iterate-independent* guarantees for SDA equipped with fully adaptive step sizes, paralleling the bounds already known for SGD with AdaGrad. Achieving this will likely require new control of the cumulative gradient history; one avenue is to adapt the bias-correction arguments introduced for momentum-type first-order methods. In particular, the analyses developed in the case of Adam and related optimizers (see, e.g., Zhou et al., 2018; Chen et al., 2019) face the same core challenge of handling updates that combine all past gradients with time-varying weights.

Another interesting direction for extension is to replace the Euclidean geometry underlying our surrogate functions with a Bregman geometry induced by a regularizer ϕ. Our analysis of SDA relies on its equivalence to online gradient descent applied to a sequence of losses

with Euclidean regularization. In the same spirit, running FTRL with a regularizer ϕ is equivalent to online mirror descent with the same regularizer. Extending our arguments to this broader setting, however, requires replacing the smoothness assumptions on f with respect to the Euclidean distance by conditions tailored to the geometry induced by ϕ (for example, by relative smoothness introduced by Lu et al. (2018)). Establishing the precise form of these arguments, as well as identifying any additional assumptions needed on ϕ is left for future work.

References

J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Rocco A. Servedio and Tong Zhang, editors, *Proc. of Conference on Learning Theory (COLT)*, pages 263–274. Omnipress, 2008.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

X. Chen and E. Hazan. Open problem: Black-box reductions & adaptive gradient methods. *Proceedings of Machine Learning Research vol*, 196:1–8, 2024.

X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

A. Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International conference on machine learning*, pages 1446–1454. PMLR, 2019.

Aaron Defazio and Samy Jelassi. A momentumized, adaptive, dual averaged gradient method. *Journal of Machine Learning Research*, 23(144):1–34, 2022.

J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4): 1469–1492, 2012.

G. J. Gordon. Regret bounds for prediction problems. In *Proc. of the twelfth annual conference on Computational learning theory (COLT)*, pages 29–40, 1999.

J Harold, G Kushner, and George Yin. Stochastic approximation and recursive algorithm and applications. *Application of Mathematics*, 35(10), 1997.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *Proc. of the 21st Conference on Learning Theory*, 2008.

J. He, X. Jia, J. Xu, L. Zhang, and L. Zhao. Make ℓ_1 regularization effective in training sparse CNN. *Computational Optimization and Applications*, 77 (1):163–182, 2020.

Zih-Syuan Huang and Ching-pei Lee. Regularized adaptive momentum dual averaging with an efficient inexact subproblem solver for training structured neural network. *Advances in Neural Information Processing Systems*, 37:128489–128515, 2024.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of machine learning research*, 18(223):1–42, 2018.

S. Jelassi and A. Defazio. Dual averaging is surprisingly effective for deep learning optimization. *arXiv preprint arXiv:2010.10502*, 2020.

X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proc. of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, 2019.

Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR, 2023.

H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

H. B. McMahan and M. J. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

H Brendan McMahan. A unified view of regularized dual averaging and mirror descent with implicit updates. *arXiv preprint arXiv:1009.3240*, 2010.

A. Mishkin. *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia, 2020.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120 (1):221–259, 2009. Received: 29 September 2005 / Accepted: 13 January 2007 / Published online: 19 June 2007.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

F. Orabona and D. Pál. Parameter-free stochastic optimization of variationally coherent functions. *arXiv preprint arXiv:2102.00236*, 2021.

F. Orabona and T. Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems*, pages 2160–2170, 2017.

M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

S. Shalev-Shwartz and Y. Singer. Online learning meets optimization in the dual. In *International Conference on Computational Learning Theory*, pages 423–437. Springer, 2006a.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In *Advances in neural information processing systems*, pages 1265–1272, 2006b.

U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

V. Solodkin, S. Chezhegov, R. Nazikov, A. Beznosikov, and A. Gasnikov. Accelerated stochastic gradient method with applications to consensus problem in Markov-varying networks. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 69–86. Springer, 2024.

M. Streeter and H. B. McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

A. S. Suggala and P. Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of SGD for over-parameterized models (and an accelerated Perceptron). In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

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

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

L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

H. Yuan, M. Zaheer, and S. Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.

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

L. Zhang and Z.-H. Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the $O(1/t)$ convergence rate. In *Conference on Learning Theory*, pages 3160–3179. PMLR, 2019.

D. Zhou, J. Chen, Y. Cao, Z. Yang, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. of the International Conference on Machine Learning*, pages 928–936, 2003.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Justification: Section 3 is dedicated to describing the problem under consideration and presenting all the assumptions used throughout the paper.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Justification: Sections 4, 5, and 6 provide convergence rate analysis and complexity bounds for SDA under different settings.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable] Justification: This is a theoretical paper without experiments.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Justification: All theorems clearly state which assumptions are required, e.g., Theorem 4.3 specifies assumptions (A1)–(A4).
   (b) Complete proofs of all theoretical results. [Yes] Justification: For each theorem, we provide a proof sketch in the main text and a complete, detailed proof in the appendix.
   (c) Clear explanations of any assumptions. [Yes] Justification: Section 3 provides detailed explanations for each assumption, including their motivation and relationship to prior work.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable] Justification: This paper is theoretical in nature, with a focus on proving the convergence of DA, a well-known algorithm in the optimization literature. We did not present any experimental results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable] Justification: No existing code, data, or model assets were used.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable] Justification: Again, the paper is purely theoretical. No crowdsourcing or human subjects research was conducted.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Appendix

A.1 Missing Proofs in Section 4

Theorem (Restatement of Theorem 4.3)**.** *Suppose* (A1)–(A4) *holds, and let* $\boldsymbol{x}_1 = \mathbf{0}$*, then SDA iterates with step sizes*

$$\eta_t = \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})},$$

where $\alpha = \min\left\{\sigma/\left(L(1+\rho)\right), 1\right\}$*, satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] \leq \left(11\Delta + \frac{81}{4}\left(\frac{\sigma^2}{L(1+\rho)} + L(1+\rho)\right)\log(1+\alpha\sqrt{T})\right)\frac{L(1+\rho)^2 + \sigma\sqrt{T}}{T},$$

where $\Delta = f(\mathbf{0}) - f^*$.

Proof. Recall that $\boldsymbol{\theta}_t = \sum_{i=1}^{t}\boldsymbol{g}_i$. We have

$$\begin{aligned}
\boldsymbol{x}_{t+1} &= -\eta_t(\boldsymbol{\theta}_{t-1} + \boldsymbol{g}_t) \\
&= -\eta_{t-1}\boldsymbol{\theta}_{t-1}\frac{\eta_t}{\eta_{t-1}} - \eta_t\boldsymbol{g}_t \\
&= \boldsymbol{x}_t\left(1 - \left(1 - \frac{\eta_t}{\eta_{t-1}}\right)\right) - \eta_t\boldsymbol{g}_t \\
&= \boldsymbol{x}_t - \eta_t\left(\boldsymbol{g}_t + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\boldsymbol{x}_t\right) \ .
\end{aligned}$$

Let $\gamma_t := 1/\eta_t - 1/\eta_{t-1}$, and $f_t(\boldsymbol{x}) = f(\boldsymbol{x}) + (\gamma_t/2)\left\|\boldsymbol{x}\right\|^2$, recall that

$$\nabla f_t(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \gamma_t\boldsymbol{x} \ .$$

Therefore, the iterates of SDA using the steps sizes η_t are the same as the iterates of SGD with steps sizes η_t on the sequence of functions $\{f_t\}_{t\geq 1}$ with stochastic gradients $\widehat{\nabla} f_t(\boldsymbol{x}_t) := \boldsymbol{g}_t + \gamma_t\boldsymbol{x}_t$.

We consider $\eta_0 = \eta_1$, and take the convention $\gamma_{-1} = 0$. Define $L_t := L + \gamma_t$, since $f(\cdot)$ is L-smooth, $f_t(\cdot)$ is L_t smooth, therefore

$$\begin{aligned}
f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}_t) &\leq \left\langle \nabla f_t(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right\rangle + \frac{L_t}{2}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right\|^2 \\
&= -\eta_t\left\langle \nabla f_t(\boldsymbol{x}_t)\,,\, \widehat{\nabla} f_t(\boldsymbol{x}_t)\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\widehat{\nabla} f_t(\boldsymbol{x}_t)\right\|^2 \\
&= -\eta_t\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 - \eta_t\left\langle \nabla f_t(\boldsymbol{x}_t)\,,\, \xi_t\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 + \frac{\eta_t^2 L_t}{2}\left\|\boldsymbol{\xi}_t\right\|^2 + \eta_t^2 L_t\left\langle \nabla f_t(\boldsymbol{x}_t)\,,\, \boldsymbol{\xi}_t\right\rangle \\
&= -\eta_t\left(1 - \frac{\eta_t L_t}{2}\right)\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 - \eta_t(1 - \eta_t L_t)\left\langle \nabla f_t(\boldsymbol{x}_t)\,,\, \boldsymbol{\xi}_t\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\boldsymbol{\xi}_t\right\|^2 \ . \qquad (10)
\end{aligned}$$

Observe that

$$\begin{aligned}
f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}_t) &= f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t) + f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t) \\
&= f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t) + \frac{\gamma_{t-1} - \gamma_t}{2}\left\|\boldsymbol{x}_t\right\|^2 \ . \qquad (11)
\end{aligned}$$

Combining (10) and (11) and rearranging we obtain

$$\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\|\nabla f_t(\boldsymbol{x}_t)\|^2+\frac{\gamma_{t-1}-\gamma_t}{2}\|\boldsymbol{x}_t\|^2\le f_{t-1}(\boldsymbol{x}_t)-f_t(\boldsymbol{x}_{t+1})-\eta_t(1-\eta_t L_t)\langle\nabla f_t(\boldsymbol{x}_t),\boldsymbol{\xi}_t\rangle+\frac{\eta_t^2 L_t}{2}\|\boldsymbol{\xi}_t\|^2\ .$$

Taking the expectation, summing over t and using the bound on the noise, we obtain

$$\begin{aligned}\sum_{t=1}^T\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\mathbb{E}\left[\|\nabla f_t(\boldsymbol{x}_t)\|^2\right]+\frac{\gamma_{t-1}-\gamma_t}{2}\mathbb{E}\left[\|\boldsymbol{x}_t\|^2\right]&\le f(\boldsymbol{0})-\mathbb{E}\left[f_T(\boldsymbol{x}_{T+1})\right]+\sum_{t=1}^T\frac{\eta_t^2 L_t}{2}\mathbb{E}\left[\mathbb{E}_{t-1}\left[\|\boldsymbol{\xi}_t\|^2\right]\right]\\&\le f(\boldsymbol{0})-f^*+\sum_{t=1}^T\frac{\eta_t^2 L_t}{2}\left(\rho\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]+\sigma^2\right)\ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{t=1}^T\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\mathbb{E}\left[\|\nabla f_t(\boldsymbol{x}_t)\|^2\right]+\sum_{t=1}^T\frac{\gamma_{t-1}-\gamma_t}{2}\mathbb{E}\left[\|\boldsymbol{x}_t\|^2\right]-\sum_{t=1}^T\frac{\rho}{2}\eta_t^2 L_t\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\\&\quad\le f(\boldsymbol{0})-f^*+\sum_{t=1}^T\sigma^2\frac{\eta_t^2 L_t}{2}\ .\end{aligned}$$

Denote

$$A_t:=\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\|\nabla f_t(\boldsymbol{x}_t)\|^2+\frac{\gamma_{t-1}-\gamma_t}{2}\|\boldsymbol{x}_t\|^2-\frac{\rho}{2}\eta_t^2 L_t\|\nabla f(\boldsymbol{x}_t)\|^2\ .$$

The last bound rewrites as

$$\sum_{t=1}^T\mathbb{E}\left[A_t\right]\le\mathbb{E}\left[f(\boldsymbol{0})-f^*\right]+\frac{\sigma^2}{2}\sum_{t=1}^T\eta_t^2 L_t\ . \tag{12}$$

We consider step sizes of the form

$$\eta_t=\frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})},$$

where $\alpha=\min\left\{\sigma/\left(L(1+\rho)\right),1\right\}$. Recall that we have

$$\|\nabla f_t(\boldsymbol{x}_t)\|^2=\|\nabla f(\boldsymbol{x}_t)+\gamma_t\boldsymbol{x}_t\|^2\ge\frac{1}{3}\|\nabla f(\boldsymbol{x}_t)\|^2-\frac{1}{2}\gamma_t^2\|\boldsymbol{x}_t\|^2\ .$$

Therefore, following Lemma A.2, which gives that $1-\eta_t L_t/2\ge 0$, we have for any $t\ge 2$

$$\begin{aligned}A_t&\ge\frac{1}{3}\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\|\nabla f(\boldsymbol{x}_t)\|^2-\frac{1}{2}\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\gamma_t^2\|\boldsymbol{x}_t\|^2+\frac{\gamma_{t-1}-\gamma_t}{2}\|\boldsymbol{x}_t\|^2-\frac{\rho}{2}\eta_t^2 L_t\|\nabla f(\boldsymbol{x}_t)\|^2\\&=\left(\frac{1}{3}\eta_t-\frac{1}{6}\eta_t^2 L_t-\frac{1}{2}\rho\eta_t^2 L_t\right)\|\nabla f(\boldsymbol{x}_t)\|^2+\left(\frac{\gamma_{t-1}-\gamma_t}{2}-\frac{1}{2}\gamma_t^2\eta_t+\frac{1}{4}\gamma_t^2\eta_t^2 L_t\right)\|\boldsymbol{x}_t\|^2\\&\ge\frac{\eta_t}{9}\|\nabla f(\boldsymbol{x}_t)\|^2\ .\end{aligned}$$

The last line follows from the lower bounds of the factors of $\|\nabla f(\boldsymbol{x}_t)\|^2$ and $\|\boldsymbol{x}_t\|^2$ provided in Lemma A.2. We conclude using (12) and then the second bound of Lemma A.2 that

$$\begin{aligned}\sum_{t=2}^T\frac{\eta_t}{9}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]&\le f(\boldsymbol{0})-f^*+\frac{\sigma^2}{2}\sum_{t=1}^T\eta_t^2 L_t\\&\le f(\boldsymbol{0})-f^*+\frac{3\sigma^2}{4L(1+\rho)}\sum_{t=1}^T\frac{1}{(1+\rho+\alpha\sqrt{t})^2}\\&\le f(\boldsymbol{0})-f^*+\frac{3\sigma^2}{4L(1+\rho)}\left(1+\frac{2}{\alpha^2}\log\left(\frac{1+\rho+\alpha\sqrt{T}}{1+\rho+\alpha}\right)\right)\\&\le f(\boldsymbol{0})-f^*+\frac{3\sigma^2}{4L(1+\rho)}\left(1+\frac{2}{\alpha^2}\log(1+\alpha\sqrt{T})\right)\ .\end{aligned}$$

Therefore,

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] \le \left\|\nabla f(\mathbf{0})\right\|^2 + \frac{9}{\eta_T}(f(\mathbf{0})-f^*) + \frac{27\sigma^2}{4\eta_T L(1+\rho)}\left(1+\frac{2}{\alpha^2}\log(1+\alpha\sqrt{T})\right)$$
$$\le \left(11(f(\mathbf{0})-f^*) + \frac{81\sigma^2\log(1+\alpha\sqrt{T})}{4L(1+\rho)\alpha^2}\right)\frac{1}{\eta_T},$$

where we used the descent lemma in the last line ($\left\|\nabla f(\mathbf{0})\right\|^2 \le 2L(f(\mathbf{0})-f^*)$). Noticing

$$\frac{x^2}{\min\{x^2,1\}} \le x^2+1,$$

we have

$$\frac{81\sigma^2\log(1+\alpha\sqrt{T})}{4L(1+\rho)\alpha^2} = \frac{81}{4}\frac{\sigma^2}{L^2(1+\rho)^2}\frac{\log(1+\alpha\sqrt{T})}{\alpha^2}L(1+\rho)$$
$$\le \frac{81}{4}L(1+\rho)\left(\frac{\sigma^2}{L^2(1+\rho)^2}+1\right)\log(1+\alpha\sqrt{T})\ .$$

The conclusion follows from the expressions of α and η_T. □

A.2 Missing Proofs in Section 5

Proof of Lemma 5.1. We start from the one-step descent of $f_t(\boldsymbol{x}_t)$. Smoothness of $f_t(\cdot)$ implies that

$$f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}_t) \le \langle\nabla f_t(\boldsymbol{x}_t)\,,\,\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{L_t}{2}\left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|^2$$
$$= -\eta_t\left\langle\nabla f_t(\boldsymbol{x}_t)\,,\,\widehat{\nabla} f_t(\boldsymbol{x}_t)\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\widehat{\nabla} f_t(\boldsymbol{x}_t)\right\|^2$$
$$= -\eta_t\left\langle\nabla f_t(\boldsymbol{x}_t)\,,\,\nabla f_t(\boldsymbol{x}_t)+\boldsymbol{\xi}_t\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\nabla f_t(\boldsymbol{x}_t)+\boldsymbol{\xi}_t\right\|^2$$
$$= -\eta_t\left(1-\frac{\eta_t L_t}{2}\right)\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 + \eta_t\left(\eta_t L_t - 1\right)\left\langle\nabla f_t(\boldsymbol{x}_t)\,,\,\boldsymbol{\xi}_t\right\rangle + \frac{\eta_t^2 L_t}{2}\left\|\boldsymbol{\xi}_t\right\|^2\ .$$

We obtain the inequality (5) by rearranging the terms. □

Proof of Theorem 5.2. We prove by induction. The base case $t = T+1$ trivially holds. Consider $1 \le t \le T$, we have

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

From the induction hypothesis we have

$$\mathbb{E}\left[\exp\left(S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \le \exp\left(3\sigma^2\sum_{i=t+1}^{T}\frac{w_i\eta_i^2 L_i}{2}\right),$$

hence

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

We have then

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

where the second line is due to (5) in Lemma 5.1 and the second to last line is due to the helper Lemma 2.2 in Liu et al. (2023) (see Lemma A.3 for a restatement). Therefore,

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

which is what we want to show. □

Proof of Corollary 5.2.1. In Theorem 5.2, let $t = 1$ we obtain

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

Hence, by Markov's inequality, we have

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

In other words, with probability at least $1-\delta$ (once condition (6) is satisfied),

$$\sum_{t=1}^{T}\left(\left(w_t\eta_t\left(1-\frac{\eta_t L_t}{2}\right) - v_t\right)\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 + w_t\left(\Delta_{t+1} f_t - \Delta_t f_t\right)\right) \leq 3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L_t}{2} + \log\frac{1}{\delta} .$$

□

Proof of Lemma 5.3. Consider $w_t = w = 1/(6\sigma^2\eta_1)$. Clearly, our choice of w_t satisfies condition (6) since

$$w_t\eta_t^2 L_t = \frac{\eta_t}{\eta_1}\cdot(\eta_t L_t)\cdot\frac{1}{6\sigma^2} \leq \frac{1}{2\sigma^2},$$

then it follows that

$$\text{LHS of (7)} \geq \underbrace{\sum_{t=1}^{T}\left(w\eta_t\left(1-\frac{\eta_t L_t}{2}\right) - 3\sigma^2 w^2\eta_t^2(\eta_t L_t - 1)^2\right)\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2}_{A} + \underbrace{\sum_{t=1}^{T} w\left(\Delta_{t+1} f_t - \Delta_t f_t\right)}_{B},$$

where

$$\begin{aligned}
A &= \sum_{t=1}^{T} w\eta_t \left(1 - \frac{\eta_t L_t}{2} - 3\sigma^2 w\eta_t \left(1-\eta_t L_t\right)^2\right) \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \\
&\geq \sum_{t=1}^{T} w\eta_t \left(1 - \frac{\eta_t L_t}{2} - 3\sigma^2 w\eta_1 \left(1-\eta_t L_t\right)^2\right) \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \\
&= \sum_{t=1}^{T} w\eta_t \left(1 - \frac{\eta_t L_t}{2} - \frac{1}{2}\left(1-\eta_t L_t\right)^2\right) \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \\
&\geq \sum_{t=1}^{T} \frac{w\eta_t}{2} \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \\
&= \frac{w}{2} \sum_{t=1}^{T} \eta_t \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \ .
\end{aligned}$$

The second inequality is due to

$$1 - \frac{\eta L_t}{2\sqrt{t}} - \frac{1}{2}\left(1-\eta_t L_t\right)^2 \geq \frac{1}{2},$$

when $0 \leq \eta_t L_t \leq 1$. We also have

$$\begin{aligned}
B &= w \sum_{t=1}^{T} \left(f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}_t)\right) \\
&= w \sum_{t=1}^{T} \left(f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t) + f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right) \\
&= w \sum_{t=1}^{T} \left(f_t(\boldsymbol{x}_{t+1}) - f_{t-1}(\boldsymbol{x}_t)\right) + w \sum_{t=1}^{T} \left(f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right) \\
&= w\left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1\right) + w \sum_{t=1}^{T} \left(f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right) \ .
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 + 2\left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1\right) + 2\sum_{t=1}^{T} \left(f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t)\right) &\leq 3\sigma^2 \sum_{t=1}^{T} \eta_t^2 L_t + \frac{2}{w} \log\frac{1}{\delta} \\
&= 3\sigma^2 \sum_{t=1}^{T} \eta_t^2 L_t + 12\sigma^2 \eta_1 \log\frac{1}{\delta} \ .
\end{aligned}$$

□

Proof of Theorem 5.4. We first take care of the LHS of (8). Note that by definition,

$$\nabla f_t(\boldsymbol{x}_t) = \nabla f(\boldsymbol{x}_t) + \gamma_t \boldsymbol{x}_t \ .$$

Invoking Lemma A.1 with $u = \nabla f(\boldsymbol{x}_t)$, $v = \gamma_t \boldsymbol{x}_t$ and $\lambda = 1/2$ gives

$$\left\|\nabla f_t(\boldsymbol{x}_t)\right\|^2 \geq \frac{1}{2}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 - \gamma_t^2 \left\|\boldsymbol{x}_t\right\|^2 \ . \tag{13}$$

Also recall that

$$\begin{aligned}
f_{t-1}(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t) &= f(\boldsymbol{x}_t) + \frac{\gamma_{t-1}}{2}\left\|\boldsymbol{x}_t\right\|^2 - \left(f(\boldsymbol{x}_t) + \frac{\gamma_t}{2}\left\|\boldsymbol{x}_t\right\|^2\right) \\
&= \frac{\gamma_{t-1} - \gamma_t}{2}\left\|\boldsymbol{x}_t\right\|^2 \ .
\end{aligned} \tag{14}$$

Obviously our choice of η_t is non-increasing and

$$\begin{aligned}\eta_t L_t &= \eta_t \left(L + \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \\ &= \eta_t L + 1 - \frac{\eta_t}{\eta_{t-1}} \\ &= \frac{L}{L+\sigma\sqrt{t}} - \frac{L+\sigma\sqrt{t-1}}{L+\sigma\sqrt{t}} + 1 \\ &\leq 1 \ .\end{aligned}$$

Now, plugging the above couple of inequalities (13) and (14) in, we obtain

$$\begin{aligned}\text{LHS of (8)} &\geq \sum_{t=1}^{T} \eta_t \left(\frac{1}{2} \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 - \gamma_t^2 \left\| \boldsymbol{x}_t \right\|^2 \right) + 2 \left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1 \right) + \sum_{t=1}^{T} \left(\gamma_{t-1} - \gamma_t \right) \left\| \boldsymbol{x}_t \right\|^2 \\ &= \frac{1}{2} \sum_{t=1}^{T} \eta_t \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 + 2 \left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1 \right) + \sum_{t=1}^{T} \left(\gamma_{t-1} - \gamma_t - \gamma_t^2 \eta_t \right) \left\| \boldsymbol{x}_t \right\|^2 \\ &\overset{(*)}{\geq} \frac{1}{2} \sum_{t=1}^{T} \eta_t \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 + 2 \left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1 \right) \\ &\geq \frac{1}{2} \sum_{t=1}^{T} \eta_T \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 + 2 \left(\Delta_{T+1} f_{T+1} - \Delta_1 f_1 \right) \\ &\geq \frac{\eta_T}{2} \sum_{t=1}^{T} \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 - 2\Delta_1 f_1,\end{aligned}$$

where $(*)$ comes from Lemma A.4, and the last inequality follows from the fact that $f_{T+1}(\boldsymbol{x}_{T+1}) \geq f^*$. Besides, by choosing $\eta_t = 1/(L+\sigma\sqrt{t})$,

$$\begin{aligned}\text{RHS of (8)} &= 3\sigma^2 \sum_{t=1}^{T} \left(\frac{1}{L+\sigma\sqrt{t}} \right)^2 \left(L + \sigma\sqrt{t} - \left(L + \sigma\sqrt{t-1} \right) \right) + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \\ &= 3\sigma^3 \sum_{t=1}^{T} \left(\frac{1}{L+\sigma\sqrt{t}} \right)^2 \left(\sqrt{t} - \sqrt{t-1} \right) + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \\ &\leq 3\sigma^3 \sum_{t=1}^{T} \left(\frac{1}{L+\sigma\sqrt{t}} \right)^2 + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \\ &\leq 3\sigma^3 \cdot \frac{2}{\sigma^2} \left(\log \left(1 + \frac{\sigma\sqrt{T}}{L} \right) + \frac{L}{L+\sigma\sqrt{T}} - 1 \right) + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \\ &\leq 6\sigma \log \left(1 + \frac{\sigma\sqrt{T}}{L} \right) + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \ .\end{aligned}$$

Now, dividing both sides by $\eta_T/2$ yields

$$\begin{aligned}\sum_{t=1}^{T} \left\| \nabla f(\boldsymbol{x}_t) \right\|^2 &\leq \frac{2}{\eta_T} \left(2 \left(f(\boldsymbol{0}) - f^* \right) + 6\sigma \log \left(1 + \frac{\sigma\sqrt{T}}{L} \right) + 12\sigma^2 \eta_1 \log \frac{1}{\delta} \right) \\ &= \left(4 \left(f(\boldsymbol{0}) - f^* \right) + 12\sigma \log \left(1 + \frac{\sigma\sqrt{T}}{L} \right) + \frac{24\sigma^2}{L+\sigma} \log \frac{1}{\delta} \right) \left(L + \sigma\sqrt{T} \right) \ .\end{aligned}$$

Factoring out an order of T gives the desired result. □

A.3 Proof of Theorem 6.1

Theorem (Restatement of Theorem 6.1)**.** *Assume* (A1)–(A4) *are satisfied. Let* $\boldsymbol{x}_1 = \mathbf{0}$*. Consider SDA iterates using step sizes*

$$\eta_t = \frac{\eta}{\sqrt{\gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2}} \ .$$

Let $\boldsymbol{x}^*$ *be an arbitrary stationary point and*

$$B_T := \max_{t \le T} \left\{ \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \right\} \ .$$

Then, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \le \frac{2\left(f(\mathbf{0}) - f^*\right)}{\eta} + 2\sqrt{2}L^2\left(\sqrt{2} + \rho\right)\mathbb{E}\left[\left(\frac{13B_T}{2\eta} + 2\eta\right)^2\right] + \frac{\gamma}{2} + 2\sqrt{2}L\sigma\left(\mathbb{E}\left[\left(\frac{13B_T}{2\eta} + 2\eta\right)^2\right]\right)^{1/2}\sqrt{T} \ .$$

Proof. We consider the update rule given by

$$\boldsymbol{x}_{t+1} = -\eta_t \sum_{i=1}^{t} \boldsymbol{g}_i = -\eta_t \boldsymbol{\theta}_t,$$

where $\boldsymbol{\theta}_t = \sum_{i=1}^{t} \boldsymbol{g}_i$, and η_t is a learning rate.

$$\begin{aligned} \boldsymbol{x}_{t+1} - \boldsymbol{x}_t &= -\eta_t \boldsymbol{\theta}_t + \eta_{t-1}\boldsymbol{\theta}_{t-1} \\ &= (\eta_{t-1} - \eta_t)\boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{g}_t \\ &= -\eta_t\left(\boldsymbol{g}_t + \left(1 - \frac{\eta_{t-1}}{\eta_t}\right)\boldsymbol{\theta}_{t-1}\right) \ . \end{aligned}$$

Denote

$$\boldsymbol{g}'_t := \boldsymbol{g}_t + \left(1 - \frac{\eta_{t-1}}{\eta_t}\right)\boldsymbol{\theta}_{t-1},$$

the updates can be written us $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}'_t$. We consider the learning rates given by

$$\eta_t = \frac{\eta}{\sqrt{\gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2}},$$

We define $S_t := \gamma + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2$. We note by $\boldsymbol{\xi}_t$ the noise

$$\boldsymbol{\xi}_t := \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t) \ .$$

Using smoothness, we have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) &\le \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle + \frac{1}{2}L\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \\ &= -\eta_t \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{g}'_t\rangle + \frac{L}{2}\eta_t^2 \|\boldsymbol{g}'_t\|^2 \ . \end{aligned}$$

Let $\Delta_t := f(\boldsymbol{x}_t) - f^*$. Using the last bound gives

$$\langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{g}'_t\rangle \le \frac{1}{\eta_t}\left(\Delta_t - \Delta_{t+1}\right) + \frac{L}{2}\eta_t \|\boldsymbol{g}'_t\|^2 \ .$$

Summing over t and taking $\eta_0 = \eta$, we have

$$\begin{aligned}\sum_{t=1}^{T} \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{g}'_t \rangle &\leq \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\Delta_t - \Delta_{t+1}\right) + \frac{L}{2} \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}'_t\right\|^2 \\ &\leq \frac{\Delta_1}{\eta} - \frac{\Delta_{T+1}}{\eta_T} + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \Delta_t + \frac{L}{2} \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}'_t\right\|^2 \\ &\leq \frac{\Delta_1}{\eta} + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \Delta_t + \frac{L}{2} \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}'_t\right\|^2 \ .\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\sum_{t=1}^{T} \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{g}_t \rangle &\leq \frac{\Delta_1}{\eta} + \sum_{t=1}^{T} \left(\frac{\eta_{t-1}}{\eta_t} - 1\right) \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{\theta}_{t-1} \rangle + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \Delta_t + \frac{L}{2} \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}'_t\right\|^2 \\ &\leq \frac{\Delta_1}{\eta} - \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_t \rangle + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \Delta_t \\ &\quad + L \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}_t\right\|^2 + L \sum_{t=1}^{T} \eta_t \left(\frac{\eta_{t-1}}{\eta_t} - 1\right)^2 \left\|\boldsymbol{\theta}_{t-1}\right\|^2 \\ &= \frac{\Delta_1}{\eta} - \sum_{t=1}^{T} \left(\langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_t \rangle - \Delta_t\right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + L \sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}_t\right\|^2 \\ &\quad + L \sum_{t=1}^{T} \eta_t \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)^2 \left\|\boldsymbol{x}_t\right\|^2 \\ &= \frac{\Delta_1}{\eta} - \sum_{t=1}^{T} \left(\langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_t \rangle - \Delta_t\right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + L\eta \sum_{t=1}^{T} \frac{\left\|\boldsymbol{g}_t\right\|^2}{\sqrt{S_t}} \\ &\quad + L \sum_{t=1}^{T} \left(1 - \frac{\eta_t}{\eta_{t-1}}\right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \left\|\boldsymbol{x}_t\right\|^2 \\ &\leq \frac{\Delta_1}{\eta} - \sum_{t=1}^{T} \left(\langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_t \rangle - \Delta_t\right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + L\eta \sum_{t=1}^{T} \frac{\left\|\boldsymbol{g}_t\right\|^2}{\sqrt{S_t}} \\ &\quad + L \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \left\|\boldsymbol{x}_t\right\|^2 \\ &\leq \frac{\Delta_1}{\eta} - \sum_{t=1}^{T} \left(\langle \nabla f(\boldsymbol{x}_t)\,,\, \boldsymbol{x}_t \rangle - \Delta_t - L \left\|\boldsymbol{x}_t\right\|^2\right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + L\eta \sum_{t=1}^{T} \frac{\left\|\boldsymbol{g}_t\right\|^2}{\sqrt{S_t}} \ . \end{aligned} \tag{15}$$

We also have

$$\sum_{t=1}^{T} \frac{\left\|\boldsymbol{g}_t\right\|^2}{\sqrt{S_t}} \leq 2 \sum_{t=1}^{T} \left(\sqrt{S_t} - \sqrt{S_{t-1}}\right) \leq 2\sqrt{S_T} \ .$$

Let $\boldsymbol{x}^*$ be an arbitrary stationary point, given that $f(\cdot)$ is L-smooth, we have

$$\left|f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) - \langle \nabla f(\boldsymbol{x}^*)\,,\, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle\right| \leq \frac{L}{2} \left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|^2 \ .$$

Therefore,

$$\left|\Delta_t\right| = \left|f(\boldsymbol{x}_t) - f^*\right| \leq \left|\langle \nabla f(\boldsymbol{x}^*)\,,\, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle\right| + \frac{L}{2} \left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|^2 = \frac{L}{2} \left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|^2 \ .$$

Moreover,

$$\left\|\nabla f(\boldsymbol{x}_t)\right\| = \left\|\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}^*)\right\| \leq L \left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\| \ .$$

Using the last two bounds we have

$$\begin{aligned}
|\langle \nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{x}_t\rangle - \Delta_t - L\left\|\boldsymbol{x}_t\right\|^2| &\le |\langle \nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{x}_t\rangle| + |\Delta_t| + L\left\|\boldsymbol{x}_t\right\|^2 \\
&\le \left\|\nabla f(\boldsymbol{x}_t)\right\|\left\|\boldsymbol{x}_t\right\| + \frac{L}{2}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 + 2L\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 + 2L\left\|\boldsymbol{x}^*\right\|^2 \\
&\le L\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|\left\|\boldsymbol{x}_t\right\| + \frac{5L}{2}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 + 2L\left\|\boldsymbol{x}^*\right\|^2 \\
&\le L\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|\left\|\boldsymbol{x}^*\right\| + \frac{7L}{2}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 + 2L\left\|\boldsymbol{x}^*\right\|^2 \\
&\le 4L\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 + \frac{5L}{2}\left\|\boldsymbol{x}^*\right\|^2 \\
&\le \frac{13L}{2}B_T, \qquad (16)
\end{aligned}$$

where we used the definition of B_T in the last line.

Using bounds (16) in (15) and the definition of B_T, we have

$$\begin{aligned}
\sum_{t=1}^{T}\langle \nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{g}_t\rangle &\le \frac{\Delta_1}{\eta} + \frac{13L}{2}B_T\sum_{t=1}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right) + 2L\eta\sqrt{S_T} \\
&\le \frac{\Delta_1}{\eta} + \frac{13}{2\eta}LB_T\sqrt{S_T} + 2L\eta\sqrt{S_T} \\
&\le \frac{\Delta_1}{\eta} + 13\sqrt{2}\frac{L}{2\eta}B_T\sqrt{\gamma + \sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + \sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2} \\
&\quad + 2\sqrt{2}L\eta\sqrt{\gamma + \sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + \sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2} \\
&\le \frac{\Delta_1}{\eta} + \sqrt{2}\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\gamma + \sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2} \\
&\quad + \sqrt{2}\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2}\,.
\end{aligned}$$

Taking the expectation both sides, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t)\,,\,\boldsymbol{g}_t\rangle\right] \\
&\le \frac{\Delta_1}{\eta} + \sqrt{2}\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\gamma + \sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}\right] \\
&\quad + \sqrt{2}\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{\xi}_t\right\|^2}\right]\,. \qquad (17)
\end{aligned}$$

Recall that using the fact that $ax - (1/4)x^2 \le a^2$ for all $x \in \mathbb{R}$ we have

$$\sqrt{2}\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\gamma + \sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2} - \frac{1}{4}\sum_{t=1}^{T}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 - \frac{\gamma}{4} \le 2\left(\frac{13L}{2\eta}B_T + 2L\eta\right)^2\,. \qquad (18)$$

Moreover, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2}\right] \\
&\leq \left(\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2\right]\right)^{1/2} \\
&\leq \left(\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\sum_{t=1}^{T}\left(\rho\|\nabla f(\boldsymbol{x}_t)\|^2 + \sigma^2\right)\right]\right)^{1/2} .
\end{aligned}$$

To ease the notation, let

$$\bar{B}_T := \left(\mathbb{E}\left[\left(\frac{13}{2\eta}B_T + 2\eta\right)^2\right]\right)^{1/2} .$$

So, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2}\right] &\leq L\bar{B}_T\sqrt{\sigma^2 T + \rho\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]} \\
&\leq L\sigma\bar{B}_T\sqrt{T} + L\bar{B}_T\sqrt{\rho}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]} .
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)\sqrt{\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2}\right] - \frac{1}{4}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&\leq L\sigma\bar{B}_T\sqrt{T} + L\bar{B}_T\sqrt{\rho}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]} - \frac{1}{4}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&\leq L\sigma\bar{B}_T\sqrt{T} + L^2\bar{B}_T^2\rho . \qquad (19)
\end{aligned}$$

Finally, using (18) and (19) in (17), we have

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] &\leq \frac{\gamma}{4} + \frac{\Delta_1}{\eta} + 2\mathbb{E}\left[\left(\frac{13L}{2\eta}B_T + 2L\eta\right)^2\right] + \sqrt{2}L\sigma\bar{B}_T\sqrt{T} + \sqrt{2}L^2\rho\bar{B}_T^2 \\
&= \frac{\gamma}{4} + \frac{\Delta_1}{\eta} + \sqrt{2}L\sigma\bar{B}_T\sqrt{T} + \sqrt{2}L^2\left(\sqrt{2}+\rho\right)\bar{B}_T^2 .
\end{aligned}$$

□

A.4 Additional Helper Lemmas

Lemma A.1. *Let $u, v \in \mathbb{R}^d$ and let $\lambda > 0$. Then*

$$\|u+v\|^2 \geq (1-\lambda)\|u\|^2 + (1-\lambda^{-1})\|v\|^2 .$$

Proof. Young's inequality says that for any $\lambda > 0$,

$$2\langle u, v\rangle \geq -\lambda^{-1}\|u\|^2 - \lambda\|v\|^2 .$$

Insert this bound into the expansion

$$\|u+v\|^2 = \|u\|^2 + 2\langle u, v\rangle + \|v\|^2 ,$$

yielding

$$\|u+v\|^2 \geq \|u\|^2 - \lambda^{-1}\|u\|^2 - \lambda\|v\|^2 + \|v\|^2 = (1-\lambda)\|u\|^2 + (1-\lambda^{-1})\|v\|^2 \ .$$

□

Lemma A.2. *Consider the step sizes*

$$\eta_t = \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})},$$

where $\alpha \in (0,1]$*, let* $\gamma_t = 1/\eta_t - 1/\eta_{t-1}$*, and* $L_t = L + \gamma_t$*. We have for any* $t \geq 2$

$$\frac{\eta_t L_t}{2} \leq 1 \tag{20}$$

$$\eta_t^2 L_t \leq \frac{3}{2L(1+\rho)(1+\rho+\alpha\sqrt{t})^2} \tag{21}$$

$$\frac{1}{3}\eta_t - \frac{1}{6}\eta_t^2 L_t - \frac{1}{2}\rho\eta_t^2 L_t \geq \frac{\eta_t}{9} \tag{22}$$

$$\frac{\gamma_{t-1}-\gamma_t}{2} - \frac{1}{2}\gamma_t^2\eta_t + \frac{1}{4}\gamma_t^2\eta_t^2 L_t \geq 0 \ . \tag{23}$$

Proof. Fix $t \geq 2$. Before proving the results of the lemma, we use the following bounds on γ_t and $\gamma_{t-1}-\gamma_t$. We have

$$\begin{aligned}
\gamma_t &= \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \\
&= L(1+\rho)\left(1+\rho+\alpha\sqrt{t}\right) - L(1+\rho)\left(1+\rho+\alpha\sqrt{t-1}\right) \\
&= L(1+\rho)\alpha\left(\sqrt{t}-\sqrt{t-1}\right) \\
&= \frac{L(1+\rho)\alpha}{\sqrt{t}+\sqrt{t-1}} \leq \frac{L(1+\rho)\alpha}{2\sqrt{t-1}} \ .
\end{aligned} \tag{24}$$

We also have

$$\begin{aligned}
\gamma_t^2\eta_t &\leq \frac{L^2(1+\rho)^2\alpha^2}{4(t-1)} \cdot \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})} \\
&\leq \frac{L(1+\rho)\alpha^2}{4(t-1)(1+\rho+\alpha\sqrt{t})} \\
&\leq \frac{L(1+\rho)\alpha}{4(t-1)\sqrt{t}} \ .
\end{aligned} \tag{25}$$

Moreover, using (24), we have

$$\begin{aligned}
\gamma_{t-1}-\gamma_t &= \frac{L(1+\rho)\alpha}{(\sqrt{t-1}+\sqrt{t-2})} - \frac{L(1+\rho)\alpha}{(\sqrt{t}+\sqrt{t-1})} \\
&= L(1+\rho)\alpha\left(\frac{1}{\sqrt{t-1}+\sqrt{t-2}} - \frac{1}{\sqrt{t}+\sqrt{t-1}}\right) \ .
\end{aligned}$$

Recall that

$$\begin{aligned}
\frac{1}{\sqrt{t-1}+\sqrt{t-2}} - \frac{1}{\sqrt{t}+\sqrt{t-1}} &= \frac{\sqrt{t}-\sqrt{t-2}}{(\sqrt{t-1}+\sqrt{t-2})(\sqrt{t}+\sqrt{t-1})} \\
&= \frac{2}{(\sqrt{t}+\sqrt{t-2})(\sqrt{t-1}+\sqrt{t-2})(\sqrt{t}+\sqrt{t-1})} \\
&\geq \frac{1}{4\sqrt{t}(t-1)} \ .
\end{aligned}$$

We conclude that

$$\gamma_{t-1} - \gamma_t \geq \frac{L(1+\rho)\alpha}{4\sqrt{t}(t-1)} \ . \tag{26}$$

Proof of bound (20)**:** We have

$$\begin{aligned}
\eta_t L_t &\leq \frac{1}{L(1+\rho)(1+\rho+\alpha\sqrt{t})}\left(L + \frac{L(1+\rho)\alpha}{\sqrt{t}}\right) \\
&\leq \frac{1}{1+\rho+\alpha\sqrt{t}}\left(\frac{1}{1+\rho} + \frac{\alpha}{\sqrt{t}}\right) \\
&= \frac{1/(1+\rho)}{1+\rho+\alpha\sqrt{t}} + \frac{\alpha}{\sqrt{t}(1+\rho+\alpha\sqrt{t})} \\
&\leq \frac{1}{(1+\rho)^2} + \frac{\alpha}{\sqrt{t}(1+\rho)+\alpha t} \\
&\leq \frac{1}{(1+\rho)^2} + \frac{1}{\sqrt{t}(1+\rho)+t} \leq 2,
\end{aligned}$$

where we used $\alpha \in [0,1]$ and the fact that

$$x \to \frac{x}{\sqrt{t}(1+\rho)+xt}$$

is increasing for positive numbers.

Proof of bound (21)**:** We have

$$\begin{aligned}
\eta_t^2 L_t &= \frac{1}{L^2(1+\rho)^2(1+\rho+\alpha\sqrt{t})^2}\left(L + \frac{L(1+\rho)\alpha}{2\sqrt{t-1}}\right) \\
&\leq \frac{1+\dfrac{1+\rho}{2}}{L(1+\rho)^2(1+\rho+\alpha\sqrt{t})^2} \\
&\leq \frac{3}{2L(1+\rho)(1+\rho+\alpha\sqrt{t})^2} \ .
\end{aligned}$$

Proof of bound (22)**:** The bound is equivalent to

$$\eta_t L_t \leq \frac{4}{3+9\rho} \ . \tag{27}$$

Using the expressions of η_t and L_t, we have

$$\eta_t L_t = \frac{1}{(1+\rho)(1+\rho+\alpha\sqrt{t})} + \frac{\alpha(\sqrt{t}-\sqrt{t-1})}{1+\rho+\alpha\sqrt{t}} := h(\alpha) \ .$$

Studying the last expression's dependence on α we have

$$\begin{aligned}
h(\alpha) &\leq \max\left\{\frac{1}{(1+\rho)^2}; \frac{1}{(1+\rho)(1+\rho+\sqrt{t})} + \frac{1}{(1+\rho+\sqrt{t})(\sqrt{t}+\sqrt{t-1})}\right\} \\
&\leq \max\left\{\frac{1}{(1+\rho)^2}; \frac{1}{(1+\rho)(1+\sqrt{2}+\rho)} + \frac{1}{(1+\sqrt{2}+\rho)(1+\sqrt{2})}\right\} \ .
\end{aligned}$$

Therefore, we need to show that each of the terms inside the maximum is smaller than $4/(3+9\rho)$.

The function

$$x \to \frac{3+9x}{4(1+x)^2}$$

is upper bounded by 1 for positive inputs. Therefore,

$$\frac{1}{(1+\rho)^2} \leq \frac{4}{3+9\rho} \ .$$

The function

$$x \to \frac{3+9x}{4(1+x)(1+\sqrt{2}+x)} + \frac{3+9x}{4(1+\sqrt{2}+x)(1+\sqrt{2})}$$

is upper bounded by 1 for positive inputs. Therefore,

$$\frac{1}{(1+\rho)(1+\sqrt{2}+\rho)} + \frac{1}{(1+\sqrt{2}+\rho)(1+\sqrt{2})} \leq \frac{4}{3+9\rho} \ .$$

We conclude that the bound (27) holds for each $\rho \geq 0$, which gives the result.

Proof of bound (23)**:** Combining the bound on $\gamma_{t-1} - \gamma_t$ given by (26), with the bound on $\gamma_t^2 \eta_t$ given in (25) yields

$$\begin{aligned}\frac{\gamma_{t-1}-\gamma_t}{2} - \frac{1}{2}\gamma_t^2\eta_t + \frac{1}{4}\gamma_t^2\eta_t^2 L_t &\geq \frac{\gamma_{t-1}-\gamma_t}{2} - \frac{1}{2}\gamma_t^2\eta_t \\ &\geq \frac{L(1+\rho)\alpha}{8\sqrt{t}(t-1)} - \frac{L(1+\rho)\alpha}{8\sqrt{t}(t-1)} = 0,\end{aligned}$$

where we used in the last line bounds (26) and (25). □

Lemma A.3 (Restatement of Lemma 2.2 in Liu et al. (2023))**.** *Suppose $X \in \mathbb{R}^d$ such that $\mathbb{E}[X] = 0$ and $\|X\|$ is a σ-sub-Gaussian random variable, then for any $a \in \mathbb{R}^d$, $0 \leq b \leq 1/(2\sigma)$,*

$$\mathbb{E}\left[\exp\left(\langle a\,,\, X\rangle + b^2\|X\|^2\right)\right] \leq \exp\left(3\left(\|a\|^2 + b^2\right)\sigma^2\right) \ .$$

Especially, when $b = 0$, we have

$$\mathbb{E}\left[\exp\left(\langle a\,,\, X\rangle\right)\right] \leq \exp\left(2\|a\|^2\sigma^2\right) \ .$$

Proof. The proof is very long and technical, we refer the reader to Liu et al. (2023, p. 14–15) for details. □

Lemma A.4. *Let*

$$\eta_t = \frac{\eta}{1+\alpha\sqrt{t}}, \quad \forall t \geq 1,$$

with $\eta > 0$ and $\alpha \geq 0$. Then, we have

$$\gamma_{t-1} - \gamma_t - \gamma_t^2\eta_t \geq 0, \quad \forall t \geq 1 \ . \tag{28}$$

Proof. Given the choice of η_t, we have access to a closed form expression for γ_t in t. Let's find it.

$$\gamma_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{\alpha}{\eta}\left(\sqrt{t} - \sqrt{t-1}\right) \ .$$

We will first examine the base cases. For $t = 1$, we need to show that

$$\gamma_0 - \gamma_1 - \gamma_1^2\eta_1 = \gamma_0 - \frac{\alpha}{\eta} - \frac{\alpha^2}{(1+\alpha)\eta} \geq 0 \ .$$

Since we have control over γ_0, we can always set it to be sufficiently large. For example, setting

$$\gamma_0 = \frac{\alpha}{\eta} + \frac{\alpha^2}{(1+\alpha)\eta} + 1$$

will do the trick. For $t \geq 2$,

$$\text{LHS of (28)} = \alpha \underbrace{\left(2\sqrt{t-1} - \sqrt{t} - \sqrt{t-2}\right)}_{A_t} + \alpha^2\sqrt{t}\underbrace{\left(4\sqrt{(t-1)} - \sqrt{(t-2)} - 3\sqrt{t} + \frac{1}{\sqrt{t}}\right)}_{B_t} \ .$$

Here we intentionally omitted the boring algebra. Notice A_t is non-negative for concavity of $\sqrt{x}$:

$$\frac{\sqrt{t}}{2} + \frac{\sqrt{t-2}}{2} \geq \sqrt{\frac{t}{2} + \frac{t-2}{2}} = \sqrt{t-1}$$
$$\implies A_t = \frac{\sqrt{t}}{2} + \frac{\sqrt{t-2}}{2} - \sqrt{t-1} \geq 0 \ .$$

To see that B_t is non-negative, let

$$\psi(\tau) := 4\sqrt{\tau-1} - \sqrt{\tau-2} - 3\sqrt{\tau} + \frac{1}{\sqrt{\tau}}, \quad \forall \tau \in (2, \infty) \ .$$

Now,

$$\psi(2) = 4 - 3\sqrt{2} + \frac{1}{\sqrt{2}} \approx 0.464 > 0,$$
$$\psi(\infty) = 0 \ .$$

To show that B_t is non-negative, it suffices to show that $\psi'(\tau)$ is non-negative for $\tau \in (2, \infty)$. Differentiating $\psi(\tau)$ gives

$$\psi'(\tau) = \frac{2}{\sqrt{\tau-1}} - \frac{1}{2\sqrt{\tau-2}} - \frac{3}{2\sqrt{\tau}} - \frac{1}{2\tau^{3/2}} \leq 0 \ .$$

Thus, ψ decreases monotonically from the positive value $\psi(2)$ down to 0, so $\psi(\tau) \geq 0$ for all $\tau \geq 2$. In particular, $B_t = \psi(t) \geq 0$ for every integer $t \geq 2$. □