ALMOST SURE CONVERGENCE RATE FOR THE LAST ITERATE OF STOCHASTIC GRADIENT DESCENT SCHEMES

MARCEL HUDIANI

Abstract. We study the almost sure convergence rate for the last iterate of stochastic gradient descent (SGD) and stochastic heavy ball (SHB) in the parametric setting when the objective function F is globally convex or non-convex whose gradient is γ-Hölder. Given the step size $O(t^{-p})$ with $p \in (\frac{1}{1+\gamma}, 1)$, we achieve $\|\nabla F(w_t)\|^2 = o(t^{p-1})$ and if F is convex, $F(w_t) - F_* = o(t^{p-1})$ using only Gronwall's inequality and classical probability theory without relying on Robbins-Siegmund theorem nor any martingale convergence theory. In addition, we proved that SHB attains a convergence rate of $O(t^{p-1}\log^2 \frac{t}{\delta})$ with probability at least $1-\delta$ when the objective function is convex and $\gamma = 1$.

1. Introduction

We study the almost sure convergence rate for the last iterate of Stochastic Gradient Descent (SGD) and Stochastic Heavy Ball (SHB) schemes associated to solving an unconstrained optimization problem involving a cost function defined on a subset of $\mathbb{R}^d$. We consider the application of these algorithms in the following context:

$$w_{t+1} - w_t = -\alpha_t \nabla \ell(Z_t, w_t) + \beta(w_t - w_{t-1}) \ , \ \beta \in [0,1) \tag{1.1}$$

where w_t is in $\mathbb{R}^d$, $Z_t \in \mathbb{R}^N$ is an i.i.d process with finite variance sampled from a continuous distribution with density ρ independent of $\mathscr{F}_t = \sigma(w_s : s \le t)$, α_t is the (non-random) step size, and ℓ is an estimator of a deterministic cost function $F : \mathcal{W} \to [0, \infty)$ defined by $F(w) = \mathbb{E}_\rho[\ell(Z, w)]$. When $\beta = 0$, eq. (1.1) is the SGD and can be rewritten as

$$w_{t+1} - w_t = -\alpha_t \nabla F(w_t) + \alpha_t\, \delta m_t \ , \ \delta m_t = \nabla F(w_t) - \nabla \ell(Z_t, w_t). \tag{1.2}$$

Algorithms of the form (1.2) have been studied extensively as part of a field called stochastic approximation ([7], [10]). However, the almost sure convergence rate for the last iterate of SGD is studied only recently. The most recent work in the parametric setting are done by Weissmann et al in [12]. Weissmann et al. considers SGD and SHB where properties of the objective function F is only known locally, i.e. properties apply in some neighborhood of a point [12, assumption 2.1.(ii)]. Under the local gradient domination [12, definition 2.2] $\|\nabla F(w_t)\|^2 \ge F(w_t) - F_*$, Weissmann et al proved the existence of neighborhoods $\mathcal{U}_1, \mathcal{U}$ of critical points in the parameter space $\mathcal{W}$ for which if $w_1 \in \mathcal{U}_1$, then the event $\Omega_{\mathcal{U}} = \{\mathcal{U} \text{ is absorbing for } w_t\}$ is highly probable and $(F(w_t) - F_*)\mathbb{1}_{\Omega_{\mathcal{U}}} = o(t^{-1+\epsilon})$ [12, table 1, theorem 5.1]. The set $\mathcal{U}$ is defined as a neighborhood of the set of minimizers with a fixed radius. We remark that Weissmann et al also examines the case when global properties of F are known [12, table 1].

When global properties of the objective function are known, there are other recent works as well. Liu and Yuan proved almost sure convergence rates for SGD, SHB, and SNAG (Stochastic Nesterov's Accelerated Gradient) by assuming that the objective function is either globally non-convex, convex, or strongly convex [6]. Liu and Yuan produced the rate $F(w_t) - F_* = o(t^{-1+\epsilon})$ for strongly convex F and $\min_{1 \le i < t-1} \|\nabla F(w_i)\|^2 = o(\sum_{i=1}^{t-1} \alpha_i)^{-1}$ for non-convex F [6, theorem 6, theorem 8]. The strongly convex rate is optimal according to Agarwal et al [1]. For a convex objective function F, Liu and Yuan proved a rate of $F(w_t) - F_* = o(t^{-\frac{1}{3}+\epsilon})$, which is not as close to the optimal rate of $t^{-1/2}$ in expectation [6, remark 14]. Liu and Yuan used the Robbins-Siegmund theorem ([10, theorem 1], [6, proposition 2]) to prove their convergence rate results. We remark that both SGD and SHB produce the same rate according to Liu and Yuan, i.e. the momentum parameter β does not affect the convergence rate [6, theorem 8, theorem 13].

Intuitively, the momentum parameter β in eq. (1.1) serves to counter slowdown when the gradient is near zero. As such, it makes sense to adjust β depending on the gradient and step sizes. Sebbouh et al studied such a case in [11]. In particular, they proved that the last iterate of SHB can achieve $o(t^{-1+\epsilon})$ rate in the convex setting when the SHB parameter β is allowed to vary as a function of t in the 'overparametrized' case [11, table 1, corollary 14, definition 3]. On the other hand, in the usual parametrized case, given step size of order $O(t^{-\frac{1}{2}-\epsilon})$, Sebbouh et al achieved $o(t^{-\frac{1}{2}+\epsilon})$ rate [11, corollary 17]. Similar to Liu and Yuan's work [6], Sebbouh et al used the Robbins-Siegmund theorem [11, lemma 6] to prove these rates.

In another work [5], Lei, Shi, and Guo obtained the last iterate almost sure convergence rate of $o(t^{\max(p-1,1-p(\gamma+1))+\epsilon})$ given step size $O(t^{-p})$ with $p \in (\frac{1}{1+\gamma}, 1)$ for a convex objective function, whose gradient is γ-Hölder with SGD in the non-parametric setting [5, theorem 6]. Moreover, in [5, corollary 11], Lei, Shi, and Guo proved that SGD converges at rate $F(w_{t+1}) - F_* = O(t^{\max(p-1,1-p(\gamma+1))} \log^2(t/\delta))$ with probability at least $1-\delta$ if $p \in (\frac{1}{1+\gamma}, 1)$. In particular, when $p = 2/3$ and $\gamma = 1$, their work matches Liu and Yuan's with $O(t^{-\frac{1}{3}+\epsilon})$ [5, theorem 6, corollary 11]. Interestingly, Lei, Shi, and Guo did not use Robbins-Siegmund theorem. However, their approach is similar to that of Robbins and Siegmund, where a supermartingale is constructed and the convergence of the iterates is implied by that of the supermartingale [5, section 4.2]. We remark that other authors have worked on the convergence rate for SGDs.

1.1. Contribution.

As remarked in Liu and Yuan's paper [6, remark 14], the rate $O(t^{-\frac{1}{3}+\epsilon})$ for the convex case is not close to the optimal rate in expectation $O(t^{-\frac{1}{2}})$ proved by Agarwal et al in [1]. Motivated by this remark, we study the almost sure convergence rates for the last iterate of SGD and SHB where the objective function F is convex and non-convex with ∇F being (γ, L)-Hölder. With these conditions, our contribution is as follows:

(1) Instead of relying on Robbins-Siegmund, or any stochastic approximation theory, we use only Gronwall's inequality and classical non-martingale probability theory to prove all 'little-o' convergence rates while using similar assumptions as in Liu and Yuan [6] (the main difference being ∇F is only assumed to be γ-Hölder in our case). By not relying on martingale convergence theory, we provide further validation and different viewpoint in the theory of convergence rate for SGDs. Also, we decouple such results from martingale convergence theory and uncover behaviors of SGD and SHB in the non-convex setting, e.g. $\|\nabla F(w_t)\|$ is decreasing eventually (proposition 4.1), which allows us to remove the 'min' in $\min_{1\le i\le t-1} \|\nabla F(w_t)\|^2 = o(t^{p-1})$ for non-convex setting, thereby improving Liu and Yuan's result in the non-convex case [6, theorem 6, theorem 8]. In addition, we are able to achieve an error rate of $o(t^{p-1})$ in the convex setting by showing that $F(w_t) - F_*$ decreases eventually.

(2) We provide a high probability convergence rate for SHB iterates w_t in eq. (1.1) with a fixed, constant $\beta \in (0, 1)$. We found that SHB converges at rate $O(t^{p-1})$ when the gradient of the objective function ∇F is Lipschitz. Our result is similar to Lei, Shi, Guo under the same assumptions. In addition, our result is consistent with the findings of SHB convergence rate thus far in table 1. While our result is restricted to the case when β is fixed, such a result is still beneficial since it provides a baseline SHB performance which can be compared to other algorithms. To the author's knowledge, this is the first high probability convergence rate result for the SHB.

(3) We use a slightly weaker assumption for ∇F compared to Liu and Yuan [6, assumption 4] that it is only γ-Hölder with $\gamma \in (0, 1]$, following Lei, Shi, and Guo [5, assumption 1]. In this manner, we provide a comprehensive analysis for general 'smooth' objective functions.

Algo	Cost Function w/ (γ, L)-smoothness	Step Size Decay Rate p in $O(t^{-p})$	Convergence Rate for $F(w_t) - F_*$ (convex case) or $\|\nabla F(w_t)\|^2$ (non-convex)	Statement
SGD	Convex, $\gamma \in (0, 1]$	$(\frac{1}{1+\gamma}, 1)$	$o(t^{p-1})$	theorem 2.5.
			$o(t^{\max(p-1, 1-(\gamma+1)p)})$	[5, theorem 6].
	Convex, $\gamma = 1$	$(\frac{2}{3}, 1)$	$o(t^{p-1})$	[6, theorem 13].
	Non-convex, $\gamma \in (0, 1]$	$(\frac{1}{1+\gamma}, 1)$	$\|\nabla F(w_t)\|^2 = o(t^{p-1})$	corollary 3.2.
	Non-convex, $\gamma = 1$	$(\frac{1}{2}, 1)$	$\min\limits_{1 \le i \le t-1} \|\nabla F(w_t)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right)$	[6, theorem 6].
SHB	Convex, $\gamma \in (0, 1]$	$(\frac{1}{1+\gamma}, 1)$	$o(t^{p-1})$	theorem 2.5.
	Convex, $\gamma = 1$	$(\frac{2}{3}, 1)$	$o(t^{p-1})$	[6, theorem 13].
		$(\frac{1}{2}, 1)$	$O(t^{p-1} \log^2 \frac{t}{\delta})$	theorem 2.6.
		$-\frac{1}{2} - \epsilon$	$o(t^{-\frac{1}{2}+\epsilon})$	[11, corollary 17].
	Non-convex, $\gamma \in (0, 1]$	$(\frac{1}{1+\gamma}, 1)$	$\|\nabla F(w_t)\|^2 = o(t^{p-1})$	corollary 4.2.
	Non-convex, $\gamma = 1$	$(\frac{1}{2}, 1)$	$\min\limits_{1 \le i \le t-1} \|\nabla F(w_t)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right)$	[6, theorem 8].

Table 1. Summary of Results.

1.2. **Acknowledgment.** I would like to thank my advisor, Sunder Sethuraman, for discussions and support throughout the writing process of this manuscript.

2. Mathematical Model and Main Results

2.1. **Notation and Assumptions.** We list notations, assumptions, results, and main mathematical arguments that we use in our work for convenience of the reader.

Assumption 2.1. *Let $\mathcal{Z} \subset \mathbb{R}^n$, $\mathcal{W} \subset \mathbb{R}^d$ and ρ be a probability density on $\mathcal{Z}$. The function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}_+$ satisfies:*

*(1) $\ell(z, \cdot)$ is (γ, L)**-smooth**, i.e. $\ell(z, \cdot)$ is differentiable and $\nabla \ell(z, \cdot)$ is (γ, L)-Hölder:*

$$\exists \gamma \in (0, 1] \text{ such that } \|\nabla \ell(z, u) - \nabla \ell(z, v)\| \le L \|u - v\|^\gamma \ \forall u, v \in \mathcal{W},\ z \in \mathcal{Z}.$$

(2) $\nabla \ell(\cdot, w) \in L^1(\mathcal{Z}, \mathscr{B}, \rho\, dz)$ for all $w \in \mathcal{W}$ where $\mathscr{B}$ is the Borel σ-algebra in $\mathbb{R}^N$.

(3) $F(w) = \mathbb{E}_\rho[\ell(Z, w)]$ is bounded below by $F_* = \inf_{w \in \mathcal{W}} F(w) > -\infty$ and a global minimizer $w_* \in \mathcal{W}$ exists, i.e. $F(w_*) = F_*$.

The assumption that $\nabla \ell$ is only γ-Hölder is used in [5, assumption 1], whose application is also mentioned in [5]. This is a weakening of the $\nabla \ell$ being Lipschitz used in [6, assumption 1]. However, thanks to a slight modification in the proof of Garrigos and Gower [2, lemma 2.25] (see proposition B.2), any function f that is (γ, L)-smooth still enjoys the property

$$f(y) \le f(x) + \langle \nabla f(x),\, y - x \rangle + \frac{L}{1 + \gamma} \|y - x\|^{1+\gamma} \ \forall x, y \in \mathbb{R}^d. \tag{2.1}$$

Symbol	Description
$\mathbb{N}_0 = \mathbb{N} \cup \{0\}$	The set of non-negative integers.
$\mathbb{R}_+$	The open interval $(0, \infty)$, i.e. the set of strictly positive real numbers.
$\lesssim$	$a \lesssim b$ means that there exists $c > 0$ such that $a \leq c \cdot b$.
$\|\cdot\|$	The standard Euclidean norm in $\mathbb{R}^d$.
$f(t) = o(g(t))$	$f(t)/g(t) \to 0$ as $t \to \infty$.
$f(t) = O(g(t))$	$\exists\, c > 0$ and $T > 0$ such that $f(t) \leq c\, g(t)$ for all $t > T$.
$f(t) = \Theta(g(t))$	$f(t) = O(g(t))$ and $g(t) = O(f(t))$.
$(\Omega, \mathscr{F}, \mathbb{P})$	The probability space for the paired process (Z_t, w_t) in eq. (1.1).
$\mathscr{F}_t$	The filtration $\{\mathscr{F}_t : t \in \mathbb{N}\}$ where $\mathscr{F}_t = \sigma\{w_s : s \leq t\}$.
$\mathscr{F}_\infty$	The σ-algebra $\bigcup_{t \in \mathbb{N}} \mathscr{F}_t$.
$m\mathscr{F}_t$	The set of measurable functions w.r.t $\mathscr{F}_t$ for $t \in \mathbb{N} \cup \{\infty\}$.
E_n eventually	The event $\bigcup_{m=1}^{\infty} \bigcap_{n \geq m} E_n$ ($\exists m > 0$ such that E_n occurs for all $n \geq m$).
$\ell(z, w)$	The estimator $\ell : \mathcal{Z} \times \mathcal{W} \to [0, \infty)$ of the objective function.
Z_t	An i.i.d sequence of random vectors in $\mathcal{Z} \subset \mathbb{R}^N$ that makes up the stochastic gradient estimator $\nabla \ell(Z_t, w_t)$.
ρ	The density function for Z_t.
F	The objective function $F : \mathcal{W} \to [0, \infty)$ defined by $F(w) = \mathbb{E}_\rho[\ell(Z_t, w)]$.
δm_t	The injected noise in stochastic algorithms, defined in eq. (1.2).

Table 2. Summary of Notations.

Now, observe that $\mathbb{E}_\rho[\nabla \ell(Z, w)]$ exists due to the second property in assumption 2.1. Such a property can be satisfied for example when $\mathcal{Z}$ is bounded with loss $\ell((x, y), w) = (y - \langle w,\, x \rangle)^2$. The assumption $\nabla \ell(\cdot, w) \in L^1(\mathcal{Z}, \mathscr{B}, \rho\, dz)$ is natural; in "averaged" gradient descent (GD)

$$w_{t+1} - w_t = -\alpha_t \frac{1}{n} \sum_{k=1}^{n} \nabla \ell(Z_k, w_t) \tag{2.2}$$

such an assumption implies convergence almost surely of the empirical average on the right hand side to $\mathbb{E}_\rho[\nabla \ell(Z, w_t)] = \nabla \mathbb{E}_\rho[\ell(Z, w_t)] = \nabla F(w_t)$ as the number of data points $n \uparrow \infty$. This suggests that (2.2) approximates the deterministic GD [2, algorithm 3.2]. In addition, the assumption $\nabla \ell(\cdot, w) \in L^1(\mathcal{Z}, \mathscr{B}, \rho\, dz)$ produces important consequences. The first consequence is due to proposition B.1: F inherits the properties in assumption 2.1 from the function $\ell(z, \cdot)$. Secondly, if the joint distribution of (Z_t, w_t) is the product measure $\rho\, dz \otimes \mu_t$, then

$$\begin{aligned} \int \mathbb{1}_{\{w_t \in A\}}\, \mathbb{E}[\nabla \ell(Z_t, w_t) \,|\, \mathscr{F}_t]\, d\mathbb{P} &= \int_A \int_{\mathcal{Z}} \nabla \ell(z, w)\, \rho(z)\, dz\, \mu_t(dw) \\ &= \int_A \mathbb{E}_\rho[\nabla \ell(Z, w)]\, \mu_t(dw) = \int_A \nabla \mathbb{E}_\rho[\ell(Z, w)]\, \mu_t(dw) \\ &= \int \mathbb{1}_{\{w_t \in A\}}\, \nabla F(w_t)\, d\mathbb{P}. \end{aligned} \tag{2.3}$$

Equation (2.3) implies $\mathbb{E}[\nabla \ell(Z_t, w_t) \,|\, \mathscr{F}_t] = \nabla F(w_t)$ and that the estimator $\nabla \ell(Z, w_t)$ is unbiased (w.r.t ρ). That $\nabla F(w_t)$ can be accessed through the estimator $\nabla \ell(Z, w_t)$ is also assumed in [6, the sentence right above assumption 4].

Assumption 2.2. *Using the same notation as in assumption 2.1, there exists $A, B, C \in \mathbb{R}_+$ such that $\mathbb{E}[\|\nabla \ell(Z_t, w_t)\|^{1+\gamma} \,|\, \mathscr{F}_t] \leq A(F(w_t) - F_*) + B\, \|\nabla F(w_t)\|^{1+\gamma} + C$ for all $t > 0$ a.s.-$\mathbb{P}$.*

The above assumption is called the ***ABC condition***, originally proposed by Khaled and Richtárik in [3], and is used by Liu and Yuan in [6, assumption 4] when F is $(1, L)$-smooth, i.e. $\gamma = 1$. It is said to be "the weakest assumption" for analysis of SGD in the non-convex setting [6, remark 1]. By

[5, lemma 14] (stated in lemma B.4), the first bounding term is a consequence of (γ, L)-smoothness and convexity:

$$\begin{aligned}\int \mathbb{1}_{\{w_t\in A\}}\, \mathbb{E}[\|\nabla\ell(Z_t,w_t)\|^{1+\gamma} \mid \mathscr{F}_t]\, d\mathbb{P} &= \int_A \mathbb{E}_\rho \|\nabla \ell(z,w)\|^{1+\gamma}\, \mu_t(dw)\\ &\le \int_A \Big\{ c_1(\beta,\gamma)(F(w)-F_*) + c_2(\beta,\gamma) + c_3(\beta,\gamma)\, \mathbb{E}_\rho[\|\nabla\ell(Z,w_*)\|^{1+\gamma}]\Big\}\, \mu_t(dw).\end{aligned}$$

Note that the last two terms in the integrand are constants and that the existence of a global minimizer $w_* \in \mathcal{W}$ is required (satisfied by assumption 2.1). When $\gamma = 1$ and the gradient estimator is unbiased, i.e. $\mathbb{E}_\rho[\nabla\ell(Z,w)] = F(w)$, the variance is $\mathbb{E}_\rho\|\nabla\ell(Z,w)-\nabla F(w)\|^2 = \mathbb{E}_\rho\|\nabla\ell(Z,w)\|^2 - \|\nabla F(w)\|^2$. Therefore, if the variance is bounded by $\sigma^2 < \infty$, then $\mathbb{E}_\rho\|\nabla\ell(Z,w)\|^2 \le \|\nabla F(w)\|^2 + \sigma^2$. This justifies the second term in the ABC assumption. For $\gamma < 1$, assuming that the gradient estimator has variance bounded by σ^2, we end up with

$$\mathbb{E}_\rho\|\nabla\ell(Z_t,w_t)\|^{1+\gamma} \le 2^{1+\gamma}(\|\nabla F(w_t)\|^{1+\gamma} + \mathbb{E}_\rho\|\delta m_t\|^{1+\gamma}) \le B\,\|\nabla F(w_t)\|^{1+\gamma} + C\,\sigma^{1+\gamma}.$$

In fact, with a simple estimate $a^{1+\gamma} \le (a+1)^{1+\gamma} \le (a+1)^2 \le 2a^2+2$, setting $a = \|\nabla F(w_t)\|$ gives

$$\mathbb{E}[\|\nabla\ell(Z_t,w_t)\|^{1+\gamma} \mid \mathscr{F}_t] \le 2B\,\|\nabla F(w_t)\|^2 + (C\,\sigma^{1+\gamma}+2) \text{ a.s.-}\mathbb{P}. \tag{2.4}$$

which has the form of the ABC condition in [6, assumption 4]. Further discussions on the ABC condition can be found in [6] and [3].

Assumption 2.3. $\sup_{z\in\mathcal{Z}} \ell(z,w_*) < \infty$.

Assumption 2.3 is also used in [5]. Such an assumption can be satisfied when $\mathcal{Z}$ is bounded. See discussions in [5]. Note that by lemma B.5, if $\ell(z,\cdot)$ is convex, then the 'noise' $\|\delta m_t\|^2$ defined in eq. (1.2) is bounded by a scalar multiple of $\|w_t - w_*\|^{2\gamma} + 1$. This assumption will only be used to produce probability estimates for the convergence rate of SHB iterates.

2.2. **Main Results.** With assumption 2.1 and assumption 2.2, in estimating $F(w_t) - F_*$, one naturally arrives at a recursive inequality. In this regard, our result of 'little-o' convergence lies in the following statement about recursive inequality, which is a deterministic, weaker version of the Robbins-Siegmund theorem [10].

Proposition 2.4. *Let X_t, Y_t, Z_t be non-negative for all $t \in \mathbb{N}_0$ and $a_t > 0$ be such that*

(1) $a_t, Z_t \in \ell^1(\mathbb{N})$.

(2) $Y_t \le (1+a_{t-1})Y_{t-1} - X_{t-1} + Z_{t-1}$.

Then $Y_t \in \ell^\infty(\mathbb{N})$ and $X_t \in \ell^1(\mathbb{N})$.

Using proposition 2.4, we conclude that $\|\nabla F(w_t)\|^2$ is decreasing eventually in corollary 3.2 and corollary 4.2. In such a case, $\|\nabla F(w_t)\|^2$ may fluctuate in $[0,\varepsilon)$ only finitely many times, but then it would have to reach and stay at zero either at finite time or at $+\infty$. This allows us to conclude a stronger result $\|\nabla F(w_t)\|^2 = o(t^{p-1})$ as $t \uparrow \infty$ when $\alpha_t = O(t^{-p})$ and $p \in (\frac{1}{1+\gamma}, 1)$.

When the objective function is convex, the set of critical points are all minima since there are no saddle points. Therefore, we have that $\|\nabla F(w_t)\|^2 \to 0$ if and only if w_t converges to a minimizer, which implies $F(w_t) - F_* \to 0$ due to the continuity of F. Moreover, we are able to prove a more direct result: $F(w_t) - F_*$ is decreasing eventually (proposition 3.5 and proposition 4.4). As a consequence, we can provide a rate of convergence $o(t^{p-1})$ consistent with results from prior works [5, theorem 6], [6, theorem 13], and [11, corollary 17].

Theorem 2.5. *Let w_t be the iterates in eq. (1.1). With the notation in table 2, if $\alpha_t = O(t^{-p})$ with $p \in (\frac{1}{1+\gamma}, 1)$, assumption 2.1 and assumption 2.2 hold, then the SHB converges almost surely at rate*

(1) $\|\nabla F(w_t)\|^2 = o(t^{p-1})$ *as* $t \uparrow \infty$.

(2) If in addition, $\ell(z, \cdot)$ *is convex for all* $z \in \mathcal{Z}$*, then* $F(w_t) - F_* = o(t^{p-1})$.

Our second result is an asymptotic statement about the chances that SHB iterates adhere to a certain convergence rate. In summary, the SHB with constant momentum parameter $\beta \in (0,1)$ is very likely to perform similar to SGD, with risk bounded by roughly $T^{p-1}(\log \frac{T}{\delta})^2$ for some tolerance level $\delta \in (0,1)$.

Theorem 2.6. *Let* z_t, w_t, v_t *be the SHB iterates as in eq.* (1.1) *and eq.* (4.2) *and* $K_0(L,\gamma,\beta)$, $K_5(L,\gamma,\beta)$ *be positive constants defined in eq.* (4.12) *and proposition 4.7. If* $\ell(z,\cdot)$ *is convex for all* $z \in \mathcal{Z}$*, assumption 2.1 and assumption 2.3 both hold with* $\gamma = 1$, $\beta \in (0,1)$*, and* α_t *satisfies*

(1) $\alpha_0 \leq \min\left(1, \dfrac{1}{2\sqrt{K_5}}, K_0\right)$.

(2) $\alpha_{t+1} \leq \alpha_t$ *for all* $t \in \mathbb{N}$.

(3) $\alpha_t = \Theta(t^{-p})$ *where* $p \in (\frac{1}{2}, 1)$.

then

$$\mathbb{P}\left[F(w_{T+1}) - F_* = O\left(T^{p-1}\left(\log\frac{T}{\delta}\right)^2\right)\right] \geq 1 - \delta.$$

Remark 2.7. *We remark that it is required to have* $\alpha_t = \Theta(t^{-p})$ *be* Θ*-bounded instead of* $O(t^{-p})$ *since we need to estimate ratio of the form* α_{t+s}/α_t *in (4.36) in addition to* $\sum \alpha_t$ *appearing in the denominator in the conclusion of theorem 4.14. If* $p = \frac{1}{2} + \epsilon$*, then* $F(w_{T+1}) - F_* = O(T^{-\frac{1}{2}+\epsilon}\left(\log\frac{T}{\delta}\right)^2)$*, consistent with the little-o convergence result of theorem 2.5 and* [11, corollary 17].

Theorem 2.6 is a consequence of theorem 4.14; a non-asymptotic estimate about the probability that at all times $T > 1$, the risk $F(w_{T+1}) - F_*$ can be bounded by $\max(T_0^{-p}, (\sum_{t=T_0}^{T}\alpha_t)^{-1})$ for any $T_0 \in (1, T]$ up to a logarithmic factor. Lemma B.3 asserts $F(w_t) - F_* \lesssim \|w_t - w_*\|^{1+\gamma}$ and $\|\nabla F(w_t)\|^2 \lesssim \|w_t - w_*\|^{2\gamma}$ which suggests that the rate of convergence for SHB and the 'magnitude' of the noise δm_t (under assumption 2.3) depend on the smoothness parameter γ and the distance to a minimizer. As such, we follow Lei, Shi, and Guo's approach in [5] to first prove proposition 4.9; up to time $T > 0$, the algorithm stays within a neighborhood of a minimum w_* up to $\log(T/\delta)$. Since the step size α_t is decaying faster than logarithm, based on eq. (1.2), intuitively, w_t would behave similar to the deterministic GD [2, algorithm 3.2] and approach a minimum. Following the same idea as Lei, Shi, Guo in [5], the key idea in proving theorem 2.6 is to employ Bernstein's and Azuma-Hoeffding's inequality to estimate martingales in processes involving SHB iterates to drop difficult terms and replace it with a term whose convergence rate is easier to compute. For example, in the following iteration,

$$\begin{aligned} F(w_{t+1}) - F_* \leq F(w_t) - F_* - \big(\alpha_t \|\nabla F(w_t)\|^2 - \alpha_t \langle \nabla F(w_t), \delta m_t\rangle\big) \\ + \beta\langle \nabla F(w_t), w_t - w_{t-1}\rangle + \frac{L}{1+\gamma}\|w_{t+1} - w_t\|^{1+\gamma} \end{aligned}$$

the martingale difference term $\alpha_t \langle \nabla F(w_t), \delta m_t\rangle$ can be made smaller than $\alpha_t \|\nabla F(w_t)\|^2$ so that the middle term can be dropped and replaced with a term whose decay rate is known (see lemma 4.12).

3. Stochastic Gradient Descent

Proposition 3.1. *If assumption 2.1, assumption 2.2 all hold and* $\alpha_t \in \ell^{1+\gamma}(\mathbb{N})$ *satisfies* $\alpha_t \leq \min\{1, \frac{1}{\sqrt[\gamma]{LB}}\}$ *for all* $t \geq 0$*, then the SGD algorithm (eq.* (1.1) *with* $\beta = 0$*) satisfies*

(1) $\sup_{t\geq 0}\mathbb{E}[F(w_t)-F_*]<\infty$ *and* $\sum_{t=1}^{\infty}\alpha_t\,\mathbb{E}\|\nabla F(w_t)\|^2<\infty$.

(2) $\|\nabla F(w_t)\|$ *is decreasing eventually.*

Proof. Due to (γ, L)-smoothness of F and assumption 2.2, by proposition B.2,

$$\begin{aligned}\mathbb{E}[F(w_t)-F_*-(F(w_{t-1})-F_*)\,|\,\mathscr{F}_{t-1}] &\leq \mathbb{E}[F(w_t)-F(w_{t-1})\,|\,\mathscr{F}_{t-1}]\\ &\leq \langle\nabla F(w_{t-1}),\,\mathbb{E}[w_t-w_{t-1}\,|\,\mathscr{F}_{t-1}]\rangle+\frac{L}{1+\gamma}\,\mathbb{E}[\|w_t-w_{t-1}\|^{1+\gamma}\,|\,\mathscr{F}_{t-1}]\\ &\leq -\alpha_{t-1}\|\nabla F(w_{t-1})\|^2+\frac{L}{1+\gamma}\,\alpha_{t-1}^{1+\gamma}\,\big(A\,(F(w_{t-1})-F_*)+B\,\|\nabla F(w_{t-1})\|^{1+\gamma}+C\big)\end{aligned}\tag{3.1}$$

By eq. (2.4), we get

$$\begin{aligned}\mathbb{E}[F(w_t)-F_*-(F(w_{t-1})-F_*)\,|\,\mathscr{F}_{t-1}] &\leq -\alpha_{t-1}\|\nabla F(w_{t-1})\|^2\\ &\quad+\frac{L}{1+\gamma}\,\alpha_{t-1}^{1+\gamma}\,\big(A\,(F(w_{t-1})-F_*)+2B\,\|\nabla F(w_{t-1})\|^2+2+C\big)\,.\end{aligned}\tag{3.2}$$

If $Y_t=\mathbb{E}[F(w_t)-F_*]$ and $LB\alpha_t^{\gamma}\leq 1$, then

$$Y_t\leq\left(1+\frac{LA}{1+\gamma}\,\alpha_{t-1}^{1+\gamma}\right)Y_{t-1}-\left(\frac{\gamma}{1+\gamma}\right)\alpha_{t-1}\,\mathbb{E}\|\nabla F(w_{t-1})\|^2+\frac{L\,(C+2)}{1+\gamma}\,\alpha_{t-1}^{1+\gamma}.\tag{3.3}$$

Applying proposition 2.4 yields the first claim.

To prove the second claim, using the fact that ∇F is (γ, L)-Hölder, we have, for any $\varepsilon>0$, $\mathbb{P}(\|\nabla F(w_t)\|-\|\nabla F(w_{t-1})\|\geq L\varepsilon^{\gamma})\leq\mathbb{P}(\|w_t-w_{t-1}\|\geq\varepsilon)$. By Chebychev's inequality, assumption 2.2, and the simple estimate $a^{1+\gamma}\leq 2a^2+2$ used to get (3.2), we obtain

$$\begin{aligned}\mathbb{P}(\|w_t-w_{t-1}\|\geq\varepsilon) &\leq\frac{\mathbb{E}[\|\alpha_{t-1}\,\nabla\ell(w_{t-1})\|^{1+\gamma}]}{\varepsilon^{1+\gamma}}\\ &\lesssim\alpha_{t-1}^{1+\gamma}\Big(\mathbb{E}[F(w_{t-1})-F_*]+2B\,\mathbb{E}\|\nabla F(w_{t-1})\|^2+(C+2)\Big).\end{aligned}\tag{3.4}$$

Since $\alpha_t\leq 1$ and $\alpha_t\in\ell^{1+\gamma}(\mathbb{N})$, by the first claim, all terms in the right hand side of (3.4) is summable. Applying Borel-Cantelli yields the desired result. □

Corollary 3.2. *With the assumptions in proposition 3.1,*

(1) $\mathbb{P}\left(\sum_{t=1}^{\infty}\alpha_t\,\|\nabla F(w_t)\|^2<\infty\right)=1$.

(2) $\mathbb{P}\left(\lim_{t\to\infty}F(w_t)-F_*<\infty\right)=1$.

(3) If $\alpha_t=\Theta(t^{-p})$ *for* $p\in(\frac{1}{1+\gamma},1)$, *then* $\|\nabla F(w_t)\|^2=o(t^{p-1})$ *almost surely.*

Proof. The first claim follows from the fact that $\sum\alpha_t\,\mathbb{E}\|\nabla F(w_t)\|^2$ is finite by proposition 3.1. The second claim follows similarly; if there is a positive probability that $\lim_{t\to\infty}F(w_t)=\infty$, then $\lim_{t\to\infty}\mathbb{E}[F(w_t)-F_*]=\infty$, contradicting proposition 3.1. For the third claim, by the first claim and proposition 3.1, $\|\nabla F(w_t)\|$ decreases eventually. This implies that there exists a random time $\tau\in m\mathscr{F}_\infty$ with $\tau<\infty$ a.s.-$\mathbb{P}$ such that $\|\nabla F(w_t)\|$ is decreasing when $t>\tau$. Therefore, $\sum_{t=\tau}^{\infty}\alpha_t\,\|\nabla F(w_t)\|^2$ is a sum of monotonically decreasing non-negative numbers that converges almost surely. By [4, theorem 3.3.1], we have $t^{1-p}\,\|\nabla F(w_t)\|^2\to 0$ almost surely. □

Remark 3.3. *Note that corollary 3.2 is a result for non-convex cost function F, comparable with* [2, thm 5.12] *by Garrigos and Gower for $\gamma=1$, which asserts that for a fixed end time $T\geq 1$, under a constant step size of the order $O(T^{-1/2})$, we have* $\min_{0\leq t<T}\mathbb{E}[\|\nabla F(w_t)\|^2]=O(T^{-1/2})$. *Moreover, for $\gamma=1$, corollary 3.2 shows that $\|\nabla F(w_t)\|^2$ converges to zero when $p\in(\frac{1}{2},1]$ which suggests that α_t must be in $(\ell^1(\mathbb{N}))^c\cap\ell^2(\mathbb{N})$, consistent with the result in* [9, theorem 2] *by Orabona. Note that*

the convergence rate of $o(t^{-\frac{1}{2}+\varepsilon})$ is achievable by setting $p = \frac{1}{2} + \varepsilon$, which is close to the optimal rate in expectation obtained by Agarwal et al. in [1].

Remark 3.4. *Note that if F is convex and (γ, L)-smooth, then F has no saddle points and all minima are global, i.e. all critical points in $\mathcal{W}^* = \{w \in \mathcal{W} : \|\nabla F(w_t)\|^2 = 0\}$ must all be minima. Therefore, $\|\nabla F(w_t)\|^2 \to 0$ if and only if there exists $w_* \in \mathcal{W}^*$ such that $\|w_t - w_*\|^2 \to 0$.*

Proposition 3.5. *If all assumptions in proposition 3.1 are met and in addition $\ell(z, \cdot)$ is convex for all $z \in \mathcal{Z}$, then*

(1) $F(w_t) - F_*$ is decreasing eventually.

(2) $\mathbb{P}\left(\sum_{t=1}^{\infty} \alpha_t (F(w_t) - F_*) < \infty\right) = 1$.

(3) If $\alpha_t = O(t^{-p})$ for $p \in (\frac{1}{1+\gamma}, 1)$, then $F(w_t) - F_* = o(t^{p-1})$ a.s.-$\mathbb{P}$.

Proof. Let $Y_t = \mathbb{E}[F(w_t) - F_*]$ and $Y^* = \sup_{t \geq 0} Y_t$. According to proposition 3.1, Y^* is finite. Observe that when F is convex and (γ, L)-smooth, using (3.3), we may bound Y_t above and below by summable terms

$$-\alpha_{t-1}\, \mathbb{E}\|\nabla F(w_{t-1})\|^2 \leq \mathbb{E}[F(w_t) - F(w_{t-1})] \leq \frac{L(AY^* + C + 2)}{1+\gamma}\, \alpha_{t-1}^{1+\gamma}. \tag{3.5}$$

This implies

$$\begin{aligned}
\mathbb{P}\left(F(w_t) - F_* - (F(w_{t-1}) - F_*) > \varepsilon\right) &= \mathbb{P}\left(F(w_t) - F(w_{t-1}) > \varepsilon\right) \\
&\leq \frac{\mathbb{E}|F(w_t) - F(w_{t-1})|}{\varepsilon} \\
&\leq \frac{1}{\varepsilon}\left(\frac{L(AY^* + C + 2)}{1+\gamma}\, \alpha_{t-1}^{1+\gamma} + \alpha_{t-1}\, \mathbb{E}\|\nabla F(w_{t-1})\|^2\right).
\end{aligned}$$

Therefore, by Borel-Cantelli, the first claim is proved. For the second claim, let $W_t = \mathbb{E}\|w_t - w_*\|^2$. By the convexity of F, we obtain the estimate

$$\begin{aligned}
W_{t+1} &= W_t + 2\, \mathbb{E}[\langle w_{t+1} - w_t,\, w_t - w_* \rangle] + \mathbb{E}\|w_{t+1} - w_t\|^2 \\
&= W_t - 2\alpha_t\, \mathbb{E}\langle \nabla F(w_t),\, w_t - w_* \rangle + \alpha_t^2 \mathbb{E}\|\nabla \ell(Z_t, w_t)\|^2 \\
&\leq W_t - 2\alpha_t\, \mathbb{E}[F(w_t) - F_*] + \alpha_t^2\left(A\, \mathbb{E}[F(w_t) - F_*] + B\, \alpha_t\, \mathbb{E}\|\nabla F(w_t)\|^2 + C\right).
\end{aligned}$$

Since $\alpha_t \in \ell^2(\mathbb{N})$, by proposition 3.1, the third term in the above estimate is summable. Therefore, we may apply proposition 2.4 and conclude $\sum \alpha_t\, \mathbb{E}[F(w_t) - F_*] < \infty$. The second claim follows. The third claim follows from the same argument as in the proof of corollary 3.2. □

4. Stochastic Heavy Ball

The algorithm for SHB is given in eq. (1.1). From [6, eq. (17) and (18)], by defining

$$\begin{cases} v_t = w_t - w_{t-1}\ ,\ \ z_t = w_t + \frac{\beta}{1-\beta}\, v_t\ ,\ \ t \geq 1. \\ w_1 = w_0. \end{cases} \tag{4.1}$$

the SHB can be rewritten as one-step iterates

$$\begin{cases} v_{t+1} = \beta\, v_t - \alpha_t\, \nabla \ell(Z_t, w_t) & t \geq 1 \\ z_{t+1} = z_t - \dfrac{\alpha_t}{1-\beta}\, \nabla \ell(Z_t, w_t) & t \geq 1 \\ z_1 = w_1\ ,\ \ v_1 = 0. \end{cases} \tag{4.2}$$

Proposition 4.1. *If assumption 2.1 and assumption 2.2 holds, then there exists a constant $b_0 > 0$ such that if $\alpha_t \in \ell^{1+\gamma}$ and $\alpha_t \le \min(1, b_0)$, then the SHB algorithm (4.2) corresponding to (1.1) satisfies*

(1) $\sup_{t\ge 0} \mathbb{E}[F(z_t) - F_* + \|v_t\|^2] < \infty$ *and* $\sum_{t=1}^{\infty} \alpha_t\, \mathbb{E}\|\nabla F(z_t)\|^2 < \infty$.

(2) $\|\nabla F(z_t)\|$ *and* $\|v_t\|$ *are both decreasing eventually.*

Proof. **Proof of claim 1.** Following [6, eq. (21)-(22)] in the proof of [6, theorem 8] and using the simple estimate $a^{1+\gamma} \le 2a^2 + 2$ as in eq. (2.4), we have

$$\|\nabla F(w_t)\|^2 \le 2\|\nabla F(z_t)\|^2 + c_1\, \|v_t\|^2 \tag{4.3}$$

$$F(w_t) - F_* \le F(z_t) - F_* + \frac{\|\nabla F(z_t)\|^2}{2} + \{c_2 + c_3\}\, \|v_t\|^2 + c_4 \tag{4.4}$$

where

$$c_1 := \frac{2L^2\beta^2}{(1-\beta)^2}\ ,\ c_2 := \frac{\beta^2}{2(1-\beta)^2}\ ,\ c_3 := \frac{L\beta^{1+\gamma}}{(1+\gamma)(1-\beta)^{1+\gamma}}\ ,\ c_4 := 2(c_3+1).$$

Observe that the left hand sides of (4.3) and (4.4) makeup the bounding terms in assumption 2.2. This implies that the upper bound of $\mathbb{E}[\nabla \ell(Z_t, w_t)\,|\,\mathscr{F}_t]$ can be expressed entirely in terms of SHB variables $F(z_t) - F_*$, $\|\nabla F(z_t)\|^2$, and $\|v_t\|^2$. As a consequence, we may estimate the iterates of the SHB variables as follows

$$\begin{aligned}
\mathbb{E}[\|v_{t+1}\|^2\,|\,\mathscr{F}_t] &\le \beta^2\,(1+\varepsilon_1)\,\|v_t\|^2 + \frac{1}{\varepsilon_1}\,\alpha_t^2\,\|\nabla F(w_t)\|^2 \\
&\quad + \alpha_t^2\left(A\,[F(w_t) - F_*] + B\,\|\nabla F(w_t)\|^2 + C\right) \\
&\le \left(\beta^2\,(1+\varepsilon_1) + \left\{\left(\frac{1}{\varepsilon_1} + B\right) c_1 + A\,c_2 + A\,c_3\right\}\,\alpha_t^2\right)\|v_t\|^2 \\
&\quad + \left(\frac{2}{\varepsilon_1} + 2B + \frac{A}{2}\right)\alpha_t^2\,\|\nabla F(z_t)\|^2 + \alpha_t^2\left(C + A\,c_4\right) \\
&\quad + A\,\alpha_t^2\,[F(z_t) - F_*]
\end{aligned} \tag{4.5}$$

where ε_1 is an arbitrary constant. In addition, following [6, eq.(20)], we get

$$\begin{aligned}
\mathbb{E}[F(z_{t+1})\,|\,\mathscr{F}_t] &\le F(z_t) - \left(\frac{1}{1-\beta} - \frac{\alpha_t\,c_5}{\varepsilon_2}\right)\alpha_t\,\|\nabla F(z_t)\|^2 + \varepsilon_2\|v_t\|^2 \\
&\quad + c_6(L,\beta,\gamma)\alpha_t^{1+\gamma}\,(A[F(w_t) - F_*] + B\|\nabla F(w_t)\|^2 + C) \\
&\le F(z_t) - \left(\frac{1}{1-\beta} - \frac{\alpha_t\,c_5}{\varepsilon_2} - \left(B\,c_6 + \frac{A}{2}\right)\alpha_t^{\gamma}\right)\alpha_t\,\|\nabla F(z_t)\|^2 \\
&\quad + \left(\varepsilon_2 + (c_1\,c_6\,B + A\,c_2 + A\,c_3)\,\alpha_t^{1+\gamma}\right)\|v_t\|^2 \\
&\quad + A\,c_6\,\alpha_t^{1+\gamma}\,[F(z_t) - F_*] + (C + A\,c_4)\,\alpha_t^{1+\gamma}
\end{aligned} \tag{4.6}$$

where ε_2 is an arbitrary constant and

$$c_5 := \frac{c_1(L,\beta)}{8(1-\beta)}\ ,\ c_6 := \frac{c_3(L,\beta,\gamma)}{\beta^{1+\gamma}}$$

Since $\beta < 1$, we take $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$ such that $\beta^2(1+\varepsilon_1) + \varepsilon_2 \le 1$, e.g. set $\varepsilon_1 = \frac{1-\beta^2}{2\beta^2}$ and $\varepsilon_2 = \frac{1-\beta^2}{2}$. Now, adding (4.5) and (4.6) yields that there exists positive constants $b_0, b_1, b_2, b_3 \in \mathbb{R}_+$ such that if $\alpha_t \leq \min(1, b_0)$, then

$$\begin{aligned} \mathbb{E}[F(z_{t+1}) - F_* + \|v_{t+1}\|^2 \,|\, \mathscr{F}_t] &\leq (1 + b_1\, \alpha_t^{1+\gamma})(F(z_t) - F_* + \|v_t\|^2) \\ &\quad - b_2\, \alpha_t\, \|\nabla F(z_t)\|^2 + b_3\, \alpha_t^{1+\gamma}. \end{aligned} \tag{4.7}$$

Applying the expectation on both sides and proposition 2.4 yields the first claim.

Proof of claim 2. To prove the second claim, observe that (4.3), (4.4), assumption 2.2, and the first claim implies that the upper bound for $\mathbb{E}[\|\nabla\ell(Z_t, w_t)\|^2]$ are bounded, except for the term $\|\nabla F(z_t)\|^2$. Therefore, by Chebychev's inequality, we have

$$\mathbb{P}(\|\alpha_t\, \nabla\ell(Z_t, w_t)\| > \varepsilon) \leq \frac{\alpha_t^{1+\gamma}\, \mathbb{E}\|\nabla\ell(Z, w_t)\|^{1+\gamma}}{\varepsilon^{1+\gamma}} \lesssim \alpha_t^{1+\gamma} + \alpha_t\, \|\nabla F(z_t)\|^2. \tag{4.8}$$

Observe that the right hand side is summable in t and that the right hand side is the SHB iterates (4.1). Since ∇F is (γ, L)-Hölder and $0 \leq \beta < 1$, we get the following estimates

$$\begin{gathered} \|\nabla F(z_t)\| - \|\nabla F(z_{t-1})\| \leq L\, \|z_t - z_{t-1}\|^\gamma \leq \|\alpha_t\, \nabla\ell(Z_t, w_t)\|^\gamma \\ \|v_t\| - \|v_{t-1}\| \leq \|v_t\| - \beta\, \|v_{t-1}\| \leq \|v_t - \beta\, v_{t-1}\| \leq \|\alpha_t\, \nabla\ell(Z_t, w_t)\|. \end{gathered}$$

Combining these estimates with (4.8) and applying Borel-Cantelli yields the second claim. □

Corollary 4.2. *With the assumptions in proposition 3.1 and w_t, z_t, v_t as in eq.* (1.1) *and eq.* (4.2)*,*

(1) $\mathbb{P}\left(\sum_{t=1}^{\infty} \alpha_t\, \|\nabla F(z_t)\|^2 < \infty\right) = 1.$

(2) $\mathbb{P}\left(\lim_{t\to\infty} F(w_t) - F_* < \infty\right) = \mathbb{P}\left(\lim_{t\to\infty} F(z_t) - F_* < \infty\right) = 1.$

(3) $\mathbb{P}\left(\sum_{t=1}^{\infty} \|v_t\|^2 < \infty\right) = 1$ *and* $\|v_t\|^2 = o(t^{-1})$ *a.s.-*$\mathbb{P}$.

(4) If $\alpha_t = O(t^{-p})$ *for* $p \in (\frac{1}{1+\gamma}, 1)$*, then* $\|\nabla F(z_t)\|^2 = o(t^{p-1})$ *and* $\|\nabla F(w_t)\|^2 = o(t^{p-1})$ *a.s.-*$\mathbb{P}$.

Proof. The first claim follows from the first conclusion in proposition 4.1. The second claim holds if the third claim is applied to (4.4). For the third claim, first, observe that the iterates of v_t satisfies

$$\begin{aligned} \mathbb{E}\|v_{t+1}\|^2 &= \beta^2\, \mathbb{E}\|v_t\|^2 - 2\beta\, \alpha_t\, \mathbb{E}[\langle\nabla F(w_t),\, v_t\rangle] + \alpha_t^2\, \mathbb{E}\|\nabla\ell(Z, w_t)\|^2 \\ &\leq \mathbb{E}\|v_t\|^2 - 2\beta\, \alpha_t\, \mathbb{E}[\langle\nabla F(w_t),\, v_t\rangle] + \alpha_t^2\, \mathbb{E}\|\nabla\ell(Z, w_t)\|^2. \end{aligned} \tag{4.9}$$

Using assumption 2.2, the first claim, along with $\ell^{1+\gamma} \subset \ell^2$, (4.3), and (4.4) yields that the last term in (4.9) is summable:

$$\sum_{t=1}^{\infty} \alpha_t^2 \left(\mathbb{E}\|\nabla F(z_t)\|^2 + \sup_{t\geq 0}\, \mathbb{E}\left[F(z_t) - F_* + \|v_t\|^2\right] + 1\right) < \infty.$$

Therefore, by proposition 2.4, the middle term in (4.9) is also summable: $\sum \alpha_t \mathbb{E}[\langle\nabla F(w_t),\, v_t\rangle] < \infty$. Now we have that $\mathbb{E}\|v_{t+1}\|^2 - \beta^2\mathbb{E}\|v_t\|^2$ corresponds to summable terms. Since $v_1 = 0$, we know $\sum_{t=1}^{\infty} \|v_t\|^2 = \sum_{t=1}^{\infty} \|v_{t+1}\|^2$. Therefore,

$$(1 - \beta^2) \sum_{t=1}^{\infty} \mathbb{E}\|v_t\|^2 = -2\beta \sum_{t=1}^{\infty} \alpha_t\, \mathbb{E}[\langle\nabla F(w_t),\, v_t\rangle] + \sum_{t=1}^{\infty} \alpha_t^2\, \mathbb{E}\|\nabla\ell(Z_t, w_t)\|^2 < \infty.$$

This implies $\sum_{t=1}^{\infty} \|v_t\|^2 < \infty$ a.s.-$\mathbb{P}$. By proposition 4.1, $\|v_t\|$ is decreasing eventually. It follows that if $\tau \in m\mathscr{F}_\infty$ is the random time where $\|v_t\|$ is decreasing for $t > \tau$, then $\tau < \infty$ a.s.-$\mathbb{P}$ so that

$$(1 - \beta^2) \left(\sum_{t=1}^{\tau} + \sum_{t=\tau}^{\infty}\right) \|v_t\|^2 \leq -2\beta \sum_{t=1}^{\infty} \alpha_t \langle\nabla\ell(Z_t, w_t),\, v_t\rangle + \sum_{t=1}^{\infty} \alpha_t^2\, \|\nabla\ell(Z_t, w_t)\|^2 < \infty. \tag{4.10}$$

Since $\tau < \infty$ a.s.-$\mathbb{P}$, the above implies $\sum_{t=\tau}^{\infty} \|v_t\|^2 < \infty$ a.s.-$\mathbb{P}$ with monotonically decreasing summands. The claim follows from [4, theorem 3.3.1]. The last claim follows from (4.3) and [4, theorem 3.3.1] using an argument similar to the proof in corollary 3.2. In particular, we have $\|\nabla F(w_t)\|^2 = o(t^{\max(p-1,-1)}) = o(t^{p-1})$. □

Remark 4.3. *Corollary 4.2 shows that the SHB converges at rate similar to that of SGD. In particular, the SHB parameter β in eq.* (1.1) *does not impact the almost sure convergence rate, consistent with Liu and Yuan's result* [6, theorem 8]*. In addition, that the SHB can achieve $o(t^{-\frac{1}{2}+\epsilon})$ when the step size is $O(t^{-\frac{1}{2}-\epsilon})$ is also consistent with* [11, corollary 17].

Proposition 4.4. *If all assumptions in proposition 4.1 are satisfied and $\ell(z,\cdot)$ is convex for all $z \in \mathcal{Z}$ and $\alpha_t = O(t^{-p})$ for $p \in (\frac{1}{1+\gamma}, 1)$ and $\alpha_t \leq \alpha_{t-1}$ for all $t > 0$, then*

(1) $F(z_t) - F_*$ is decreasing eventually.

(2) $\mathbb{P}\left(\sum \alpha_t \left(F(w_t) - F_*\right) < \infty\right) = 1$.

(3) $F(w_t) - F_* = o(t^{p-1})$ a.s.-$\mathbb{P}$.

Proof. To prove the first claim, if we define $\zeta_t = \mathbb{E}[F(z_t) - F_*]$, then by the convexity of F,

$$\begin{aligned}\zeta_{t+1} - \zeta_t &\geq \mathbb{E}[\langle \nabla F(z_t),\, z_{t+1} - z_t\rangle] \\ &= -\alpha_t\, \mathbb{E}[\langle \nabla F(z_t),\, \nabla F(w_t)\rangle] \\ &\geq -\frac{\alpha_t}{2}\Big(\mathbb{E}\|\nabla F(z_t)\|^2 + \mathbb{E}\|\nabla F(w_t)\|^2\Big).\end{aligned}$$

By proposition 4.1, corollary 4.2, and (4.3), the right hand side is summable. Also, by (γ, L)-smoothness, we have $\zeta_{t+1} - \zeta_t \lesssim \mathbb{E}[\langle \nabla F(z_t),\, z_{t+1} - z_t\rangle] + \alpha_t^2 \mathbb{E}\|\nabla \ell(Z_t, w_t)\|^2$ where both terms on the right hand side are summable. Therefore, by Chebychev's inequality,

$$\begin{aligned}\mathbb{P}\left(F(z_{t+1}) - F(z_t) > \varepsilon\right) &\leq \frac{\mathbb{E}|F(z_{t+1}) - F(z_t)|}{\varepsilon} \\ &\leq \frac{1}{\varepsilon}\left(\alpha_t\, \mathbb{E}\|\nabla F(w_t)\|^2 + \alpha_t\, \mathbb{E}\|\nabla F(z_t)\|^2 + \alpha_t^2\, K_{ABC}\right).\end{aligned} \tag{4.11}$$

Note that we have used proposition 4.1 and assumption 2.2 for the second inequality to conclude that $\mathbb{E}\|\nabla \ell(Z_t, w_t)\|^2 \leq K_{ABC} < \infty$ is uniformly bounded by a constant K_{ABC}. Since the right hand side in (4.11) is summable, by Borel-Cantelli, the first claim is proved.

For the second claim, if we define $\eta_t = \mathbb{E}\|z_t - w_*\|^2$, then by the convexity of F,

$$\begin{aligned}\eta_{t+1} &= \eta_t - 2\alpha_t\, \mathbb{E}[\langle \nabla F(w_t),\, z_t - w_*\rangle] + \alpha_t^2\, \mathbb{E}\|\nabla \ell(Z_t, w_t)\|^2 \\ &= \eta_t - 2\alpha_t\left(\mathbb{E}[\langle \nabla F(w_t),\, w_t - w_*\rangle] + \frac{\beta}{1-\beta}\mathbb{E}[\langle \nabla F(w_t),\, v_t\rangle]\right) + \alpha_t^2 \mathbb{E}\|\nabla \ell(Z_t, w_t)\|^2 \\ &\leq \eta_t - 2\,\alpha_t\, \mathbb{E}[F(w_t) - F_*] + \left(\frac{2\beta}{1-\beta}\alpha_t(F(w_{t-1}) - F(w_t)) + \alpha_t^2\, K_{ABC}\right) \\ &\leq \eta_t - 2\,\alpha_t\, \mathbb{E}[F(w_t) - F_*] + \left(\frac{2\beta}{1-\beta}\Big(\alpha_{t-1}\mathbb{E}[F(w_{t-1})] - \alpha_t \mathbb{E}[F(w_t)]\Big) + \alpha_t^2\, K_{ABC}\right).\end{aligned}$$

Note that we have used convexity of F and that α_t is decreasing in the second inequality. Note that terms in the parentheses are summable. Applying proposition 2.4 yields that $\sum \alpha_t \mathbb{E}[F(w_t) - F_*] < \infty$ for which the second claim follows.

For the third claim, observe that (γ, L)-smoothness gives

$$\begin{aligned}F(z_t) - F_* &\leq F(w_t) - F_* + \langle \nabla F(w_t),\, z_t - w_t\rangle + \frac{L}{1+\gamma}\|z_t - w_t\|^{1+\gamma} \\ &\leq F(w_t) - F_* + \frac{\|F(w_t)\|^2}{2} + \frac{\beta^2}{2(1-\beta)^2}\|v_t\|^2 + \frac{L}{1+\gamma}\left(\frac{\beta}{1-\beta}\right)^{1+\gamma}\|v_t\|^{1+\gamma}.\end{aligned}$$

Multiplying both sides by α_t yields that the right hand side is summable by the second claim. Therefore, we have $\sum \alpha_t\,(F(z_t) - F_*) < \infty$ a.s.-$\mathbb{P}$. Together with the first claim, following the same argument as in the proof of corollary 4.2, by [4, theorem 3.3.1], we have $F(z_t) - F_* = o(t^{p-1})$. We may estimate $F(w_t) - F_*$ in terms of $F(z_t) - F_*$ as follows

$$\begin{aligned}
F(w_t) - F_* &\le F(z_t) - F_* + \langle \nabla F(z_t),\, w_t - z_t\rangle + \frac{L}{1+\gamma}\|w_t - z_t\|^{1+\gamma} \\
&\le F(z_t) - F_* + \frac{\|F(z_t)\|^2}{2} + \frac{\beta^2}{2(1-\beta)^2}\|v_t\|^2 + \frac{L}{1+\gamma}\left(\frac{\beta}{1-\beta}\right)^{1+\gamma}\|v_t\|^{1+\gamma}.
\end{aligned}$$

By corollary 4.2, all terms on the right hand side are known to be $o(t^{p-1})$. Therefore, the third claim follows. □

4.1. **Convergence Rate with High Probability in the Convex Setting.** In this section, we compute the probability of last iterate SHB error rate via Azuma-Hoeffding and Bernstein's inequality following the work by Lei, Shi, Guo in [5]. We will assume throughout this section that $\ell(z,\cdot)$ is convex for all $z \in \mathcal{Z}$, which by proposition B.1, makes F also convex. It will be convenient to define once and for all the following constants and notations. Let a_1 and a_2 be the constants defined in lemma B.5. Also, define the following

$$\begin{aligned}
&k_0 := \frac{\beta}{1-\beta}\ ,\ k_1 := \frac{a_1(L,\gamma)}{(1-\beta)^2}\ ,\ k_2 := \frac{a_2(L,\gamma)}{(1-\beta)^2}\ ,\ k_3 := \frac{2}{1-\beta}\cdot \sup_{z\in\mathcal{Z}} \ell(z, w_*) \\
&k_4 := k_0 + \beta\, k_0^2\ ,\ k_5 := k_2 + k_3 + a_2\, k_0^2 \\
&K_0 := \min\left(1, \frac{1}{(a_1\, k_0^2 + a_1\, k_0 + a_1 + k_1)(1-\beta)}\right) \\
&\Delta z_t = \|z_t - w_*\|^2 - \|z_{t-1} - w_*\|^2\ ,\ \Delta v_t = \|v_t\|^2 - \|v_{t-1}\|^2
\end{aligned} \tag{4.12}$$

Lemma 4.5. *Let z_t, w_t, v_t be the SHB iterates as in eq.* (1.1) *and eq.* (4.2)*. With the notation in eq.* (4.12)*, if $\ell(z,\cdot)$ is convex for all $z \in \mathcal{Z}$, assumption 2.1 and assumption 2.3 both hold, and α_t satisfies*

(1) $\alpha_0 \le K_0$.

(2) $\alpha_{t+1} \le \alpha_t$ *for all* $t \ge 0$.

(3) $\alpha_t \in \ell^2(\mathbb{N})$.

then there exists constants $K_1(L,\beta,\gamma,w_0) > 0$ and $K_2(L,\beta,\gamma,w_0) > 0$ such that for all $t > 0$,

(1) $\max\left\{\|w_t - w_*\|^2, \|v_t\|^2, \|z_t\|^2\right\} \le K_1 \sum_{s=0}^{t-1} \alpha_s$ *a.s.-*$\mathbb{P}$.

(2) $\sum_{s=1}^{t-1} \alpha_s^2\, \ell(Z_s, w_s) \le K_2$ *a.s.-*$\mathbb{P}$.

Proof. Consider eq. (4.2) for the iterates of v_t. Take the squared-norm. Applying lemma B.5 yields

$$\begin{aligned}
\|v_{t+1}\|^2 - \|v_t\|^2 &\le \|v_{t+1}\|^2 - \beta^2\,\|v_t\|^2 \\
&= -2\,\alpha_t\,\beta\,\langle \nabla \ell(Z_t, w_t),\, v_t\rangle + \alpha_t^2\,\|\nabla \ell(Z_t, w_t)\|^2 \\
&\le -2\,\alpha_t\,\beta\,\langle \nabla \ell(Z_t, w_t),\, v_t\rangle + a_1\,\alpha_t^2\,\ell(Z_t, w_t) + a_2\,\alpha_t^2
\end{aligned} \tag{4.13}$$

Using the notation in eq. (4.12), convexity and lemma B.5 yields

$$\begin{aligned}
\Delta z_{t+1} &= -2\,\alpha_t\,\langle \nabla\ell(Z_t,w_t),\, \frac{\beta v_t + w_t - w_*}{1-\beta}\rangle + \frac{1}{(1-\beta)^2}\,\alpha_t^2\,\|\nabla\ell(Z_t,w_t)\|^2 \\
&\le -2\,\alpha_t\,\langle \nabla\ell(Z_t,w_t),\, \frac{\beta v_t}{1-\beta}\rangle + \frac{2}{1-\beta}\,\alpha_t\Big(\ell(Z_t,w_*) - \ell(Z_t,w_t)\Big) + k_1\,\alpha_t^2\,\ell(Z_t,w_t) + k_2\,\alpha_t^2 \\
&\le -2\,\alpha_t\,\langle \nabla\ell(Z_t,w_t),\, \frac{\beta v_t}{1-\beta}\rangle + \alpha_t\,(k_3 + k_2\,\alpha_t) - \alpha_t\,\ell(Z_t,w_t)\bigg(\frac{2}{1-\beta} - \alpha_t\,k_1\bigg).
\end{aligned} \tag{4.14}$$

Now, let $A_t = \sum_{s=1}^{t-1}\alpha_s$. Note that $v_1 = 0$ and $z_1 = w_0$. Multiplying (4.13) by k_0^2, adding (4.14), and summing from $s=1$ to $s=t-1$ yields

$$\begin{aligned}
\|z_t - w_*\|^2 + k_0^2\,\|v_t\|^2 \le \|w_0 - w_*\|^2 &+ k_5\,A_t \\
&- 2\beta\left(\frac{1}{1-\beta} + k_0^2\right)\sum_{s=1}^{t-1}\alpha_s\,\langle\nabla\ell(Z_s,w_s),\, v_s\rangle \\
&- \sum_{s=1}^{t-1}\alpha_s\,\ell(Z_s,w_s)\bigg(\frac{2}{1-\beta} - \alpha_s\,(k_1 + a_1 k_0^2)\bigg).
\end{aligned} \tag{4.15}$$

Let X_t be the last term in (4.15) and $Y_t = 2k_0\langle w_0 - w_*,\, v_t\rangle$. Using (4.15) and $z_1 = w_0$, we get the estimate

$$\begin{aligned}
\|w_t - w_*\|^2 &= \|z_t - w_*\|^2 + k_0^2\,\|v_t\|^2 - 2k_0\,\langle z_t - w_*,\, v_t\rangle \\
&= \|z_t - w_*\|^2 + k_0^2\,\|v_t\|^2 - 2k_0\langle z_1 + w_*,\, v_t\rangle - 2k_0\sum_{s=1}^{t-1}\langle z_{s+1} - z_s,\, v_s\rangle \\
&= \|z_t - w_*\|^2 + k_0^2\,\|v_t\|^2 - 2k_0\langle w_0 - w_*,\, v_t\rangle + \frac{2k_0}{1-\beta}\sum_{s=1}^{t-1}\alpha_s\,\langle\nabla\ell(Z_s,w_s),\, v_s\rangle \\
&\le \|w_0 - w_*\|^2 + k_5\,A_t + X_t - Y_t + 2\,\beta\,k_0\sum_{s=1}^{t-1}\alpha_s\,\langle\nabla\ell(Z_s,w_s),\, v_s\rangle.
\end{aligned} \tag{4.16}$$

Now, recalling $v_1 = 0$ and rearranging (4.13) and summing from $s=1$ to $s=t-1$ yields

$$2\beta\sum_{s=1}^{t-1}\alpha_s\langle\nabla\ell(Z_s,w_s),\, v_s\rangle \le -\|v_t\|^2 + a_1\sum_{s=1}^{t-1}\alpha_s^2\,\ell(Z_s,w_s) + a_2\sum_{s=1}^{t-1}\alpha_s^2.$$

Combining this with (4.16) yields that the middle term $a_1\sum\alpha_s^2\ell(Z_s,w_s)$ can be collected with X_t so that

$$\begin{aligned}
\|w_t - w_*\|^2 \le \|w_0 - w_*\|^2 + k_5\,A_t - Y_t - 2k_0\|v_t\|^2 - k_0 a_2\sum_{s=1}^{t-1}\alpha_s^2 \\
- \sum_{s=1}^{t-1}\alpha_s\,\ell(Z_s,w_s)\bigg(\frac{2}{1-\beta} - \alpha_s\,(k_1 + a_1 k_0^2 + a_1 k_0)\bigg).
\end{aligned} \tag{4.17}$$

Requiring α_t to be small enough, we may drop the last two terms on the right hand side of (4.17). Now, using the inequality $2|\langle w_0 - w_*,\, v_t\rangle| \le \varepsilon\|w_0 - w_*\|^2 + \frac{1}{\varepsilon}\|v_t\|^2$ and setting $\varepsilon = 1/2$ yields

$$\begin{aligned}
\|w_t - w_*\|^2 &\le \|w_0 - w_*\|^2 + k_5\,A_t - Y_t - 2k_0\|v_t\|^2 \\
&\le \|w_0 - w_*\|^2 + k_5\,A_t + 2k_0|\langle w_0 - w_*,\, v_t\rangle| - 2k_0\|v_t\|^2 \\
&\le \|w_0 - w_*\|^2 + k_5\,A_t + \frac{k_0}{2}\|w_0 - w_*\|^2.
\end{aligned} \tag{4.18}$$

Since we define $\alpha_0 > 0$, we have $\|w_t - w_*\|^2 \lesssim \sum_{s=0}^{t-1} \alpha_s$ as desired. Since $v_t = w_t - w_{t-1}$, we know that $\|v_t\|^2$ has the same order as $\|w_t - w_*\|^2$. Similarly, since $z_t = w_t + k_0\, v_t$, it follows that $\|z_t\|^2$ also has the same order as $\|w_t - w_*\|^2$.

Now, define $\Delta v_{t+1}^{\beta} = \beta^2 \|v_t\|^2 - \|v_{t+1}\|^2$. Observe that

$$\begin{aligned}\Delta z_{t+1} - \Delta v_{t+1}^{\beta} &\leq (k_3 + k_2\, \alpha_t)\alpha_t - \alpha_t\, \ell(Z_t, w_t) \left(\frac{2}{1-\beta} - \alpha_t\, k_1\right) - \alpha_t^2 \|\nabla \ell(Z_t, w_t)\|^2 \\ &\leq k_5\, \alpha_t - \alpha_t\, \ell(Z_t, w_t) \left(\frac{2}{1-\beta} - \alpha_t\, k_1\right) + \alpha_t^2 \|\nabla \ell(Z_t, w_t)\|^2 \\ &\leq (k_5 + a_2 \alpha_t)\, \alpha_t - \alpha_t\, \ell(Z_t, w_t) \left(\frac{2}{1-\beta} - \alpha_t\, (k_1 + a_1)\right).\end{aligned}$$

If α_t is small enough and is decreasing, rearranging the above by moving the last term to the left hand side and multiplying by α_t yields

$$\begin{aligned}\left(\frac{1}{1-\beta}\right) \alpha_t^2\, \ell(Z_t, w_t) &\leq (k_5 + a_2\, \alpha_t)\, \alpha_t^2 - \Delta z_{t+1} + \Delta v_{t+1}^{\beta} \\ &= (k_3 + a_2\, \alpha_t)\, \alpha_t^2 - \alpha_t \Delta z_{t+1} + \alpha_t \beta^2 \|v_t\|^2 - \alpha_t \|v_{t+1}\|^2 \\ &\leq (k_3 + a_2)\, \alpha_t^2 + \alpha_t \|z_t\|^2 - \alpha_{t+1} \|z_{t+1}\|^2 + \alpha_t \|v_t\|^2 - \alpha_{t+1} \|v_{t+1}\|^2.\end{aligned} \tag{4.19}$$

Since $\alpha_t \in \ell^2$, summing the above and making use of the first claim completes the proof. □

Remark 4.6. *Note that the β-parameter does not affect the long term fluctuation of the iterates as can be seen from the conclusion of lemma 4.5 and from setting $\beta = 0$ directly in estimates (4.18) and (4.19).*

In the sequel, we will require estimating constants of the form $\alpha_t \sum_{s=1}^{t} \alpha_s$. Under the assumption $\alpha_t \in \ell^{1+\gamma}(\mathbb{N})$ for $\gamma \in (0, 1]$, we have $\alpha_t \sum_{s=0}^{t-1} \alpha_s \leq \alpha_t^{1-\gamma} \sum_{s=0}^{t-1} \alpha_s^{1+\gamma} < \infty$. Moreover, such a sum converges to zero for $\gamma \in (0, 1)$. If $\gamma = 1$, $\alpha_0 = K_0$, $\alpha_t \leq K_0\, t^{-p}$ for $t > 0$, then

$$\alpha_t \sum_{s=0}^{t-1} \alpha_s \leq K_0^2\, t^{-p} \left(1 + \sum_{s=1}^{t-1} s^{-p}\right) \lesssim K_0^2\, \max(t^{-p}, (t-1)^{-2p+1}) \leq K_0^2 \leq K_0\ ,\ \forall t > 0\ . \tag{4.20}$$

Proposition 4.7. *With the notation in lemma 4.5, if all assumptions in lemma 4.5 holds and in addition $\alpha_t \leq K_0 t^{-p}$ for $p \in (\frac{1}{2}, 1)$, then there exists positive constants $K_3, K_4, K_5 \in \mathbb{R}_+$ depending only on L, γ, β such that for any $t > 0$ and $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\|z_t - w_*\|^2 + \left(\frac{\beta}{1-\beta}\right)^2 \|v_t\|^2 \leq K_3 + K_4 \log \frac{1}{\delta} + K_5 \sum_{s=1}^{t-1} \alpha_s^2\, \|w_s - w_*\|^2\right) \geq 1 - \delta.$$

Proof. We use the notation of constants in eq. (4.12). Using the convexity of F, summing (4.13) and (4.14) yields

$$\begin{aligned}\Delta z_{t+1} + k_0^2\, \Delta v_{t+1} &\leq -2\, \alpha_t\, \beta \left(\frac{1}{1-\beta} + k_0^2\right) \langle \nabla \ell(Z_t, w_t),\, v_t \rangle - \frac{2\alpha_t}{1-\beta} \langle \nabla \ell(Z_t, w_t),\, w_t - w_* \rangle + \varphi_t \\ &= -2\, \alpha_t\, k_4\, \langle \nabla F(w_t),\, w_t - w_{t-1} \rangle - \frac{2\alpha_t}{1-\beta}\, \langle \nabla F(w_t),\, w_t - w_* \rangle + \frac{2\xi_t}{1-\beta} + \varphi_t \\ &\leq 2\, k_4\, \alpha_t \big(F(w_{t-1}) - F(w_t)\big) + \frac{2\alpha_t}{1-\beta} \big(F(w_*) - F(w_t)\big) + \frac{2\xi_t}{1-\beta} + \varphi_t \\ &\leq 2\, k_4\, \big(\alpha_{t-1} F(w_{t-1}) - \alpha_t F(w_t)\big) + \frac{2}{1-\beta}\, \big(\alpha_t\, \big(F(w_*) - F(w_t)\big) + \xi_t\big) + \varphi_t\end{aligned} \tag{4.21}$$

where ξ_t and φ_t are defined as $\mathscr{F}_t$-adapted processes

$$\begin{aligned}\xi_t &= (1-\beta)k_4\alpha_t\langle \delta m_t,\, v_t\rangle + \alpha_t\langle \delta m_t,\, w_t - w_*\rangle.\\ \varphi_t &= \alpha_t^2\left(\left(a_1k_0^2+k_1\right)\ell(Z_t,w_t) + a_2k_0^2 + k_2\right).\end{aligned} \tag{4.22}$$

By lemma 4.5, we have

$$\sum_{t=1}^{\infty}\varphi_t \le (a_1k_0^2+k_1)K_2 + (a_2k_0^2+k_2) < \infty \text{ a.s.-}\mathbb{P}. \tag{4.23}$$

Also, we have $\mathbb{E}[\xi_t \mid \mathscr{F}_t] = 0$. We now show that we may bound $\sum_{k=1}^{t}\xi_k - \mathbb{E}[\xi_k \mid \mathscr{F}_k]$ using [5, lemma 21]. By lemma B.5, lemma 4.5, and (4.20),

$$\begin{aligned}\xi_t - \mathbb{E}[\xi_t \mid \mathscr{F}_t] = \xi_t &\le \alpha_t\|\delta m_t\|\left(k_4\,\|v_t\| + \|w_t - w_*\|\right)\\ &\le \alpha_t\sqrt{6L^2\|w_t-w_*\|^{\gamma} + a_4}\,(k_4+1)\sqrt{\sum_{s=0}^{t-1}\alpha_s}\\ &\le (k_4+1)\left(6L^2\sqrt{\alpha_t^{1-\gamma}\left(\alpha_t\sum_{s=0}^{t-1}\alpha_s\right)^{\gamma+1}} + a_4\sqrt{\alpha_t^2\sum_{s=0}^{t-1}\alpha_s}\right)\\ &\le (k_4+1)\left(6L^2\sqrt{K_0^{1+\gamma}} + a_4\sqrt{K_0}\right).\end{aligned} \tag{4.24}$$

As for the conditional variance, we apply Cauchy-Schwarz, lemma B.5, and lemma 4.5 to get

$$\begin{aligned}\mathbb{E}[\xi_t^2 \mid \mathscr{F}_t] &\le 2\alpha_t^2\left(k_4^2\,\mathbb{E}[\langle\delta m_t,\, v_t\rangle^2 \mid \mathscr{F}_t] + \mathbb{E}[\langle \delta m_t,\, w_t-w_*\rangle^2 \mid \mathscr{F}_t]\right)\\ &\le 2\left(k_4^2+1\right)\alpha_t^2\,\mathbb{E}[\|\delta m_t\|^2 \mid \mathscr{F}_t]\left(\|v_t\|^2 + \|w_t-w_*\|^2\right)\\ &\le 2\left(k_4^2+1\right)\alpha_t^2\left(a_6\left(F(w_t)-F_*\right)+a_7\right)\left(\left(\sum_{s=0}^{t-1}\alpha_s\right) + \|w_t-w_*\|^2\right).\end{aligned}$$

Recall, by lemma 4.5, we get $\alpha_t\|w_t - w_*\|^2 \le \alpha_t\sum_{s=0}^{t-1}\alpha_s \le K_0$. Let $k_{11} = 2\left(k_4^2+1\right)$. There exists $c>0$ such that

$$\begin{aligned}\mathbb{E}[\xi_t^2 \mid \mathscr{F}_t] &\le 2k_{11}\,a_6\,K_0\,\alpha_t\left(F(w_t)-F_*\right) + k_{11}\,K_0\,a_7 + k_{11}\,a_7\,\alpha_t^2\,\|w_t-w_*\|^2\\ &\le c\left(1+\alpha_t\left(F(w_t)-F_*\right)+\alpha_t^2\,\|w_t-w_*\|^2\right).\end{aligned} \tag{4.25}$$

Let $b>0$ be the right hand side of (4.24). Applying [5, lemma 21.b], we get for a choice of tolerance level $\delta\in(0,1)$ and all $q>0$, the following holds with probability at least $1-\delta$:

$$\sum_{s=1}^{t-1}\xi_s \le \frac{(e^q-q-1)}{q}\cdot\frac{c}{b}\left(1+\sum_{s=1}^{t-1}\alpha_s\left(F(w_s)-F_*\right)+\sum_{s=1}^{t-1}\alpha_s^2\,\|w_s-w_*\|^2\right)+\frac{b}{q}\cdot\frac{\log(1/\delta)}{q}. \tag{4.26}$$

Now, we want to 'remove' the $\alpha_t\left(F(w_*)-F(w_t)\right)$ term in (4.21) so that the upper bound of $\Delta z_{t+1} + k_0^2\Delta v_{t+1}$ is of the order $O(1+\sum_{k=0}^{t-1}\alpha_k^2\|w_k-w_*\|^2)$. To do so, we choose $q>0$ such that

$$\frac{e^q-q-1}{q}\cdot\frac{c}{b} < 1 \tag{4.27}$$

The above inequality is solvable since $(e^q-q-1)/q \to 0$ as $q\to 0$. Moreover, q is independent of t and α_t since b and c are both independent of t and α_t. Therefore, all constants on the right hand side of (4.26) are independent of t and α_t. Since $F(w_t) > F_*$, the term $\alpha_t(F_* - F(w_t))$ in (4.21) dominates over that in (4.26). Using the estimate in (4.26) and summing (4.21) from 1 to $t-1$ yields the desired result. □

Remark 4.8. *Observe that the constant K_5 is independent of the step sizes. This can be seen from (4.27). The constants b and c on the right hand side of (4.24) and (4.25) respectively are independent of the step sizes α_t.*

Proposition 4.9. *With the notation in proposition 4.7, if all assumptions in lemma 4.7 holds and $\alpha_0^2 \leq \frac{1}{4K_5}$, then for any $\delta \in (0,1)$ there exists a constant $K_6(L,\gamma,\beta) > 0$ such that for any $T > 0$*

$$\mathbb{P}\left(\max_{1\leq t\leq T}\{\|w_t - w_*\|^2, \|z_t - w_*\|^2, \|v_t\|^2\} \leq K_6 \log\left(\frac{T}{\delta}\right)\right) \geq 1-\delta.$$

Proof. It suffices to prove the conclusion for $\|w_t - w_*\|^2$ since $\|z_t - w_*\|^2 \leq 2\|w_t - w_*\|^2 + 2k_0^2\|v_t\|^2$ and $v_t = (w_t - w_*) - (w_{t-1} - w_*)$ gives the following inequalities

$$\max_{1\leq t\leq T}\left(\|v_t\|^2, \|z_t - w_*\|^2\right) \leq 10\max(1,k_0^2)\max_{1\leq t\leq T}\|w_t - w_*\|^2.$$

As a consequence of proposition 4.7 and $\|w_t - w_*\|^2 \leq 2\|z_t - w_*\|^2 + 2k_0^2\|v_t\|^2$, we have

$$\mathbb{P}\left(\frac{\|w_t - w_*\|^2}{2} > K_3 + K_4\log\left(\frac{1}{\delta}\right) + K_5\sum_{s=1}^{t-1}\alpha_s^2\|w_s - w_*\|^2\right) < \delta\ ,\ \forall t \geq 1. \tag{4.28}$$

Let $t_1 = \min\{t \in [1,T] : K_5\sum_{s=t}^{T}\alpha_s^2 \leq \frac{1}{4}\}$. Note that t_1 is well-defined since $\alpha_T^2 \leq \alpha_0^2 \leq \frac{1}{4K_5}$. In particular, t_1 depends only on L,γ,β since K_5 depends only on those parameters. Using lemma 4.5, with probability greater than $1-\delta$, for any $t \geq 1$,

$$\begin{aligned}
\frac{\|w_t - w_*\|^2}{2} &\leq K_3 + K_4\ \log\left(\frac{T}{\delta}\right) + K_5\ \sum_{s=1}^{t_1}\alpha_s^2\|w_s - w_*\|^2 + K_5\ \sum_{s=t_1}^{T}\alpha_s^2\|w_s - w_*\|^2 \\
&\leq K_3 + K_4\ \log\left(\frac{T}{\delta}\right) + K_5\sum_{s=1}^{t_1}\alpha_s^2\left(\sum_{k=1}^{s}\alpha_k\right) + \frac{1}{4}\max_{t_1\leq t\leq T}\|w_t - w_*\|^2 \\
&\leq \left(K_3 + K_5\,K_0\sum_{s=1}^{t_1}\alpha_s\right) + K_4\ \log\left(\frac{T}{\delta}\right) + \frac{1}{4}\max_{1\leq t\leq T}\|w_t - w_*\|^2.
\end{aligned}$$

Since the right hand side is independent of t, we know $\frac{1}{2}\max_{1\leq t\leq T}\|w_t - w_*\|^2$ is also bounded by the right hand side. Solving the inequality for $\max_{1\leq t\leq T}\|w_t - w_*\|^2$ yields the desired result, where $K_6 = \max\{4K_3 + t_1K_0, 4K_4, 10k_0^2, 10\}$ for all $t > 0$. □

Proposition 4.10. *With the notation in lemma 4.5, if all of the assumptions in proposition 4.9 holds, then there exists a constant $K_7(L,\gamma,\beta,w_0) > 0$ such that for any $T > 0$ and $0 < \delta < \min(1, 2T/e)$,*

$$\mathbb{P}\left(\sum_{t=1}^{T}\alpha_t\left(F(w_t) - F_*\right) \leq K_7\left(\log\frac{2T}{\delta}\right)^{3/2}\right) \geq 1-\delta.$$

Proof. Using the notation in eq. (4.12), rearranging terms in (4.21) and using $F(w_*) = F_*$ gives us

$$\frac{2}{1-\beta}\,\alpha_t(F(w_t) - F_*) \leq 2\,k_4\left(\alpha_{t-1}F(w_{t-1}) - \alpha_tF(w_t)\right) - \Delta z_{t+1} - k_0^2\Delta v_{t+1} + \frac{2\,\xi_t}{1-\beta} + \varphi_t \tag{4.29}$$

where ξ_t and φ_t are defined in eq. (4.22). By lemma B.5, we have

$$\begin{aligned}\xi_t &= (1-\beta)k_4\,\alpha_t\langle\delta m_t,\, v_t\rangle + \,\alpha_t\langle\delta m_t,\, w_t - w_*\rangle \\ &\le (k_4+1)\;\alpha_t\,\|\delta m_t\|\;(\|v_t\| + \|w_t - w_*\|) \\ &\le (k_4+1)\,\alpha_t\,\sqrt{6L^2\|w_t-w_*\|^{2\gamma} + a_4}\,(\|v_t\| + \|w_t - w_*\|)\end{aligned} \tag{4.30}$$

Now, for a fixed $\delta \in (0,1)$ and $T \ge t$, define

$$c = 4\sqrt{6L^2+a_4}\,(k_4+1)\ \ ,\ \ \xi_t' = \xi_t \mathbb{1}_{\{\max(\|w_t-w_*\|^2, \|v_t\|^2) \le K_6\log(2T/\delta)\}}.$$

Since $\delta < 2T/e$, we have $\|w_t - w_*\|^{2\gamma} \le \log(2T/\delta) \le (\log(2T/\delta))^2$. Therefore, by (4.30), we get

$$\xi_t' \le 2\sqrt{2}\,(k_4+1)\sqrt{2(6L^2+a_4)\,\alpha_t^2\cdot\left(\log\frac{2T}{\delta}\right)^2}.$$

Applying [5, lemma 21.a] to ξ_t' yields

$$\sum_{t=1}^{T}\xi_t' \le \left(c\log\frac{2T}{\delta}\right)\left(2\log\frac{1}{\delta}\sum_{t=1}^{T}\alpha_t^2\right)^{\frac{1}{2}} \le c\sqrt{2}\left(\log\frac{2T}{\delta}\right)^{\frac{3}{2}}. \tag{4.31}$$

Let $K' > 0$ be the constant on the right hand side of (4.23). Applying proposition 4.9 and substituting (4.31) to (4.29) yields that the following occurs with probability at least $1-\delta$:

$$\sum_{t=1}^{T}\alpha_t(F(w_t) - F_*) \le \frac{1-\beta}{2}\left(\alpha_0 F(w_0) + \|z_1 - w_*\|^2 + \|v_1\|^2 + \frac{2}{1-\beta}\cdot c\sqrt{2}\left(\log\frac{2T}{\delta}\right)^{\frac{3}{2}} + K'\right).$$

Since $z_1 = w_0$, $v_1 = 0$, and $2T/\delta > e$, we take $K_7 = 4\max\left(\alpha_0\,F(w_0), \|w_0 - w_*\|^2, c\sqrt{2}, K'\right)$. □

Proposition 4.11. *With the notation in lemma 4.5, if all assumptions of proposition 4.9 holds, $\beta \in (0,1)$, and $\alpha_t = \Theta(t^{-p})$ where $p \in (\frac{1}{2}, 1)$, then there exists a constant $K_8(L,\beta,\gamma,w_0) > 0$ such that for any $\delta \in (0,1)$, $T > 1$, and $\tau \in [1,T]$,*

$$\mathbb{P}\left[\sum_{t=\tau}^{T}\|v_{t+1}\|^2 \le K_8\,\left(\log\frac{2T}{\delta}\right)^{1+\gamma}\max\left(\tau^{-p},\,\beta^{2\tau}\right)\right] \ge 1-\delta.$$

Proof. We have the recursion $\|v_{t+1}\|^2 = \beta^2\|v_t\|^2 - 2\beta\alpha_t\langle\nabla\ell(Z_t,w_t),\, v_t\rangle + \alpha_t^2\|\nabla\ell(Z_t,w_t)\|^2$. Such a difference equation can be rewritten into an integral equation by proposition C.1:

$$\begin{cases}\|v_{t+1}\|^2 = \sum_{s=0}^{t-1}\beta^{2s}X_{t-s}. \\ X_t = -2\beta\alpha_t\langle\nabla\ell(Z_t,w_t),\, v_t\rangle + \alpha_t^2\|\nabla\ell(Z_t,w_t)\|^2.\end{cases}$$

It follows that

$$\begin{aligned}\sum_{t=\tau}^{T}\|v_{t+1}\|^2 &= \sum_{t=\tau}^{T}\sum_{s=1}^{t}\beta^{2(s-1)}X_{t-s+1} \\ &= \left(\sum_{s=1}^{\tau}\sum_{t=\tau}^{T} + \sum_{s=\tau}^{T}\sum_{t=s}^{T}\right)\beta^{2(s-1)}X_{t-s+1} \\ &= \sum_{s=1}^{\tau}\beta^{2(s-1)}\left(\sum_{t=\tau}^{T}X_{t-s+1}\right) + \beta^{2(\tau-1)}\sum_{s=0}^{T-\tau}\beta^{2s}\left(\sum_{t=1}^{T-\tau-s+1}X_t\right).\end{aligned} \tag{4.32}$$

We estimate the parentheses in the first term. By the decomposition $\nabla\ell(Z_t, w_t) = \nabla\ell(Z_t, w_*) + (\nabla\ell(Z_t, w_t) - \nabla\ell(Z_t, w_*))$, assumption 2.3, and lemma B.5, we get $\|\nabla\ell(Z_t, w_t)\|^2 \le 6L^2\|w_t - w_*\|^{2\gamma} + a_4$. Together with the convexity of F, we get

$$\begin{aligned} X_t &= -2\beta\,\alpha_t\,\langle\nabla F(w_t),\, v_t\rangle + 2\,\alpha_t\,\beta\langle\delta m_t,\, v_t\rangle + \alpha_t^2\,\|\nabla\ell(Z_t, w_t)\|^2 \\ &\le 2\beta\alpha_t\,(F(w_{t-1}) - F(w_t) + \langle\delta m_t,\, v_t\rangle) + \alpha_t^2\,(6L^2\|w_t - w_*\|^{2\gamma} + a_4) \\ &\le 2\beta(\alpha_{t-1}F(w_{t-1}) - \alpha_t F(w_t)) + 2\beta\alpha_t\langle\delta m_t,\, v_t\rangle + \alpha_t^2\,(6L^2\|w_t - w_*\|^{2\gamma} + a_4). \end{aligned} \tag{4.33}$$

By proposition 4.9, the event $\mathcal{E} = \{\max_{1\le t\le T}\{\|w_t - w_*\|^2, \|v_t\|^2\} \le K_6\log(2T/\delta)\}$ occurs with probability at least $1 - \frac{\delta}{2}$. Now, define $\psi_t = \alpha_t\langle\delta m_t,\, v_t\rangle\mathbb{1}_{\mathcal{E}}$. By lemma B.5, we know that $\|\delta m_t\| \le \sqrt{6L^2\|w_t - w_*\|^{2\gamma} + a_4}$. Therefore, we have

$$\psi_t \le \alpha_t\cdot\|\delta m_t\|\cdot\|v_t\| \le \sqrt{6L^2 + a_4}\,\alpha_t\left(\log\frac{2T}{\delta}\right)^{(1+\gamma)/2}. \tag{4.34}$$

By [5, lemma 21.a], there exists an event $\mathcal{E}'$ such that $\mathbb{P}(\mathcal{E}') \ge 1 - \frac{\delta}{2}$ where the following occurs

$$\sum_{t=\tau}^{T}\psi_{t-s+1} \lesssim \left(\log\frac{2T}{\delta}\right)^{1+\gamma}\left(\sum_{t=\tau}^{T}\alpha_{t-s+1}^2\right)^{1/2}. \tag{4.35}$$

Now, under the event $\mathcal{E}\cap\mathcal{E}'$ where $\mathbb{P}(\mathcal{E}\cap\mathcal{E}') \ge 1 - \delta$, the sum of X_{t-s} in eq. (4.33) is

$$\begin{aligned} \sum_{s\le t=\tau}^{T} X_{t-s+1} &\lesssim \alpha_{\tau-s}F(w_{\tau-s}) + \left(\log\frac{2T}{\delta}\right)^{1+\gamma}\sqrt{\sum_{s\le t=\tau}^{T}\alpha_{t-s+1}^2} + \left(\log\frac{2T}{\delta}\right)^{\gamma}\sum_{s\le t=\tau}^{T}\alpha_{t-s+1}^2 \\ &\lesssim \left(\log\frac{2T}{\delta}\right)^{1+\gamma}\alpha_{\tau-s}\left(1 + \sqrt{\sum_{s\le t=1}^{T-\tau+1}\frac{\alpha_{\tau-s+t}^2}{\alpha_{\tau-s}^2}} + \sum_{t=1}^{T-\tau+1}\frac{\alpha_{\tau-s+t}^2}{\alpha_{\tau-s}^2}\right). \end{aligned} \tag{4.36}$$

Since $F(w_t) = F(w_t) - F_* + F_*$, by lemma B.3 and proposition 4.9, under the event $\mathcal{E}$, we have $F(w_t) \lesssim F_* + \|w_t - w_*\|^{1+\gamma} = O(\log^{\frac{1+\gamma}{2}}(2T/\delta))$. Also, since $2T/\delta > e$, the coefficients are $O(\log^{1+\gamma}(2T/\delta))$. If $\alpha_t = \Theta(t^{-p})$ and $p \in (\frac{1}{2}, 1)$, then the ratio of step sizes on the right hand side are of the order t^{-2p} which makes the summation terms in (4.36) summable as $T\uparrow\infty$. Therefore, under the event $\mathcal{E}\cap\mathcal{E}'$,

$$\sum_{s=1}^{\tau}\sum_{t=\tau}^{T}\beta^{2(s-1)}X_{t-s+1} \lesssim \left(\log\frac{2T}{\delta}\right)^{1+\gamma}\tau^{-p}\sum_{s=1}^{\tau}\beta^{2(s-1)}\left(\frac{\tau}{\tau - s}\right)^{p}.$$

Now, the ratio on the right hand side converges to 1 as $\tau\uparrow\infty$. As a consequence, the sum $\sum_{s=1}^{\tau}\beta^{2(s-1)}(\tau/(\tau - s))^p$ converges as $\tau\uparrow\infty$ by the dominated convergence theorem.

Now, examine the second term on the right hand side of (4.32). By lemma B.5, we have $\alpha_t^2\,\|\nabla\ell(Z_t, w_t)\|^2 \le \alpha_t^2\,(a_1\ell(Z_t, w_t) + a_2)$. By lemma 4.5, we get $\sum\alpha_t^2\,\ell(Z_t, w_t) < K_2$. Therefore, using the expression for X_t in (4.33) and the estimate (4.35), under the event $\mathcal{E}\cap\mathcal{E}'$,

$$\sum_{t=1}^{T-\tau-s+1} X_t \lesssim F(w_0) + \left(\log\frac{2T}{\delta}\right)^{1+\gamma}\sqrt{\sum_{t=1}^{\infty}\alpha_t^2} + \left(a_1\,K_2 + a_2\sum_{t=0}^{\infty}\alpha_t^2\right).$$

Therefore, the claim is proved. □

Lemma 4.12. *With the notation as in lemma 4.5, if all assumptions in proposition 4.9 holds, then there exists a constant $K_9(L, \gamma, \beta) > 0$ such that for any $\delta \in (0, 1)$, $T > 1$, and $\tau \in [0, T]$,*

$$\mathbb{P}\left[\sum_{t=\tau}^{T}\langle\nabla F(w_t),\, -\alpha_t\nabla\ell(Z_t, w_t)\rangle \le K_9\cdot\alpha_\tau\left(\log\frac{2T}{\delta}\right)^{1+\gamma}\right] \ge 1 - \delta.$$

Proof. By proposition 4.9, there exists an event $\mathcal{E} \in \mathscr{F}_T$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\delta}{2}$ such that the following inequality holds

$$\max_{1 \leq t \leq T} \{\|w_t - w_*\|^2, \|v_t\|^2\} \leq K_6 \log \frac{2T}{\delta} \tag{4.37}$$

We may decompose $-\alpha_t \langle \nabla F(w_t), \nabla \ell(Z_t, w_t) \rangle = -\alpha_t \|\nabla F(w_t)\|^2 + \psi_t$ where we define

$$\psi'_t = \psi_t \mathbb{1}_{\mathcal{E}} \ , \ \psi_t = \alpha_t \langle \nabla F(w_t), \delta m_t \rangle. \tag{4.38}$$

By Cauchy-Schwarz, γ-Hölder property of ∇F, lemma B.5, and (4.37), for all $t \in [\tau, T]$,

$$\begin{aligned} |\psi'_t| &\leq \mathbb{1}_{\mathcal{E}} \cdot \alpha_t \cdot \|\nabla F(w_t)\| \cdot \|\delta m_t\| \\ &\leq L \mathbb{1}_{\mathcal{E}} \cdot \alpha_t \sqrt{6L^2 \|w_t - w_*\|^{4\gamma} + a_4 \|w_t - w_*\|^{2\gamma}} \\ &\leq L \sqrt{K_6^\gamma (6L^2 K_6^\gamma + a_4)} \cdot \alpha_\tau \cdot \left(\log \frac{2T}{\delta} \right)^\gamma . \end{aligned} \tag{4.39}$$

We apply Cauchy-Schwarz, lemma B.5, estimate (4.37), and set $\gamma = 1$

$$\begin{aligned} \mathbb{E}[(\psi'_t)^2 \mid \mathscr{F}_t] &\leq \alpha_t^2 \cdot \|\nabla F(w_t)\|^2 \cdot \mathbb{E}[\|\delta m_t\|^2 \mid \mathscr{F}_t] \\ &\leq \alpha_\tau \cdot \alpha_t \, \|\nabla F(w_t)\|^2 \cdot (6L^2 \|w_t - w_*\|^{2\gamma} + a_4) \\ &\leq \alpha_\tau \cdot \alpha_t \, \|\nabla F(w_t)\|^2 \cdot (6L^2 + a_4) \cdot \left(\log \frac{2T}{\delta} \right)^\gamma . \end{aligned} \tag{4.40}$$

For shorthand notation, denote by $b = L\sqrt{K_6^\gamma (6L^2 K_6 + a_4)}$ and $c = 6L^2 + a_4$ as the constants appearing on the right hand side of (4.39) and (4.40). By [5, lemma 21.b], for any $q > 0$ and $\delta \in (0, 1)$, there exists an event $\mathcal{E}' \in \mathscr{F}_T$ with $\mathbb{P}(\mathcal{E}') \geq 1 - \frac{\delta}{2}$ where $\mathcal{E}'$ is the event

$$\sum_{t=\tau}^{T} \psi'_t \leq \frac{e^q - q - 1}{q} \cdot \frac{c}{b} \left(\sum_{t=\tau}^{T} \alpha_t \, \|\nabla F(w_t)\|^2 \right) + \alpha_\tau \cdot \frac{b}{q} \left(\log \frac{2T}{\delta} \right)^{1+\gamma} . \tag{4.41}$$

Since $(e^q - q - 1)/q \to 0$ as $q \to 0$, may choose $q > 0$ such that

$$\frac{e^q - q - 1}{q} \cdot \frac{c}{b} < 1.$$

Under the event $\mathcal{E} \cap \mathcal{E}'$, we have $\psi'_t = \psi_t$ and the following occurs with probability $\mathbb{P}(\mathcal{E} \cap \mathcal{E}') \geq 1 - \delta$

$$\begin{aligned} \sum_{t=\tau}^{T} \langle \nabla F(w_t), \, -\alpha_t \nabla \ell(Z_t, w_t) \rangle &= \sum_{t=\tau}^{T} \left(-\alpha_t \, \|\nabla F(w_t)\|^2 + \psi'_t \right) \\ &\leq -\left(\left(1 - \frac{e^q - q - 1}{q} \cdot \frac{c}{b} \right) \sum_{t=\tau}^{T} \alpha_t \, \|\nabla F(w_t)\|^2 \right) + \frac{b}{q} \cdot \alpha_\tau \left(\log \frac{2T}{\delta} \right)^{1+\gamma} \\ &\leq \frac{b}{q} \cdot \alpha_\tau \left(\log \frac{2T}{\delta} \right)^{1+\gamma} . \end{aligned}$$

The proof is complete. □

Proposition 4.13. *With the notation in lemma 4.5, if all the assumptions in proposition 4.9 holds and $\beta \in (0, 1)$, $\gamma = 1$, $\alpha_t = \Theta(t^{-p})$ where $p \in (\frac{1}{2}, 1)$, then there exists a constant $K_{10}(L, \beta, \gamma, w_0) > 0$ such that for any $\delta \in (0, 1)$, $T > 1$, and $\tau \in (1, T]$,*

$$\mathbb{P} \left[\sum_{t=\tau}^{T} \langle \nabla F(w_t), \, v_t \rangle \leq K_{10} \, \left(\log \frac{2T}{\delta} \right)^2 \max \left(\tau^{-p}, \, \beta^{2\tau} \right) \right] \geq 1 - \delta.$$

Proof. Fix $T > 1$ and $\tau > 1$. Let $\Lambda_t = \langle \nabla F(w_t),\, -\alpha_t \nabla \ell(Z_t, w_t)\rangle$. By lemma 4.12 and proposition 4.11, there exists an event $\mathcal{E} \in \mathscr{F}$ such that with probability $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, both of the following occurs:

$$\sum_{t=\tau}^{T} \|v_{t+1}\|^2 \leq K_8 \cdot \max(\tau^{-p}, \beta^{2\tau}) \left(\log \frac{2T}{\delta}\right)^{1+\gamma} \quad, \ \forall\ \tau \in [1, T]. \tag{4.42}$$

$$\sum_{t=t_0}^{T} \Lambda_t \leq K_9 \cdot \alpha_{t_0} \left(\log \frac{2T}{\delta}\right)^{1+\gamma} \quad, \ \forall\ t_0 \in [0, T]. \tag{4.43}$$

Since $\gamma = 1$, i.e. ∇F is Lipschitz, $\langle \nabla F(w_t),\, v_t\rangle$ satisfies the following inequality for any $t > 1$,

$$\begin{aligned}
\langle \nabla F(w_t),\, v_t\rangle &= \langle \nabla F(w_t) - \nabla F(w_{t-1}) + \nabla F(w_{t-1}),\, v_t\rangle \\
&= \langle \nabla F(w_{t-1}),\, \beta v_{t-1} - \alpha_{t-1}\nabla \ell(Z_{t-1}, w_{t-1})\rangle + \langle \nabla F(w_t) - \nabla F(w_{t-1}),\, v_t\rangle \\
&= \beta \langle \nabla F(w_{t-1}),\, v_{t-1}\rangle + \Lambda_{t-1} + \langle \nabla F(w_t) - \nabla F(w_{t-1}),\, v_t\rangle \\
&\leq \beta\, \langle \nabla F(w_{t-1}),\, v_{t-1}\rangle + \Lambda_{t-1} + \|v_t\|^2.
\end{aligned}$$

Also, since $v_1 = 0$, we have $\langle \nabla F(w_1),\, v_1\rangle = 0$. Using proposition C.1 and switching the order of integration, under the event $\mathcal{E}$, the following occurs with probability $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ for all $\tau > 1$:

$$\begin{aligned}
\sum_{t=\tau}^{T} \langle \nabla F(w_t),\, v_t\rangle &\leq \sum_{s=0}^{\tau-2} \beta^s \left(\sum_{t=\tau}^{T} \Lambda_{t-s-1} + \|v_{t-s}\|^2\right) + \sum_{s=\tau-2}^{T-2} \beta^s \left(\sum_{t=s+2}^{T} \Lambda_{t-s-1} + \|v_{t-s}\|^2\right) \\
&= \sum_{s=0}^{\tau-2} \beta^s \left(\sum_{t=\tau}^{T} \Lambda_{t-s-1} + \|v_{t-s}\|^2\right) + \beta^\tau \sum_{s=0}^{T-\tau} \beta^{s-2} \left(\sum_{t=\tau}^{T} \Lambda_{t-\tau+1} + \|v_{t-\tau+2}\|^2\right) \\
&\leq 2\max(K_8, K_9) \left(\log \frac{2T}{\delta}\right)^{1+\gamma} \left(\sum_{s=0}^{\tau-2} \beta^s \cdot \alpha_{\tau-s-1} + \beta^\tau \sum_{s=0}^{\infty} \beta^{s-2}\right) \\
&\lesssim 2\max(K_8, K_9) \cdot \max(\tau^{-p}, \beta^\tau) \cdot \left(\log \frac{2T}{\delta}\right)^{1+\gamma} \left(\sum_{s=0}^{\tau-2} \beta^s \left(\frac{\tau}{\tau - s - 1}\right)^p + \frac{1}{\beta^2(1-\beta)}\right).
\end{aligned}$$

The summation on the right hand side remains bounded as $\tau \uparrow \infty$ by the dominated convergence theorem. The proof is complete. □

Theorem 4.14. *If all assumptions in proposition 4.9 and proposition 4.11 both holds, and $\gamma = 1$, then there exists a constant $K_{11}(L, \beta, \gamma, w_0) > 0$ such that for all $T > 1$, $T_0 \in (1, T]$, and any $\delta \in (0, 1)$*

$$\mathbb{P}\left(F(w_{T+1}) - F_* \leq K_{11} \left(\log \frac{2T}{\delta}\right)^2 \max\left\{\frac{1}{\sum_{t=T_0}^{T} \alpha_t}\,,\ T_0^{-p}\,,\ \beta^{T_0}\right\}\right) \geq 1 - \delta.$$

Proof. Let $\Delta_t F = F(w_t) - F_*$. By proposition 4.10, proposition 4.11, lemma 4.12, and proposition 4.13, there exists an event $\mathcal{E} \in \mathscr{F}_T$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ such that all of the following holds

$$\sum_{t=1}^{T} \alpha_t\, \Delta_t F \leq K_7 \left(\log \frac{2T}{\delta}\right)^{\frac{3}{2}}. \tag{4.44}$$

$$\sum_{t=\tau}^{T} \langle \nabla F(w_t),\, -\alpha_t \nabla \ell(Z_t, w_t)\rangle \leq K_9 \cdot \alpha_\tau \left(\log \frac{2T}{\delta}\right)^{1+\gamma}. \tag{4.45}$$

$$\sum_{t=\tau}^{T} \left\{\|v_{t+1}\|^2 + \langle \nabla F(w_t),\, v_t\rangle\right\} \leq (K_8 + K_{10}) \left(\log \frac{2T}{\delta}\right)^2 \max\left(\tau^{-p}, \beta^\tau\right). \tag{4.46}$$

If we fix $T > 1$ and $T_0 \in (1, T]$, then under the event $\mathcal{E}$, there exists $\tau \in [T_0, T]$ such that

$$\Delta_\tau F \le \sum_{t=T_0}^{T} \frac{\alpha_t \, \Delta_t F}{\sum_{t=T_0}^{T} \alpha_t} \le \frac{K_7 \left(\log \frac{2T}{\delta}\right)^{\frac{3}{2}}}{\sum_{t=T_0}^{T} \alpha_t}$$

where the middle term is the weighted average of $\Delta_t F$ from $\tilde{t}$ to T. Using (γ, L)-smoothness of F along with estimates (4.44), (4.45), and (4.46), on the event $\mathcal{E}_1$, we have

$$\begin{aligned}\Delta_{T+1} F &\le \Delta_\tau F + \sum_{t=\tau}^{T} \langle \nabla F(w_t), \, w_{t+1} - w_t \rangle + \frac{L}{1+\gamma} \sum_{t=\tau}^{T} \|v_{t+1}\|^{1+\gamma} \\ &= \Delta_\tau F - \sum_{t=\tau}^{T} \left\{ \alpha_t \, \|\nabla F(w_t)\|^2 - \psi_t \right\} + \beta \sum_{t=\tau}^{T} \langle \nabla F(w_t), \, v_t \rangle + \frac{L}{2} \sum_{t=\tau}^{T} \|v_{t+1}\|^2. \\ &\lesssim \max\left(K_7, K_9, \frac{L+2\beta}{2}(K_8 + K_{10})\right) \left(\log \frac{2T}{\delta}\right)^2 \max\left\{ \frac{1}{\sum_{t=T_0}^{T} \alpha_t}, \alpha_\tau, \max(\tau^{-p}, \beta^\tau) \right\}.\end{aligned}$$

The proof is complete. □

Remark 4.15. *Theorem 2.6 follows from setting $T_0 = \lfloor T/2 \rfloor$. Observe that $\sum_{t=\lfloor T/2 \rfloor}^{T} \alpha_t \ge \alpha_T \lfloor \frac{T}{2} \rfloor$. In addition, if $T^{-p} \ge T^{p-1}$, then necessarily $p \le 1/2$. Therefore, setting $T_0 = \lfloor T/2 \rfloor$ in the conclusion of theorem 4.14 yields*

$$\max\left(T^{-p}, T^{p-1}, \beta^T\right) = O(T^{p-1}) \ , \ p \in \left(\frac{1}{2}, 1\right).$$

Therefore, theorem 2.6 follows.

Appendix A. Proof of Proposition 2.4

Proof. Since X_t is non-negative, we may drop X_t from the assumed inequality on Y_t and obtain $Y_t - Y_{t-1} \le a_{t-1} Y_{t-1} - X_{t-1} + Z_{t-1} \le a_{t-1} Y_{t-1} + Z_{t-1}$. We may sum up both sides to obtain the recursive relation

$$Y_t \le Y_0 + \sum_{s=1}^{t} a_{s-1} Y_{s-1} + \sum_{s=1}^{t} Z_{s-1}. \tag{A.1}$$

Since $Z_t \in \ell^1(\mathbb{N})$, the above is a Gronwall-type inequality. By proposition C.2, we get

$$Y_t \lesssim (Y_0 + \|Z\|_{\ell^1}) \prod_{s=1}^{\infty} (1 + a_{s-1}).$$

Since $a_t \in \ell^1$, the right hand side converges. Therefore, the first and second claim are proved. To prove the second claim, repeating the same argument to produce (A.1), but without dropping X_t, we get

$$\sum_{s=1}^{t} X_{s-1} \le Y_t + \sum_{s=1}^{t} X_{s-1} \le Y_0 + \sum_{s=1}^{t} a_{s-1} Y_{s-1} + \sum_{s=1}^{t} Z_{s-1}.$$

By the first claim, Y_t is uniformly bounded. Therefore, the right hand side is finite as $t \to \infty$. The proof is complete. □

Appendix B. Some Facts Used in Proofs

Proposition B.1. *If the first two properties in assumption 2.1 are both satisfied, then F is convex and (γ, L)-smooth.*

Proof. By the convexity of $\ell(z,\cdot)$, we get $\ell(z, a\,w_1+(1-a)w_2) \le a\,\ell(z,w_1)+(1-a)\ell(z,w_2)$. Applying the expectation w.r.t ρ proves that F is convex. Now, observe that we have

$$\begin{aligned}\|\nabla F(w_1) - \nabla F(w_2)\| &= \|\mathbb{E}_\rho[\nabla \ell(Z,w_1) - \nabla \ell(Z,w_2)]\| \\ &\le \mathbb{E}_\rho \|\nabla \ell(Z,w_1) - \nabla \ell(Z,w_2)\| \\ &\le L\,\|w_1 - w_2\|^\gamma.\end{aligned}$$

The proof is complete. □

Proposition B.2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is (γ, L)-smooth, then eq.* (2.1) *holds.*

Proof. Let $\phi(t) = f(x+t(y-x))$. Note that ϕ is continuously differentiable. Using the fundamental theorem of calculus on ϕ yields

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x),\, y-x\rangle \, dt \\ &\le f(x) + \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|\,\|y-x\|\,dt \\ &\le f(x) + L\,\|y-x\|^{1+\gamma}\int_0^1 t^\gamma\,dt.\end{aligned}$$

Integrating the right hand side yields the desired result. □

Lemma B.3. *(Lei, Shi, Guo* [5, lemma 13]*). If $f : \mathbb{R}^d \to \mathbb{R}$ is (γ, L)-smooth and convex, then*

$$\frac{\gamma\,\|\nabla f(y) - \nabla f(x)\|^{\frac{1+\gamma}{\gamma}}}{(1+\gamma)L^{1/\gamma}} \le f(y) - f(x) - \langle \nabla f(x),\, y-x\rangle \le \frac{L\,\|y-x\|^{1+\gamma}}{1+\gamma}.$$

Lemma B.4. *(Lei, Shi, Guo* [5, lemma 14]*). If assumption 2.1 holds and $\ell(z,\cdot)$ is convex for all $z \in \mathcal{Z}$, then for all $\theta \in (0,1]$,*

$$\mathbb{E}_\rho[\|\nabla \ell(Z,w)\|^{1+\theta}] \le 2^\theta L^{1/\gamma}(1+\theta)\,[F(w) - F_*] + \frac{2^\theta(1-\gamma\theta)}{1+\gamma} + 2^\theta\,\mathbb{E}_\rho[\|\nabla \ell(Z,w_*)\|^{1+\theta}].$$

Lemma B.5. *Assume assumption 2.1 holds and $\ell(z,\cdot)$ is convex for all $z \in \mathcal{Z}$.*

(1) There exists $a_0, a_1 \in \mathbb{R}_+$ such that $\|\nabla F(w)\|^2 \le a_1\,F(w) + a_0$ where a_1 is in (2a).

(2) If in addition assumption 2.3 holds, then there exists $a_1, a_2, a_3, a_4, a_5, a_6, a_7 \in \mathbb{R}_+$ such that

- *(a) $\|\nabla \ell(z,w)\|^2 \le a_1\,\ell(z,w) + a_2$.*
- *(b) $\mathbb{E}_\rho[\|\nabla \ell(Z,w)\|^2] \le a_3\,(F(w) - F_*) + a_1\,F_* + a_2$.*
- *(c) $\|\nabla F(w) - \nabla \ell(z,w)\|^2 \le 6L^2\,\|w - w_*\|^{2\gamma} + a_4$.*
- *(d) $\|\nabla F(w) - \nabla \ell(z,w)\|^2 \le 2a_1\,(F(w) - F_*) + 2a_1\,\ell(z,w) + a_5$.*
- *(e) $\mathbb{E}[\|\nabla F(w_t) - \nabla \ell(Z,w_t)\|^2 \,|\, \mathscr{F}_t] \le a_6\,(F(w_t) - F_*) + a_7$.*

Proof. **Proof of (1).** We use the same argument as (B.1), but replace $\ell(z, w_*)$ with F_*.

Proof of (2.a). Applying lemma B.3 and subsequently Young's inequality [5, eq 4.3] with $p = (1+\gamma)/(1-\gamma)$ and $q = (1+\gamma)/(2\gamma)$ for $\gamma \in (0,1]$ yields

$$\begin{aligned}
\|\nabla \ell(z,w)\|^2 &\le \left(\frac{1+\gamma}{\gamma}\right)^{\frac{2\gamma}{1+\gamma}} L^{\frac{2}{1+\gamma}} \left(\ell(z,w) - \ell(z,w_*)\right)^{\frac{2\gamma}{1+\gamma}} \\
&\le \frac{1-\gamma}{1+\gamma}\left(\frac{1+\gamma}{\gamma}\cdot L^{1/\gamma}\right)^{\frac{2\gamma}{1-\gamma}} + \frac{2\gamma}{1+\gamma}\left(\ell(z,w) - \ell(z,w_*)\right) \qquad \text{(B.1)} \\
&\le \frac{2\gamma}{1+\gamma}\,\ell(z,w) + \frac{1-\gamma}{1+\gamma}\left(\frac{1+\gamma}{\gamma}\cdot L^{1/\gamma}\right)^{\frac{2\gamma}{1-\gamma}} + \frac{2\gamma}{1+\gamma}\cdot \sup_{z\in\mathcal{Z}} \ell(z,w_*).
\end{aligned}$$

Proof of (2.b). If the constant a_2 is the sum of the second and third term on the right hand side of (B.1), then applying the expectation $\mathbb{E}_\rho$ on both sides yields

$$\mathbb{E}_\rho \|\nabla \ell(Z,w)\|^2 \le \frac{2\gamma}{1+\gamma}\left(F(w) - F_*\right) + \frac{2\gamma}{1+\gamma} F_* + a_2.$$

Proof of (2.c). Observe that if assumption 2.3 and a minimizer $w_* \in \mathcal{W}$ exists and $\ell(z,\cdot)$ is convex in the second slot, then applying (2a) yields

$$\begin{aligned}
\|\nabla F(w) - \nabla \ell(z,w)\|^2 &\le 2\,\|\nabla F(w)\|^2 + 2\,\|\nabla \ell(z,w)\|^2 \\
&\le 2L^2\,\|w_t - w_*\|^{2\gamma} + 4\,\|\nabla \ell(z,w_*)\|^2 + 4\|\nabla \ell(z,w) - \nabla \ell(z,w_*)\|^2 \\
&\le 6L^2 \|w_t - w_*\|^{2\gamma} + 4\,a_1 \sup_{z\in\mathcal{Z}} \ell(z,w_*) + 4\,a_2.
\end{aligned}$$

Proof of (2.d). Another way of estimating $\|\nabla F(w) - \nabla \ell(z,w)\|^2$ is by using (1)

$$\begin{aligned}
\|\nabla F(w) - \nabla \ell(z,w)\|^2 &\le 2\|\nabla F(w_t)\|^2 + 2\|\nabla \ell(Z_t,w_t)\|^2 \\
&\le 2a_1\left(F(w) - F_*\right) + 2a_1 F_* + 2a_0 + 2a_1\,\ell(z,w) + 2a_2.
\end{aligned}$$

Proof of (2.e). Applying $\mathbb{E}[\cdot \,|\, \mathscr{F}_t]$ to (2d) yields

$$\begin{aligned}
\mathbb{E}[\|\delta m_t\|^2 \,|\, \mathscr{F}_t] &\le 2a_1\left(F(w_t) - F_*\right) + 2a_1\, F(w_t) + a_5 \\
&= 4a_1\left(F(w_t) - F_*\right) + (a_5 + 2a_1\, F_*).
\end{aligned}$$

This completes the proof. □

Proposition B.6. *Let δm_t be as in eq.* (1.2) *corresponding to the SHB iterates w_t in eq.* (1.1)*.*

(1) $\mathbb{E}[\delta m_t \,|\, \mathscr{F}_s] = \mathbf{0}$ and $\mathbb{E}[\delta m_s^{(i)}\, \delta m_t^{(j)}] = 0$ for all $s < t$ and $i \neq j$.

(2) $\mathbb{E}[\langle \nabla F(w_t),\, \delta m_t\rangle \,|\, \mathscr{F}_t] = 0$ for all $t > 0$.

Proof. By eq. (2.3), we have $\mathbb{E}[\delta m_t \,|\, \mathscr{F}_t] = \nabla F(w_t) - \mathbb{E}[\nabla \ell(Z,w_t) \,|\, \mathscr{F}_t] = 0$. Furthermore,

$$\mathbb{E}[\delta m_t \,|\, \mathscr{F}_{t-1}] = \mathbb{E}[\nabla F(w_t) \,|\, \mathscr{F}_{t-1}] - \mathbb{E}[\nabla \mathbb{E}_\rho[\ell(Z,w_t)] \,|\, \mathscr{F}_{t-1}] = 0.$$

$$\mathbb{E}[\delta m_s^{(i)}\, \delta m_t^{(j)}] = \mathbb{E}[\delta m_s^{(i)}\, \mathbb{E}[\delta m_t^{(j)} \,|\, \mathscr{F}_{t-1} \,|\, \mathscr{F}_s]] = 0\ ,\ s < t.$$

Since the standard inner product on $\mathbb{R}^d$ is a finite sum and the conditional expectation is linear, applying the first claim yields

$$\mathbb{E}[\langle \nabla F(w_t),\, \delta m_t\rangle \,|\, \mathscr{F}_t] = \sum_{i=1}^d \mathbb{E}[\nabla F(w_t)^{(i)} \delta m_t^{(i)} \,|\, \mathscr{F}_t] = \langle \nabla F(w_t),\, \mathbb{E}[\delta m_t \,|\, \mathscr{F}_t]\rangle = 0.$$

This completes the proof. □

Appendix C. Gronwall-type Inequalities

Proposition C.1. *Let $\{X_n \in \mathbb{R} : n \in \mathbb{N}_0\}$ and $\{Y_n \in \mathbb{R} : n \in \mathbb{N}_0\}$ be sequences, $n_0 \in \mathbb{N}_0$, and $\beta > 0$. The sequence*

$$Y_{n+n_0} = \beta^{2n} y + \mathbb{1}_{\mathbb{N}}(n) \sum_{k=0}^{n-1} \beta^{2k} X_{n_0+n-1-k} \ , \ n \geq 0$$

is the unique solution to the difference equation

$$\begin{cases} Y_{n_0+n} = \beta^2\, Y_{n_0+n-1} + X_{n_0+n-1} \ , \ n > 0. \\ Y_{n_0} = y. \end{cases} \tag{C.1}$$

If Z_n satisfies $Z_{n_0+n} \leq \beta^2\, Z_{n_0+n-1} + X_{n_0+n-1}$ (resp. $\geq$) and $Z_{n_0} \leq y$ (resp. $\geq$), then $Z_{n_0+n} \leq Y_{n_0+n}$ (resp. $\geq$).

Proof. We have $Y_{n_0} = \beta^0 y = y$ and $Y_{n_0+1} = \beta^2 y + X_{n_0} = \beta^2 Y_{n_0} + X_{n_0}$. Moreover, for $n > 0$,

$$\begin{aligned} X_{n_0+n} + \beta^2 Y_{n_0+n} &= X_{n_0+n} + \beta^{2(n+1)} y + \sum_{k=0}^{n-1} \beta^{2k+2} X_{n_0+n-1-k} \\ &= X_{n_0+n} + \beta^{2(n+1)} y + \sum_{k=1}^{n} \beta^{2k} X_{n_0+n-k} \\ &= Y_{n_0+n+1}. \end{aligned}$$

For uniqueness, if Z_{n_0+n} satisfies (C.1), then $Z_{n_0+n} - Y_{n_0+n} = \beta^2(Z_{n_0+n-1} - Y_{n_0+n-1})$ for all $n > 0$ with $Z_{n_0} - Y_{n_0} = 0$. This implies $Z_{n_0+n} - Y_{n_0+n} = (Z_{n_0} - Y_{n_0})\, \beta^{2n} = 0$ for all $n \geq 0$.

For the second part, we show by induction that $Z_{n_0+n} \leq Y_{n_0+n}$. Note that $Z_{n_0} \leq y = Y_{n_0}$ and $Z_{n_0+1} \leq \beta^2 y + X_{n_0} = Y_{n_0+1}$. For the inductive step, assume $Z_{n_0+n} \leq Y_{n_0+n}$ for $n \leq N$. From the first claim, $Y_{n_0+n+1} = \beta^2 Y_{n_0+n} + X_{n_0+n}$. Therefore, we have

$$Z_{n_0+N+1} \leq \beta^2\, Z_{n_0+N} + X_{n_0+N} \leq \beta^2\, Y_{n_0+N} + X_{n_0+N} = Y_{n_0+N+1}.$$

On the other hand, if $Z_{n_0+n} \geq \beta^2 Z_{n_0+n-1} + X_{n_0+n}$, then we may consider

$$\begin{cases} -Z_{n_0+n} \leq \beta^2(-Z_{n_0+n-1}) + (-X_{n_0+n-1}) \\ -Z_{n_0} \leq -y. \end{cases}$$

We can apply the previous claim and get $-Z_{n_0+n} \leq -Y_{n_0+n}$. □

Proposition C.2. *Let $A \geq 0$, $X_n \geq 0$, $c_n \geq 0$.*

$$X_n \leq A + \sum_{k=1}^{n} c_{k-1}\, X_{k-1} \ , \ X_0 \leq A \quad \Longrightarrow \quad X_n \leq A \prod_{k=1}^{n} (1 + c_{k-1})\,.$$

Proof. Define the sequence

$$Y_n = A + \sum_{k=1}^{n} c_{k-1}\, X_{k-1} \ , \ Y_0 = A$$

Observe that $Y_n - Y_{n-1} = c_{n-1} X_{n-1} \leq c_{n-1} Y_{n-1}$ for all $n \geq 1$, which implies

$$\frac{Y_n}{Y_{n-1}} \leq 1 + c_{n-1} \ , \ \forall n \geq 1.$$

The above is a telescoping product. Along with the assumption $X_n \leq Y_n$, the above argument yields the desired result. □

References

[1] Agarwal, Alekh and Bartlett, Peter L. and Ravikumar, Pradeep and Wainwright, Martin J. (2012) Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization, IEEE Transactions on Information Theory, 58(5):3235-3249, 2012.

[2] Garrigos, Guillaume, Gower, Robert M. (2024) Handbook of Convergence Theorems for (Stochastic) Gradient Methods, arXiv, https://arxiv.org/abs/2301.11235.

[3] Ahmed Khaled and Peter Richtárik (2023) Better Theory for SGD in the Nonconvex World, Transactions on Machine Learning Research.

[4] Knopp, K. (1956) Infinite Sequences and Series, Dover Books on Mathematics, Dover Publications.

[5] Yunwen Lei and Lei Shi and Zheng-Chu Guo (2018). Convergence of Unregularized Online Learning Algorithms, Journal of Machine Learning Research, http://jmlr.org/papers/v18/17-457.html.

[6] Liu, Jun, Yuan, Ye (2024). Almost Sure Convergence Rates Analysis and Saddle Avoidance of Stochastic Gradient Methods, Journal of Machine Learning Research, http://jmlr.org/papers/v25/23-1436.html.

[7] Ljung, L. and Pflug, G.C. and Walk, H. (1992). Stochastic Approximation and Optimization of Random Systems, Oberwolfach Seminars, Birkhäuser Basel.

[8] Lam M. Nguyen and Phuong Ha Nguyen and and Peter Richtárik and Katya Scheinberg and Martin Takáč and Marten van Dijk (2019). New Convergence Aspects of Stochastic Gradient Algorithms, arXiv, https://arxiv.org/abs/1811.12403.

[9] Orabona, Francesco (2020). Almost sure convergence of SGD on smooth nonconvex functions. Blogpost on http://parameterfree.com, available at https://parameterfree.com/ 2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/

[10] Robbins, H. and Siegmund, D. (1971). A convergence Theorem for Non-negative Almost Supermartingales and Some Applications, Optimizing Methods in Statistics.

[11] Sebbouh, Othmane and Gower, Robert M and Defazio, Aaron (2021). Almost sure convergence rates for Stochastic Gradient Descent and Stochastic Heavy Ball, Proceedings of Thirty Fourth Conference on Learning Theory, Vol 134, 3935–3971.

[12] Simon Weissmann and Sara Klein and Waïss Azizian and Leif Döring (2025). Almost sure convergence rates of stochastic gradient methods under gradient domination, arXiv, https://arxiv.org/abs/2405.13592.

[13] Yiming Ying and Ding-Xuan Zhou (2017).Unregularized online learning algorithms with general loss functions, Applied and Computational Harmonic Analysis, 42:224-244.

Department of Mathematics, University of Arizona, 621 N. Santa Rita Ave., Tucson, AZ 85721-0089, USA

Email address: marcelh@arizona.edu