STOCHASTIC HEAVY BALL WITH POLYAK STEP SIZE AND ARMIJO LINE SEARCH: A GENERAL CONVERGENCE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Polyak step size (PS) and Armijo line search (ALS) have received increasing attention in stochastic optimization, with encouraging empirical performance and theoretical guarantees. However, their convergence theory for stochastic heavy ball (SHB) methods remains limited. In this work, we develop a unified convergence analysis for SHB equipped with PS and ALS. To this end, we introduce a modified Armijo rule that closely parallels the Polyak step size, together with a decoupling analysis that isolates the historical dependence induced by momentum. For SHB with standard PS and ALS, we establish expected convergence for strongly convex, convex, and non-convex objectives without interpolation or restrictive conditions on the momentum parameter. Under interpolation or strong growth, we further strengthen the results to almost sure rates and last-iterate convergence. Moreover, for general settings beyond interpolation, we prove almost sure convergence to the exact optimum or to stationarity for SHB with diminishing variants of PS and ALS. These results provide a more comprehensive theoretical view of Polyak step size and Armijo line search for stochastic heavy ball methods.

1 INTRODUCTION

We consider the stochastic optimization problem

$$\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{E}_{x\sim p_x} f(\boldsymbol{\theta}; x). \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameter to be optimized, and x denotes a random data sample drawn from the data distribution p_x. This formulation captures a broad class of fundamental problems in machine learning and related fields (Shalev-Shwartz et al., 2009; Bottou et al., 2018; Krizhevsky et al., 2012; He et al., 2016; Ren et al., 2015; Lin et al., 2017; Vaswani et al., 2017). A widely used method for solving such problems is the stochastic heavy ball (SHB) method (Polyak, 1964; Sutskever et al., 2013; Gadat et al., 2018), whose iteration is given by

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_k \nabla f(\boldsymbol{\theta}_k; x_k) + \beta(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}), \tag{2}$$

where x_k is sampled i.i.d. from p_x, $\beta \in [0, 1)$ is the momentum parameter, and γ_k is the step size at iteration k. We initialize $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$. While the momentum term often improves practical convergence (Sutskever et al., 2013; Gitman et al., 2019), the resulting historical dependence makes SHB substantially harder to analyze than standard stochastic gradient descent (SGD) (Yang et al., 2016; Loizou & Richtárik, 2020; Sebbouh et al., 2021; Ganesh et al., 2023).

In this paper, we focus on SHB with adaptive step sizes, specifically Polyak step size (PS) (Polyak, 1987) and Armijo line search (ALS) (Armijo, 1966). For SGD without momentum, PS and ALS have shown strong empirical performance and enjoy solid theoretical guarantees (Vaswani et al., 2019b; Berrada et al., 2020; Loizou et al., 2021). Motivated by the success in SGD, several recent works have explored variants of PS and ALS for momentum-based optimization algorithms (Wang et al., 2023; Fan et al., 2023; Oikonomou & Loizou, 2025; Zhang et al., 2025; Lapucci & Pucci, 2025; 2026). Nevertheless, a general convergence theory for SHB with PS and ALS remains incomplete. Existing guarantees typically apply only to restricted settings, such as convex objectives or functions satisfying the PL condition, and often rely on interpolation or additional assumptions on the momentum parameter β (Oikonomou & Loizou, 2025; Lapucci & Pucci, 2025; 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Summary of the main results in this paper. $\bar{\boldsymbol{\theta}}_k$ and $\bar{\boldsymbol{\theta}}_k^{\mathrm{w}}$ denote the uniform and weighted averages of the iterates, respectively. S_k denotes the cumulative upper bound of the effective step sizes. Under interpolation or strong growth, we further establish last-iterate convergence. For the diminishing variants, the corresponding convergence rates can be sharpened from big-$\mathcal{O}$ to little-o bounds.

Objective	Noise condition	Guarantee	Theorem
Standard PS / ALS: convergence in expectation			
Strongly convex	$\sigma^2 < \infty$	$\mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|^2 = \mathcal{O}(r^k + \sigma^2)$	1
Convex	$\sigma^2 < \infty$	$\mathbb{E}[F(\bar{\boldsymbol{\theta}}_k) - F(\boldsymbol{\theta}^\star)] = \mathcal{O}(k^{-1} + \sigma^2)$	2
Non-convex	Weak growth	$\min_{0\le m<k} \mathbb{E}\|\nabla F(\boldsymbol{\theta}_m)\|^2 = \mathcal{O}(k^{-1} + \delta)$	3
Standard PS / ALS: almost sure convergence			
Convex	Interpolation	$F(\bar{\boldsymbol{\theta}}_k) - F(\boldsymbol{\theta}^\star) = \mathcal{O}(k^{-1})$, a.s.	4
Non-convex	Strong growth	$\min_{0\le m<k} \|\nabla F(\boldsymbol{\theta}_m)\|^2 = \mathcal{O}(k^{-1})$, a.s.	5
Diminishing PS / ALS: almost sure convergence			
Convex	$\sigma^2 < \infty$	$F(\bar{\boldsymbol{\theta}}_k^{\mathrm{w}}) - F(\boldsymbol{\theta}^\star) = \mathcal{O}(S_k^{-1})$, a.s.	6
Non-convex	Weak growth	$\min_{0\le m\le k} \|\nabla F(\boldsymbol{\theta}_m)\|^2 = \mathcal{O}(S_k^{-1})$, a.s.	7

We address this gap by developing a more general convergence analysis for SHB with PS and ALS. To analyze the two step-size rules in a unified way, we first introduce a modified Armijo rule whose key step-size properties align with those of PS under the SHB iteration. We then provide two technical lemmas that decouple the momentum-dependent terms in SHB: one controls the difference between successive iterates, and the other controls the change in the objective value along the SHB trajectory. Building on these ingredients, we derive convergence rates for SHB with PS and ALS under strongly convex, convex, and general non-convex objectives. These results do not impose the interpolation assumption and apply for any fixed $\beta \in [0, 1)$, without an additional explicit upper bound on the momentum hyperparameter. Under interpolation or strong growth, we further strengthen these results to almost sure rates and last-iterate convergence in both convex and non-convex settings, which is particularly relevant to overparameterized deep learning models. Moreover, beyond interpolation, we analyze SHB with diminishing variants of PS and ALS and establish almost sure exact-convergence guarantees covering both convex and general non-convex objectives.

We summarize our main results in Table 1, and our contributions are as follows:

- We develop a unified convergence framework for SHB with PS and ALS, using a modified Armijo rule to align their key step-size properties and two decoupling lemmas to isolate the dependence between current iterates and historical gradients induced by momentum.
- We establish expected convergence guarantees for strongly convex, convex, and non-convex objectives without interpolation, for any fixed momentum parameter $\beta \in [0, 1)$ under suitable step-size conditions. Under interpolation or strong growth, we further obtain almost sure last-iterate convergence in objective value or gradient norm, respectively.
- Beyond the interpolation or strong growth, we establish almost-sure convergence for SHB equipped with diminishing variants of PS and ALS under a summable momentum schedule. Specifically, we obtain almost-sure rates for the weighted-average objective gap in the convex setting and the best-iterate squared gradient norm in the general non-convex setting.

2 Preliminaries

2.1 Polyak step size and Armijo line search for SHB

Polyak step size (PS) was originally proposed for subgradient methods (Polyak, 1987). PS determines the step size using the objective value and the norm of the subgradient, and enjoys particularly strong convergence guarantees for convex problems. Armijo line search (ALS), on the other hand, selects a step size satisfying a sufficient-decrease condition through a backtracking procedure (Armijo, 1966). ALS is mainly designed for differentiable objectives, which ensures that

a step size satisfying the Armijo rule can be found. Both PS and ALS have been successfully extended to SGD and shown to enjoy favorable convergence guarantees (Vaswani et al., 2019b; Loizou et al., 2021). Throughout, we assume that F is bounded below and $f(\cdot; x)$ is bounded below for p_x-almost every x. Let $\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$ be a global minimizer of F, and denote $f^\star(x) := \inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; x) > -\infty$.

We now introduce the forms of PS and ALS for SHB considered in this paper, termed SHB-PS and SHB-ALS, respectively. Following Oikonomou & Loizou (2025), the form of SHB-PS is consistent with that of PS in SGD (Loizou et al., 2021). On top of this form, we include an additional scaling factor depending on the momentum parameter β, leading to

$$\gamma_{k,\mathrm{SHB-PS}} = \min\left\{ \frac{\left(1-\sqrt{\beta}\right)^2 \left(f(\theta_k; x_k) - f^\star(x_k)\right)}{2c\|\nabla f(\theta_k; x_k)\|^2}, \gamma_{\max} \right\}, \tag{3}$$

where c is a hyperparameter, and $\gamma_{\max}$ is a prescribed upper bound to prevent the step size from becoming unbounded. In practice, $f^\star(x)$ can also be replaced by any valid lower bound on $f(\cdot; x)$ (Orvieto et al., 2022). Note that the resulting step size is equivalent to those in (Oikonomou & Loizou, 2025; Loizou et al., 2021), since the additional constant factor can always be absorbed into the hyperparameter c.

For SHB-ALS, we introduce a slightly modified Armijo rule as

$$f\left(\tilde{\boldsymbol{\theta}}_{k+1}; x_k\right) \le f(\boldsymbol{\theta}_k; x_k) - c\tilde{\gamma}_k \|\nabla f(\boldsymbol{\theta}_k; x_k)\|^2, \tag{4}$$

where $c \in (0, 1)$, and

$$\tilde{\gamma}_k = \frac{2\gamma_k}{\left(1-\sqrt{\beta}\right)^2}, \qquad \tilde{\boldsymbol{\theta}}_{k+1} = \boldsymbol{\theta}_k - \tilde{\gamma}_k \nabla f(\boldsymbol{\theta}_k; x_k). \tag{5}$$

We define $\gamma_{k,\mathrm{SHB-ALS}}$ as the largest γ_k satisfying the above condition, obtained by a backtracking search initialized at $\gamma_{\max}$ with decay factor $\omega \in (0, 1)$.

From these definitions, we immediately obtain the following basic lemma.

Lemma 1 *For $\gamma_k = \gamma_{k,\mathrm{SHB-PS}}$ or $\gamma_k = \gamma_{k,\mathrm{SHB-ALS}}$, we have*

$$\gamma_k \|\nabla f(\boldsymbol{\theta}_k; x_k)\|^2 \le \frac{(1-\sqrt{\beta})^2}{2c} \left(f(\boldsymbol{\theta}_k; x_k) - f^\star(x_k)\right). \tag{6}$$

Lemma 1 follows directly from the definitions of SHB-PS and SHB-ALS. This unified formula allows us to analyze the two algorithms within the same framework in the sequel.

2.2 Assumptions and basic bounds for the step sizes

Throughout the analysis, we assume that the stochastic objective is uniformly Lipschitz smooth.

Assumption 1 (Smoothness) *Assume that $f(\boldsymbol{\theta}; x)$ has $L(x)$-Lipschitz continuous gradients with respect to $\boldsymbol{\theta}$ for p_x-almost every x, i.e.,*

$$\|\nabla f(\boldsymbol{\theta}; x) - \nabla f(\boldsymbol{\theta}'; x)\| \le L(x)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}'. \tag{7}$$

Moreover, we assume that

$$L := \sup_{x \sim p_x} L(x) < \infty. \tag{8}$$

Under Assumption 1, the step sizes of SHB-PS and SHB-ALS are uniformly bounded from below and above. The proof follows the standard arguments for SPS and stochastic line search (Loizou et al., 2021; Vaswani et al., 2019b).

Lemma 2 (Bounds on the step sizes) *Suppose Assumption 1 holds. For SHB-PS, we have*

$$\gamma_{\min,\mathrm{PS}} := \min\left\{ \frac{\left(1-\sqrt{\beta}\right)^2}{4cL}, \gamma_{\max} \right\} \le \gamma_{k,\mathrm{SHB-PS}} \le \gamma_{\max}. \tag{9}$$

For SHB-ALS, we have

$$\gamma_{\min,\mathrm{LS}} := \min\left\{\gamma_{\max}, \left(1-\sqrt{\beta}\right)^2 \frac{\omega(1-c)}{L}\right\} \le \gamma_{k,\mathrm{SHB-ALS}} \le \gamma_{\max}. \tag{10}$$

Below, we use $\gamma_{\min}$ to denote the lower bound on both step sizes in the unified analysis.

For the analysis of convex objectives, we assume that each stochastic component is convex.

Assumption 2 (Convexity) *Assume $f(\boldsymbol{\theta}; x)$ is convex with respect to $\boldsymbol{\theta}$ for p_x-almost every x, i.e.,*

$$f(\boldsymbol{\theta}'; x) \ge f(\boldsymbol{\theta}; x) + \langle \nabla f(\boldsymbol{\theta}; x), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle, \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}'. \tag{11}$$

For the strongly convex analysis, we additionally use the following standard definition.

Definition 1 (Strong convexity) *A differentiable function F is μ-strongly convex if there exists $\mu > 0$ such that, for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$,*

$$F(\boldsymbol{\theta}') \ge F(\boldsymbol{\theta}) + \langle \nabla F(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \frac{\mu}{2}\left\|\boldsymbol{\theta}' - \boldsymbol{\theta}\right\|^2. \tag{12}$$

We also assume that the optimal difference is finite.

Assumption 3 (Finite Optimal Difference) *We assume that*

$$\sigma^2 := \mathbb{E}_{x\sim p_x}\left[f(\boldsymbol{\theta}^\star; x) - f^\star(x)\right] < \infty. \tag{13}$$

The case $\sigma^2 = 0$ corresponds to the interpolation setting (Vaswani et al., 2019b), where the global minimizer of F also minimizes each stochastic objective almost surely.

For general non-convex objectives, we impose the weak growth condition (Loizou et al., 2021).

Assumption 4 (Weak Growth Condition) *We say Problem (1) satisfies the weak growth condition if there exist constants $\rho > 0$ and $\delta \ge 0$ such that, for all $\boldsymbol{\theta}$,*

$$\mathbb{E}_{x\sim p_x}\left[\|\nabla f(\boldsymbol{\theta}; x)\|^2\right] \le \rho\|\nabla F(\boldsymbol{\theta})\|^2 + \delta. \tag{14}$$

When $\delta = 0$, the weak growth condition reduces to the strong growth condition (Schmidt & Roux, 2013), which can be viewed as a non-convex analogue of interpolation.

3 General convergence with standard PS and ALS

3.1 Decoupling lemmas for the SHB dynamics

The main difficulty in analyzing SHB is that the historical gradients accumulated in the momentum term are strongly coupled with the current iterate. We handle this issue by directly analyzing the SHB recursion and establishing two decoupling lemmas tailored to the convex and non-convex analyses, respectively. The following lemma decouples the current-iterate difference from the accumulated momentum term.

Lemma 3 *Let $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ be the iterates generated by SHB (2). Then,*

$$\begin{aligned}&\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star, \sum_{m=0}^{k} \beta^{k-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m)\right\rangle \\ &= \sum_{m=0}^{k} \beta^{k-m}\gamma_m \left\langle \boldsymbol{\theta}_m - \boldsymbol{\theta}^\star, \nabla f(\boldsymbol{\theta}_m; x_m)\right\rangle - \sum_{n=0}^{k-1}\beta^{k-n}\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m; x_m)\right\|^2.\end{aligned} \tag{15}$$

We note that the first term on the right-hand side of Lemma 3 is naturally connected to convexity, while the second term can be controlled using the basic property of PS and ALS in Lemma 1. This decomposition therefore enables our analysis for convex objectives.

The next lemma provides an upper bound on the difference between successive objective values.

Lemma 4 *Let $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ be the iterates generated by SHB (2). Under Assumption 1, we have*

$$F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_k) \leq -\sum_{m=0}^{k} \beta^{k-m}\gamma_m \left\langle \nabla F(\boldsymbol{\theta}_m), \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle + L\sum_{n=0}^{k} \beta^{k-n} \left\| \sum_{m=0}^{n} \beta^{n-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2. \tag{16}$$

Lemma 4 characterizes the relationship between the gradients of the objective function and the successive objective values along the SHB trajectory. This connection allows us to translate the decrease in objective values into a best-iterate stationarity guarantee for general non-convex objectives. The complete proofs of Lemma 3 and Lemma 4 are deferred to Appendix A.

3.2 Expected convergence rates for SHB-PS and SHB-ALS

We are now ready to present the general convergence rates for SHB-PS and SHB-ALS. First, we consider strongly convex objectives, for which the decoupling analysis yields a linear convergence guarantee in expectation for the last iterate.

Theorem 1 *Under Assumptions 1, 2, and 3, suppose further that F is μ-strongly convex (Definition 1). Then, for SHB-PS or SHB-ALS with $c \geq 1/2$, we have*

$$\mathbb{E}\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\right\|^2 \leq Br^k + \frac{2\gamma_{\max}}{\mu\gamma_{\min}}\sigma^2, \tag{17}$$

where $B > 0$ is a constant depending on the initialization, and

$$r = \frac{1 - \mu\gamma_{\min} + \sqrt{(1-\mu\gamma_{\min})^2 - 4\mu\gamma_{\min}\beta}}{2} \in [0, 1). \tag{18}$$

Proof sketch. Combining Lemma 1 and Lemma 3 with strong convexity yields a recursive inequality for the expected squared distance to $\boldsymbol{\theta}^\star$. After subtracting the residual term, the resulting sequence can be controlled through a second-order recurrence induced by momentum, giving the stated linear rate. The complete proof is deferred to Appendix B.1.

Theorem 1 shows that SHB-PS and SHB-ALS converge linearly to a neighborhood of the optimum determined by the optimal difference σ^2. In particular, under interpolation ($\sigma^2 = 0$), the last iterate converges linearly to the exact optimum in expectation. We also note that when $\beta = 0$, SHB reduces to SGD and the contraction factor reduces to $r = 1 - \mu\gamma_{\min}$, recovering the classical SPS rate $\mathcal{O}((1-\mu\gamma_{\min})^k)$ as shown in Loizou et al. (2021).

The following theorem gives the rate of SHB-PS and SHB-ALS for general convex objectives.

Theorem 2 *Under Assumptions 1, 2, and 3, for SHB-PS or SHB-ALS with $c > 1/2$, we have*

$$\mathbb{E}\left[F\left(\bar{\boldsymbol{\theta}}_k\right) - F(\boldsymbol{\theta}^\star)\right] \leq \frac{A}{k} + M \cdot \frac{k+1}{k}\sigma^2, \tag{19}$$

where $\bar{\boldsymbol{\theta}}_k = \frac{1}{k}\sum_{m=0}^{k-1} \boldsymbol{\theta}_m$, and

$$A = \frac{\left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\right\|^2}{(1+\beta)(2-1/c)\gamma_{\min}}, \quad M = \frac{1}{1-\beta^2}\left(\frac{2\gamma_{\max}}{\gamma_{\min}(2-1/c)} - 1\right). \tag{20}$$

Proof sketch. Lemma 3 together with Lemma 1 controls the momentum terms and yields a recursion for the cumulative suboptimality. Taking expectation, summing over iterations, and applying Jensen's inequality to $\bar{\boldsymbol{\theta}}_k$ gives the desired results. The complete proof is deferred to Appendix B.2.

Theorem 2 shows that SHB-PS and SHB-ALS achieve an $\mathcal{O}(1/k)$ rate for convex objectives up to a neighborhood governed by the optimal difference. For SHB-PS, this result is closely related to the rate established for $\mathrm{MomSPS}_{\max}$ (Oikonomou & Loizou, 2025). However, the analysis of $\mathrm{MomSPS}_{\max}$ requires an upper bound on the momentum parameter β. In comparison, our result does not impose an additional upper bound and thus applies to any fixed $\beta \in [0, 1)$.

For general non-convex objectives satisfying the weak growth condition, the following theorem gives a best-iterate stationarity guarantee.

Theorem 3 *Under Assumptions 1 and 4, suppose that $c \geq \rho/4$ for SHB-PS and $c > 1 - 1/(\omega\rho)$ for SHB-ALS. Let $\kappa = 1/c$ for SHB-PS and $\kappa = 4\omega(1-c)$ for SHB-ALS, and further assume that*

$$0 < \gamma_{\max} < \frac{(1-\sqrt{\beta})^2}{4\rho L}\left(1 - \rho + \sqrt{(1-\rho)^2 + 2\rho(1+\rho)\kappa}\right). \tag{21}$$

Then, we have

$$\min_{0\leq m\leq k-1} \mathbb{E}\|\nabla F(\boldsymbol{\theta}_m)\|^2 \leq \frac{F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^\star)}{k(1+\beta)s_1} + \frac{k+1}{k}\frac{s_2\delta}{(1+\beta)s_1}, \tag{22}$$

where

$$s_1 = \frac{\gamma_{\max} + \gamma_{\min}}{2} - \rho q, \qquad s_2 = \frac{q}{1-\beta}, \qquad q = \frac{L\gamma_{\max}^2}{(1-\sqrt{\beta})^2} + \frac{\gamma_{\max} - \gamma_{\min}}{2}. \tag{23}$$

Proof sketch. Combining Lemma 4 with the Cauchy–Schwarz inequality yields a descent bound in terms of a weighted sum of expected gradient norms. Summing this bound and taking the minimum gives the stated best-iterate convergence rate. The complete proof is deferred to Appendix B.3.

Theorem 3 shows that the best iterate of SHB-PS and SHB-ALS approaches a neighborhood of stationarity at an $\mathcal{O}(1/k)$ rate. The hyperparameter conditions in Theorem 3 ensure that the constants s_1 and s_2 are positive, and the detailed verification is also deferred to Appendix B.3. Specifically, the rate of SHB-PS matches the order of the non-convex guarantee of SPS (Loizou et al., 2021). This is also consistent with existing theoretical observations that momentum does not necessarily improve the theoretical order of convergence rates (Ganesh et al., 2023; Zhang et al., 2025).

3.3 Almost sure convergence under interpolation and strong growth

Under interpolation or strong growth, stochastic optimization algorithms often enjoy stronger convergence guarantees (Ma et al., 2018; Vaswani et al., 2019a;b; Berrada et al., 2020). In particular, exact convergence to the optimum or to stationarity can be obtained without diminishing step sizes, as the residual neighborhood terms in Theorems 1, 2 and 3 vanish. Moreover, we can further strengthen the convergence-in-expectation results to almost sure convergence. The convex and non-convex cases are presented in the following two theorems, respectively.

Theorem 4 *Assume that interpolation holds, i.e., $\sigma^2 = 0$, and that the conditions in Theorem 2 hold. Then, for SHB-PS or SHB-ALS, we have*

$$F(\boldsymbol{\theta}_k) \to F(\boldsymbol{\theta}^\star) \quad a.s., \qquad \text{and} \qquad F(\bar{\boldsymbol{\theta}}_k) - F(\boldsymbol{\theta}^\star) = \mathcal{O}(1/k) \quad a.s. \tag{24}$$

Theorem 5 *Assume that the strong growth holds, i.e., $\delta = 0$, and that the conditions in Theorem 3 hold. Then, for SHB-PS or SHB-ALS, we have*

$$\|\nabla F(\boldsymbol{\theta}_k)\| \to 0 \quad a.s., \qquad \text{and} \qquad \min_{0\leq m\leq k-1} \|\nabla F(\boldsymbol{\theta}_m)\|^2 = \mathcal{O}(1/k) \quad a.s. \tag{25}$$

The proofs of Theorems 4 and 5 essentially arise as byproducts of the analysis in the general setting. Under interpolation or strong growth, the corresponding residual terms vanish, allowing us to show that a nonnegative sequence is summable and thereby establish almost sure convergence. The detailed proofs are deferred to Appendices B.4 and B.5. Notably, under interpolation or strong growth, both SHB-PS and SHB-ALS also guarantee last-iterate convergence, which is particularly relevant in practice since the final iterate is typically used as the output of the optimization procedure.

4 Exact convergence with diminishing step sizes

The preceding results establish convergence guarantees for SHB-PS and SHB-ALS under general stochastic objectives. In the non-interpolation regime, however, the residual stochasticity at the optimum generally leads to convergence only to a neighborhood of the optimum or a stationary point. To achieve exact asymptotic convergence, we therefore consider variants of SHB-PS and SHB-ALS following the diminishing step sizes in Sebbouh et al. (2021).

Specifically, we consider the stochastic heavy-ball iteration with a time-varying momentum,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_k \nabla f(\boldsymbol{\theta}_k; x_k) + \beta_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}), \tag{26}$$

where $\beta_k \in [0, 1)$ denotes the momentum parameter at iteration k. The diminishing variant of SHB-PS is defined as

$$\gamma_{k,\text{SHB-PS}_{\text{dec}}} = \eta_k \min\left\{ \frac{f(\boldsymbol{\theta}_k; x_k) - f^*(x_k)}{c\|\nabla f(\boldsymbol{\theta}_k; x_k)\|^2}, \gamma_{\max,k} \right\}, \tag{27}$$

where both η_k and $\gamma_{\max,k}$ are iteration-dependent sequences. Similarly, let $\gamma_{k,\text{SHB-ALS}}(\gamma_{\max,k})$ denote the largest γ_k obtained by backtracking from $\gamma_{\max,k}$ with backtracking factor ω, such that

$$f\left(\tilde{\boldsymbol{\theta}}_{k+1}; x_k\right) \leq f(\boldsymbol{\theta}_k; x_k) - c\gamma_k \|\nabla f(\boldsymbol{\theta}_k; x_k)\|^2, \quad \tilde{\boldsymbol{\theta}}_{k+1} = \boldsymbol{\theta}_k - \gamma_k \nabla f(\boldsymbol{\theta}_k; x_k). \tag{28}$$

The diminishing variant of SHB-ALS is then defined as

$$\gamma_{k,\text{SHB-ALS}_{\text{dec}}} = \eta_k \gamma_{k,\text{SHB-ALS}}(\gamma_{\max,k}). \tag{29}$$

The following lemma provides lower and upper bounds for the diminishing step-size rules, mirroring Lemma 2 for the non-diminishing case.

Lemma 5 *Under Assumption 1, we have*

$$\eta_k \min\left\{ \frac{1}{2cL}, \gamma_{\max,k} \right\} \leq \gamma_{k,\text{SHB-PS}_{\text{dec}}} \leq \eta_k \gamma_{\max,k}, \tag{30}$$

and

$$\eta_k \min\left\{ \gamma_{\max,k}, \frac{2\omega(1-c)}{L} \right\} \leq \gamma_{k,\text{SHB-ALS}_{\text{dec}}} \leq \eta_k \gamma_{\max,k}. \tag{31}$$

Similarly, below we use the unified bounds for both step sizes as

$$\eta_k \gamma_{\min,k} \leq \gamma_k \leq \eta_k \gamma_{\max,k}. \tag{32}$$

Throughout the analysis for diminishing step sizes, we impose the following conditions on the momentum and step-size schedules.

Assumption 5 *We assume*

(i) $\{\beta_k\}_{k\geq 1}$ *is non-increasing,* $\beta_k \in [0, 1)$*, and* $\sum_{k=1}^{\infty} \beta_k < \infty$.

(ii) $\{\eta_k \gamma_{\max,k}\}_{k\geq 1}$ *is non-increasing, and* $\lim_{k\to\infty} \gamma_{\max,k} = 0$.

(iii) $\sum_{k=1}^{\infty} \eta_k \gamma_{\max,k} = \infty$, *and* $\sum_{k=1}^{\infty} \eta_k^2 \gamma_{\max,k} < \infty$.

We note that the summability condition on $\{\beta_k\}$ in (i) has also been adopted in previous convergence analyses of momentum methods (Sun et al., 2019; 2020; Rosasco et al., 2015; Reddi et al., 2018; Luo et al., 2019). Decay conditions on $\gamma_{\max,k}$ in (ii), together with the step-size conditions in (iii), are also used in the analysis of SHB in Sebbouh et al. (2021). These conditions are closely related to the Robbins–Monro condition in stochastic approximation methods (Robbins & Monro, 1951).

We are now ready to present the results of SHB-PS$_{\text{dec}}$ and SHB-ALS$_{\text{dec}}$. First, we establish convergence to the optimal objective value for a weighted average of the iterates in the convex setting.

Theorem 6 *Suppose Assumptions 1, 2, 3, and 5 hold. Assume further that there exists a sufficiently large integer* K *such that*

$$c \geq \sup_{k>K} \frac{\eta_k}{1 + \beta_k - \frac{\beta_k}{1+\beta_k} \frac{\gamma_{\max,k}}{\gamma_{\min,k}}}. \tag{33}$$

Then, for either SHB-PS$_{dec}$ or SHB-ALS$_{dec}$, we have

$$F(\bar{\boldsymbol{\theta}}_k^{\text{w}}) - F(\boldsymbol{\theta}^\star) = \mathcal{O}\left(\frac{1}{\sum_{m=1}^{k} \eta_m \gamma_{\max,m}} \right) \quad \textit{a.s.}, \tag{34}$$

where

$$\bar{\boldsymbol{\theta}}_k^{\text{w}} := \frac{\sum_{m=1}^{k} \eta_m \gamma_{\min,m} \boldsymbol{\theta}_m}{\sum_{m=1}^{k} \eta_m \gamma_{\min,m}}. \tag{35}$$

Proof sketch. We first derive recursive bounds for the squared distance to an optimum and the squared difference between successive iterates. Applying the Robbins–Siegmund theorem twice establishes almost sure summability of the squared iterate differences and weighted objective gaps. Applying Jensen's inequality gives the stated rate. The complete proof is deferred to Appendix C.2.

Here we give a quick remark on the condition on c in (33). Under Assumption 5, the denominator in (33) converges to 1. Consequently, if $\sup_{k>K} \eta_k < 1$ for some K, we can always choose a $c \in (0, 1)$ satisfying (33). This is particularly relevant for SHB-ALS$_{\mathrm{dec}}$, which requires $c < 1$ to ensure that backtracking finds a step size satisfying the Armijo rule.

The following theorem then establishes the convergence rate for the best-iterate squared gradient norm in the non-convex setting.

Theorem 7 *Suppose Assumptions 1, 4, and 5 hold. For either SHB-PS$_{dec}$ or SHB-ALS$_{dec}$, we have*

$$\min_{1\le m\le k} \|\nabla F(\boldsymbol{\theta}_m)\|^2 = \mathcal{O}\left(\frac{1}{\sum_{m=1}^{k} \eta_m \gamma_{\max,m}}\right) \quad \textit{a.s.} \tag{36}$$

Proof sketch. Using smoothness and the weak growth condition, we derive a Lyapunov recursion combining the objective gap with a weighted difference between successive iterates. The assumptions ensure that the error terms are summable. Applying the Robbins–Siegmund theorem and a weighted minimum argument yields the stated rate. The complete proof is deferred to Appendix C.3.

For the decaying variants of SHB-PS and SHB-ALS, Theorem 6 and 7 both establish almost sure convergence. Oikonomou & Loizou (2025) also considers decaying variants of SPS, namely MomDecSPS and MomAdaSPS, which likewise guarantee convergence to the exact solution, but only in expectation. We would like to note that the $\mathcal{O}$ rates in Theorem 6 and 7 can be further strengthened to o rates by exploiting the tail property and smoothness. We provide a more detailed proof of the o-refinement in Appendices C.4 and C.5.

5 Numerical validation

We examine whether the convergence behavior predicted by our analysis is observed on controlled finite-sum problems satisfying the theoretical assumptions. In all experiments, $\boldsymbol{\theta} \in \mathbb{R}^{20}$ and p_x is the uniform distribution on $\{1, \dots, n\}$, with $n = 512$. For convex objectives, we consider

$$f(\boldsymbol{\theta}; i) = \frac{1}{2}\left(\mathbf{a}_i^\top \boldsymbol{\theta} - b_i\right)^2, \qquad b_i = \mathbf{a}_i^\top \boldsymbol{\theta}^\star + \nu\epsilon_i. \tag{37}$$

The matrix $\mathbf{A}$ with rows $\mathbf{a}_i^\top$ has full column rank, and the vector $\boldsymbol{\epsilon}$ with elements ϵ_i satisfies $\mathbf{A}^\top\boldsymbol{\epsilon} = \mathbf{0}$ and $\|\boldsymbol{\epsilon}\|^2/n = 1$. Thus, $\boldsymbol{\theta}^\star$ is the unique minimizer of F. We use $\nu = 0$ as interpolation and $\nu = 0.3$ as non-interpolation. We draw $\boldsymbol{\theta}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{20})$ once and keep it fixed across all runs.

For non-convex experiments, we use

$$f(\boldsymbol{\theta}; i) = s_i \sum_{j=1}^{20} \left[\frac{1}{2}\theta_j^2 + 2(1 - \cos\theta_j)\right] - \boldsymbol{\xi}_i^\top \boldsymbol{\theta}. \tag{38}$$

Under strong growth, we set $\boldsymbol{\xi}_i = \mathbf{0}$ and assign $s_i = 0.5$ and $s_i = 1.5$ for 256 components, respectively. Beyond interpolation, we set $s_i = 1$ and use perturbations $\boldsymbol{\xi}_i \in \{-0.3, 0.3\}^{20}$ occurring in opposite pairs. $\boldsymbol{\theta}^\star$ is set to $\mathbf{0}$. The two settings satisfy Assumption 4 with $(\rho, \delta) = (1.25, 0)$ and $(1, 1.8)$, respectively. We initialize $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0 = \pi\mathbf{1}$.

We implement PS and ALS as defined in Sections 2 and 4, with $c = 0.75$ and $\omega = 0.5$. For the standard (non-diminishing) variants, we use $\gamma_{\max} = 1$ in the convex setting and 0.8 times the upper bound in Theorem 3 in the non-convex setting. We compare $\beta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ under interpolation or strong growth, and use $\beta = 0.9$ otherwise. For the diminishing variants, we set

$$\eta_k = 0.5(1 + k/10^4)^{-0.4}, \quad \beta_k = 0.9(1 + k/10^4)^{-1.1}, \quad \gamma_{\max,k} = \gamma_{\max,0}(1 + k/10^4)^{-0.3}.$$

We set $\gamma_{\max,0} = 0.02$ and $\gamma_{\max,0} = 2\gamma_{\max}^{\mathrm{std}}$ for convex and non-convex objectives, respectively. Here $\gamma_{\max}^{\mathrm{std}}$ denotes the upper bound used by the corresponding standard step sizes at $\beta = 0.9$. All experiments use 20 random seeds, and the metrics are chosen to match the corresponding theorems.

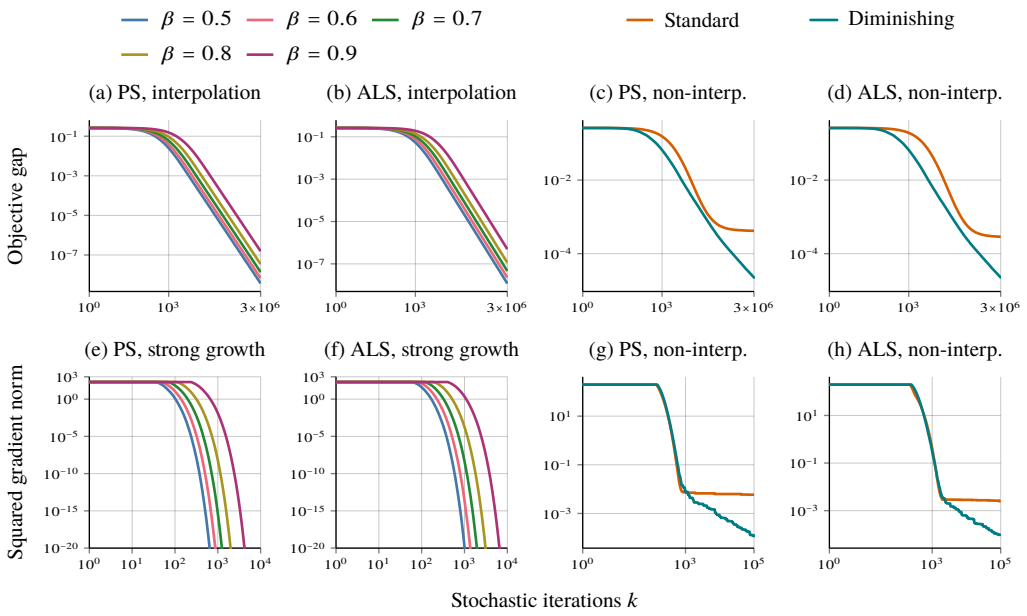


Figure 1: Convergence of SHB with PS and ALS on convex (top) and non-convex (bottom) objectives. Panels (a)–(b) and (e)–(f) show the standard, non-diminishing variants, while (c)–(d) and (g)–(h) additionally include the diminishing variants. For the standard step sizes, we report the empirical counterparts of the expected quantities in Theorems 2 and 3: the objective gap at the uniformly averaged iterate for convex objectives, and the running minimum of the squared full-gradient norm for non-convex objectives. For the diminishing step sizes, which admit almost-sure convergence guarantees, we report the pointwise *maximum* as a proxy across runs of the corresponding pathwise quantities: the weighted-average objective gap for convex objectives and the best-iterate squared gradient norm for non-convex objectives, following Theorems 6 and 7, respectively.

Figure 1 illustrates the behaviour of adaptive step sizes under different assumptions. (a)–(b) and (e)–(f) validate the general convergence of non-diminishing SHB-PS and SHB-ALS under different constant β, which is consistent with Theorem 2, 3 and the discussion in Section 3.2. Beyond interpolation, (c)–(d) and (g)–(h) show that non-diminishing step sizes converge only to a neighborhood, while the diminishing variants continuously converge, which is consistent with Theorem 6 and 7. More numerical results in machine learning tasks can be found in Appendix E.

6 Conclusion

We develop a unified convergence analysis for stochastic heavy ball methods with the Polyak step size and Armijo line search. Building on a modified Armijo rule and two decoupling lemmas that isolate the effect of the momentum term, we establish expected convergence guarantees for both convex and non-convex objectives without imposing restrictive conditions on the momentum parameter. Under interpolation or strong growth, our analysis further yields almost sure rates and last-iterate convergence. Beyond the interpolation regime, we show that diminishing variants of both step sizes exhibit almost-sure exact convergence. Together, these results provide a more comprehensive theoretical understanding of Polyak step size and Armijo line search in stochastic heavy ball.

Limitations and future work. Our analysis of the diminishing variants of SHB-PS and SHB-ALS still requires the momentum parameters β_k to be summable. This excludes commonly used constant momentum choices in practice, such as $\beta_k = 0.9$. Developing a more refined analysis that can accommodate non-summable or constant momentum therefore remains an important direction toward further narrowing the gap between theory and practice. Furthermore, since optimization algorithms are typically evaluated using their final iterates, extending the results to establish last-iterate convergence in the general setting would also be of significant interest.

AI USE STATEMENT

We use generative AI tools to assist with language editing, technical review, and the development of experimental code. The authors have reviewed the AI-assisted material and take full responsibility for the final content of this paper, including its proofs, technical claims, and experimental results.

ETHICS STATEMENT

This work primarily develops convergence theory for stochastic optimization methods, supported by numerical experiments on synthetic problems and public benchmark datasets. It does not involve human-subject research or the collection of personal data. We do not identify specific ethical concerns arising directly from the theoretical contributions.

REPRODUCIBILITY STATEMENT

The algorithmic definitions and assumptions are provided in Section 2, and the convergence results and conditions for diminishing step sizes are stated in Sections 3 and 4. Complete proofs are provided in Appendix A, B, and C. The numerical objectives, parameter choices, and reported statistics are described in Section 5, with additional experimental settings provided in Appendix E.

REFERENCES

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

Stefania Bellavia, Nataša Krejić, Nataša Krklec Jerinkić, and Marcos Raydan. SLiSeS: subsampled line search spectral gradient method for finite sums. *Optimization Methods and Software*, 41(2): 524–549, 2026.

Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Training neural networks for and by interpolation. In *International Conference on Machine Learning*, 2020.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2), 2018.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

Chen Fan, Sharan Vaswani, Christos Thrampoulidis, and Mark Schmidt. Msl: An adaptive momentum-based stochastic line-search framework. In *NeurIPS 2023 Workshop on Optimization for Machine Learning*, 2023.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

Swetha Ganesh, Rohan Deb, Gugan Thoppe, and Amarjit Budhiraja. Does momentum help in stochastic optimization? a sample complexity analysis. In *Uncertainty in Artificial Intelligence*, 2023.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Xiaowen Jiang and Sebastian U Stich. Adaptive SGD with Polyak stepsize and line-search: Robust convergence and variance reduction. In *Advances in Neural Information Processing Systems*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Matteo Lapucci and Davide Pucci. Convergence conditions for stochastic line search based optimization of over-parametrized models. *Optimization*, 2025.

Matteo Lapucci and Davide Pucci. Effectively leveraging momentum terms in stochastic line search frameworks for fast optimization of finite-sum problems. *Computational Optimization and Applications*, 94(2):543–569, 2026.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017.

Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Liangchen Luo, Yuanhao Xiong, and Yan Liu. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, 2018.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Dimitris Oikonomou and Nicolas Loizou. Stochastic polyak step-sizes and momentum: Convergence guarantees and practical performance. In *International Conference on Learning Representations*, 2025.

Antonio Orvieto, Simon Lacoste-Julien, and Nicolas Loizou. Dynamics of SGD with stochastic Polyak stepsizes: Truly adaptive variants and convergence to exact solution. In *Advances in Neural Information Processing Systems*, 2022.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

Michal Rolinek and Georg Martius. L4: Practical loss-based stepsize adaptation for deep learning. In *Advances in Neural Information Processing Systems*, 2018.

Lorenzo Rosasco, Silvia Villa, and Bang Công Vu. Stochastic inertial primal-dual algorithms. *arXiv preprint arXiv:1507.00852*, 2015.

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

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, 2021.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, 2009.

Tao Sun, Penghang Yin, Dongsheng Li, Chun Huang, Lei Guan, and Hao Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. In *AAAI Conference on Artificial Intelligence*, 2019.

Tao Sun, Linbo Qiao, and Dongsheng Li. Nonergodic complexity of proximal inertial gradient descents. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4613–4626, 2020.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, 2019b.

Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized polyak step size for first-order optimization with momentum. In *International Conference on Machine Learning*, 2023.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

Jiawei Zhang, Cheng Jin, and Yuantao Gu. Adaptive polyak step-size for momentum-accelerated stochastic gradient descent with general convergence guarantee. *IEEE Transactions on Signal Processing*, 73, 2025.

A PROOF OF TECHNICAL LEMMAS

We first provide the proofs of all lemmas used in our analysis. The proofs of Lemmas 1 and 2 follow directly from the definitions or standard arguments in prior work (Vaswani et al., 2019b; Loizou et al., 2021; Wang et al., 2023; Oikonomou & Loizou, 2025; Zhang et al., 2025). We include them here for completeness.

A.1 PROOF OF LEMMA 1

For SHB-PS, it follows directly from (3) that

$$\gamma_{k,\text{SHB-PS}} \leq \frac{(1-\sqrt{\beta})^2 \left(f(\boldsymbol{\theta}_k; x_k) - f^\star(x_k)\right)}{2c \left\|\nabla f(\boldsymbol{\theta}_k; x_k)\right\|^2}. \tag{39}$$

Rearranging the above inequality gives the desired result.

For SHB-ALS, by (4) and (5), any accepted step size satisfies

$$\frac{2c\gamma_{k,\text{SHB-ALS}}}{(1-\sqrt{\beta})^2} \left\|\nabla f(\boldsymbol{\theta}_k; x_k)\right\|^2 \leq f(\boldsymbol{\theta}_k; x_k) - f(\tilde{\boldsymbol{\theta}}_{k+1}; x_k) \leq f(\boldsymbol{\theta}_k; x_k) - f^\star(x_k).$$

The claimed bound follows immediately.

A.2 PROOF OF LEMMA 2

Under Assumption 1, the descent lemma (Nesterov, 2018) gives

$$f(\boldsymbol{\theta}'; x) \leq f(\boldsymbol{\theta}; x) + \langle \nabla f(\boldsymbol{\theta}; x), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2} \left\|\boldsymbol{\theta}' - \boldsymbol{\theta}\right\|^2. \tag{40}$$

It further implies (Loizou et al., 2021)

$$f(\boldsymbol{\theta}_k; x_k) - f^\star(x_k) \geq \frac{1}{2L} \left\|\nabla f(\boldsymbol{\theta}_k; x_k)\right\|^2. \tag{41}$$

Substituting this inequality into the definition of SHB-PS yields

$$\gamma_{k,\text{SHB-PS}} \geq \min\left\{\frac{(1-\sqrt{\beta})^2}{4cL}, \gamma_{\max}\right\} = \gamma_{\min,\text{PS}}. \tag{42}$$

For SHB-ALS, we first show that any step size satisfying

$$\gamma_k \leq (1-\sqrt{\beta})^2 \frac{1-c}{L} \tag{43}$$

satisfies the modified Armijo rule (4). By (40), we have

$$f(\boldsymbol{\theta}_k; x_k) - f\left(\boldsymbol{\theta}_k - \tilde{\gamma}_k \nabla f(\boldsymbol{\theta}_k; x_k); x_k\right) \geq \tilde{\gamma}_k \left(1 - \frac{L\tilde{\gamma}_k}{2}\right) \left\|\nabla f(\boldsymbol{\theta}_k; x_k)\right\|^2.$$

Since $\tilde{\gamma}_k = 2\gamma_k/(1-\sqrt{\beta})^2$, the above condition (43) on γ_k ensures $1 - L\tilde{\gamma}_k/2 \geq c$, and hence (4) holds. Accounting for the worst-case reduction by the backtracking factor ω therefore gives

$$\gamma_{k,\text{SHB-ALS}} \geq \min\left\{\gamma_{\max}, (1-\sqrt{\beta})^2 \frac{\omega(1-c)}{L}\right\}, \tag{44}$$

which completes the proof.

A.3 PROOF OF LEMMA 3

Recall that the SHB recursion, with the initialization $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$, implies

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \sum_{m=0}^{k-1} \beta^{k-1-m} \gamma_m \nabla f(\boldsymbol{\theta}_m; x_m). \tag{45}$$

Therefore,

$$\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star, \sum_{m=0}^{k} \beta^{k-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle - \beta \left\langle \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^\star, \sum_{m=0}^{k-1} \beta^{k-1-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle$$
$$= \gamma_k \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star, \nabla f(\boldsymbol{\theta}_k; x_k) \right\rangle - \beta \left\| \sum_{m=0}^{k-1} \beta^{k-1-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 . \tag{46}$$

Recursively applying the above identity yields

$$\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star, \sum_{m=0}^{k} \beta^{k-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle$$
$$= \sum_{m=0}^{k} \beta^{k-m}\gamma_m \left\langle \boldsymbol{\theta}_m - \boldsymbol{\theta}^\star, \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle - \sum_{n=0}^{k-1} \beta^{k-n} \left\| \sum_{m=0}^{n} \beta^{n-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 , \tag{47}$$

which proves the lemma.

A.4 Proof of Lemma 4

Under Assumption 1, F is also L-smooth. Therefore,

$$F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_k) \le \left\langle \nabla F(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \right\rangle + \frac{L}{2} \left\| \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \right\|^2$$
$$= - \sum_{m=0}^{k} \beta^{k-m}\gamma_m \left\langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle + \frac{L}{2} \left\| \sum_{m=0}^{k} \beta^{k-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 . \tag{48}$$

Similarly,

$$F(\boldsymbol{\theta}_{k-1}) - F(\boldsymbol{\theta}_k)$$
$$\le \sum_{m=0}^{k-1} \beta^{k-1-m}\gamma_m \left\langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle + \frac{L}{2} \left\| \sum_{m=0}^{k-1} \beta^{k-1-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 . \tag{49}$$

Multiplying (49) by β and adding it to (48), one cancels the historical gradient terms, yielding

$$F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_k) - \beta \left(F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_{k-1}) \right)$$
$$\le - \gamma_k \left\langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_k; x_k) \right\rangle + \frac{L}{2} \left\| \sum_{m=0}^{k} \beta^{k-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2$$
$$+ \frac{L\beta}{2} \left\| \sum_{m=0}^{k-1} \beta^{k-1-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 . \tag{50}$$

Recursively applying the above inequality gives

$$F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_k) - \beta^k \left(F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_0) \right)$$
$$\le - \sum_{m=1}^{k} \beta^{k-m}\gamma_m \left\langle \nabla F(\boldsymbol{\theta}_m), \nabla f(\boldsymbol{\theta}_m; x_m) \right\rangle + \frac{L}{2} \sum_{n=1}^{k} \beta^{k-n} \left\| \sum_{m=0}^{n} \beta^{n-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2$$
$$+ \frac{L\beta}{2} \sum_{n=1}^{k} \beta^{k-n} \left\| \sum_{m=0}^{n-1} \beta^{n-1-m}\gamma_m \nabla f(\boldsymbol{\theta}_m; x_m) \right\|^2 . \tag{51}$$

For the initial step, the descent lemma gives

$$F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_0) \le -\gamma_0 \left\langle \nabla F(\boldsymbol{\theta}_0), \nabla f(\boldsymbol{\theta}_0; x_0) \right\rangle + \frac{L}{2} \left\| \gamma_0 \nabla f(\boldsymbol{\theta}_0; x_0) \right\|^2 . \tag{52}$$

Substituting (52) into (51) and collecting the weighted terms proves the lemma.

B Proofs for standard SHB-PS and SHB-ALS in Section 3

B.1 Proof of Theorem 1

Unrolling the SHB recursion (2) and applying Lemma 3, we have

$$\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2 \le \|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2 - 2\sum_{m=0}^{k}\beta^{k-m}\gamma_m\left\langle \boldsymbol{\theta}_m-\boldsymbol{\theta}^\star, \nabla f(\boldsymbol{\theta}_m;x_m)\right\rangle + 2\sum_{n=0}^{k}\beta^{k-n}\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2. \tag{53}$$

By the Cauchy–Schwarz inequality,

$$\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2 \le \left(\sum_{m=0}^{n}\beta^{\frac{3}{2}(n-m)}\gamma_m^2\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2\right)\left(\sum_{m=0}^{n}\beta^{\frac{1}{2}(n-m)}\right) \le \frac{1}{1-\sqrt{\beta}}\sum_{m=0}^{n}\beta^{\frac{3}{2}(n-m)}\gamma_m^2\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2. \tag{54}$$

Reordering the summation and applying Lemma 1 then gives

$$\sum_{n=0}^{k}\beta^{k-n}\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2 \le \frac{1}{(1-\sqrt{\beta})^2}\sum_{m=0}^{k}\beta^{k-m}\gamma_m^2\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2 \le \frac{1}{2c}\sum_{m=0}^{k}\beta^{k-m}\gamma_m\left(f(\boldsymbol{\theta}_m;x_m)-f^\star(x_m)\right). \tag{55}$$

Combining (53) and (55), and using $c \ge 1/2$, yields

$$\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2 \le \|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2 - 2\sum_{m=0}^{k}\beta^{k-m}\gamma_m\left(\langle\boldsymbol{\theta}_m-\boldsymbol{\theta}^\star,\nabla f(\boldsymbol{\theta}_m;x_m)\rangle - f(\boldsymbol{\theta}_m;x_m)+f(\boldsymbol{\theta}^\star;x_m)\right) + 2\sum_{m=0}^{k}\beta^{k-m}\gamma_m\left(f(\boldsymbol{\theta}^\star;x_m)-f^\star(x_m)\right). \tag{56}$$

The two summations are nonnegative by convexity and the definition of $f^\star(x_m)$. Hence, using Lemma 2, taking expectation, and applying the μ-strong convexity of F, we obtain

$$\mathbb{E}\left\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\right\|^2 \le \mathbb{E}\left\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\right\|^2 - \mu\gamma_{\min}\sum_{m=0}^{k}\beta^{k-m}\mathbb{E}\left\|\boldsymbol{\theta}_m-\boldsymbol{\theta}^\star\right\|^2 + 2\gamma_{\max}\sigma^2\sum_{m=0}^{k}\beta^{k-m}. \tag{57}$$

Let

$$e_k := \mathbb{E}\left\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\right\|^2 - \frac{2\gamma_{\max}}{\mu\gamma_{\min}}\sigma^2. \tag{58}$$

Then (57) becomes

$$e_{k+1} \le e_k - \mu\gamma_{\min}\sum_{m=0}^{k}\beta^{k-m}e_m. \tag{59}$$

We first note that once e_k becomes nonpositive, it remains nonpositive thereafter. For $\beta = 0$, this follows immediately from $e_{k+1} \le (1-\mu\gamma_{\min})e_k$. For $\beta > 0$, let k_0 be the first index such that $e_{k_0} \le 0$, and define

$$S_k := \sum_{m=k_0}^{k}\beta^{-m}e_m, \qquad S_{k_0-1} := 0. \tag{60}$$

By (59),

$$\beta S_{k+1} - (\beta + 1 - \mu\gamma_{\min})\, S_k + S_{k-1} \leq 0. \tag{61}$$

Let

$$z_{\pm} = \frac{\beta + 1 - \mu\gamma_{\min} \pm \sqrt{(\beta + 1 - \mu\gamma_{\min})^2 - 4\beta}}{2\beta}. \tag{62}$$

Since $\mu\gamma_{\min} \leq (1 - \sqrt{\beta})^2$, the roots are real and positive, with $z_+ > 1$. Thus (61) gives

$$S_{k+1} - z_+ S_k \leq z_- \left(S_k - z_+ S_{k-1}\right). \tag{63}$$

Since $S_{k_0} = \beta^{-k_0} e_{k_0} \leq 0$, recursively applying the above inequality yields $S_{k+1} - z_+ S_k \leq 0$ and $S_k \leq 0$ for all $k \geq k_0$. Therefore,

$$e_{k+1} = \beta^{k+1}(S_{k+1} - S_k) \leq 0, \tag{64}$$

which proves the claim.

It remains to consider the case where $e_0, \ldots, e_k > 0$. Retaining only the two most recent terms in (59) gives

$$e_{k+1} \leq (1 - \mu\gamma_{\min}) e_k - \mu\gamma_{\min}\beta e_{k-1}. \tag{65}$$

Let

$$r_{\pm} = \frac{1 - \mu\gamma_{\min} \pm \sqrt{(1 - \mu\gamma_{\min})^2 - 4\mu\gamma_{\min}\beta}}{2}. \tag{66}$$

The step-size bounds in Lemma 2 and $c \geq 1/2$ imply

$$\mu\gamma_{\min} \leq \frac{\mu}{2L}(1 - \sqrt{\beta})^2 \leq \frac{1}{2}(1 - \sqrt{\beta})^2, \tag{67}$$

which guarantees that the discriminant is positive and $0 \leq r_- \leq r_+ < 1$. Moreover, $r_+ + r_- = 1 - \mu\gamma_{\min}$ and $r_+ r_- = \mu\gamma_{\min}\beta$. Hence (65) can be written as

$$e_{k+1} - r_- e_k \leq r_+ \left(e_k - r_- e_{k-1}\right). \tag{68}$$

Iterating the above inequality gives

$$e_k \leq \frac{e_1 - r_- e_0}{r_+ - r_-} r_+^k + \frac{r_+ e_0 - e_1}{r_+ - r_-} r_-^k \leq B r_+^k, \tag{69}$$

for some constant $B \geq 0$ determined by the initialization. Combining this bound with the nonpositive case above, and recalling the definition of e_k, yields

$$\mathbb{E}\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\right\|^2 \leq B r^k + \frac{2\gamma_{\max}}{\mu\gamma_{\min}}\sigma^2, \tag{70}$$

where $r = r_+$, which proves the theorem.

B.2 Proof of Theorem 2

Following the steps in Appendix B.1, combining (53) and (55), together with convexity, yields

$$\begin{aligned}\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^\star\right\|^2 \leq \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\right\|^2 &- \left(2 - \frac{1}{c}\right) \sum_{m=0}^{k} \beta^{k-m} \gamma_m \left(f(\boldsymbol{\theta}_m; x_m) - f^\star(x_m)\right) \\ &+ 2 \sum_{m=0}^{k} \beta^{k-m} \gamma_m \left(f(\boldsymbol{\theta}^\star; x_m) - f^\star(x_m)\right).\end{aligned} \tag{71}$$

Since $c > 1/2$, using the step-size bounds in Lemma 2 gives

$$\begin{aligned}\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^\star\right\|^2 \leq \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\right\|^2 &- \left(2 - \frac{1}{c}\right) \gamma_{\min} \sum_{m=0}^{k} \beta^{k-m} \left(f(\boldsymbol{\theta}_m; x_m) - f(\boldsymbol{\theta}^\star; x_m)\right) \\ &+ \left[2\gamma_{\max} - \left(2 - \frac{1}{c}\right)\gamma_{\min}\right] \sum_{m=0}^{k} \beta^{k-m} \left(f(\boldsymbol{\theta}^\star; x_m) - f^\star(x_m)\right).\end{aligned} \tag{72}$$

Taking expectation and using Assumption 3, we obtain

$$\left(2-\frac{1}{c}\right)\gamma_{\min}\sum_{m=0}^{k}\beta^{k-m}\mathbb{E}\left[F(\boldsymbol{\theta}_m)-F(\boldsymbol{\theta}^\star)\right]\leq\mathbb{E}\left\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\right\|^2-\mathbb{E}\left\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\right\|^2 + \frac{2\gamma_{\max}-\left(2-\frac{1}{c}\right)\gamma_{\min}}{1-\beta}\sigma^2. \tag{73}$$

Retaining the two most recent terms in the weighted sum and summing (73) over the iterations gives

$$(1+\beta)\sum_{m=0}^{k-1}\mathbb{E}\left[F(\boldsymbol{\theta}_m)-F(\boldsymbol{\theta}^\star)\right]\leq\frac{\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^\star\right\|^2}{\left(2-\frac{1}{c}\right)\gamma_{\min}}+(k+1)\left(\frac{2\gamma_{\max}}{\left(2-\frac{1}{c}\right)\gamma_{\min}}-1\right)\frac{\sigma^2}{1-\beta}. \tag{74}$$

Finally, by convexity of F and Jensen's inequality,

$$F(\bar{\boldsymbol{\theta}}_k)-F(\boldsymbol{\theta}^\star)\leq\frac{1}{k}\sum_{m=0}^{k-1}\left(F(\boldsymbol{\theta}_m)-F(\boldsymbol{\theta}^\star)\right). \tag{75}$$

Combining this with (74) yields

$$\mathbb{E}\left[F(\bar{\boldsymbol{\theta}}_k)-F(\boldsymbol{\theta}^\star)\right]\leq\frac{A}{k}+M\frac{k+1}{k}\sigma^2, \tag{76}$$

with A and M defined in Theorem 2, which completes the proof.

B.3 Proof of Theorem 3

By Lemma 4, we have

$$F(\boldsymbol{\theta}_{k+1})-F(\boldsymbol{\theta}_k)\leq-\sum_{m=0}^{k}\beta^{k-m}\gamma_m\left\langle\nabla F(\boldsymbol{\theta}_m),\nabla f(\boldsymbol{\theta}_m;x_m)\right\rangle + L\sum_{n=0}^{k}\beta^{k-n}\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2. \tag{77}$$

For the first term, using $\gamma_{\min}\leq\gamma_m\leq\gamma_{\max}$ gives

$$-\gamma_m\left\langle\nabla F(\boldsymbol{\theta}_m),\nabla f(\boldsymbol{\theta}_m;x_m)\right\rangle\leq\frac{\gamma_{\max}-\gamma_{\min}}{2}\left(\left\|\nabla F(\boldsymbol{\theta}_m)\right\|^2+\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2\right) - \gamma_{\max}\left\langle\nabla F(\boldsymbol{\theta}_m),\nabla f(\boldsymbol{\theta}_m;x_m)\right\rangle. \tag{78}$$

For the second term, using (54) and reordering the summation gives

$$L\sum_{n=0}^{k}\beta^{k-n}\left\|\sum_{m=0}^{n}\beta^{n-m}\gamma_m\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2\leq\frac{L}{1-\sqrt{\beta}}\sum_{m=0}^{k}\beta^{k-m}\gamma_m^2\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2\sum_{n=m}^{k}\beta^{\frac{1}{2}(n-m)} \leq\frac{L\gamma_{\max}^2}{\left(1-\sqrt{\beta}\right)^2}\sum_{m=0}^{k}\beta^{k-m}\left\|\nabla f(\boldsymbol{\theta}_m;x_m)\right\|^2. \tag{79}$$

Combining (77), (78) and (79), taking expectation, and applying Assumption 4, we obtain

$$\mathbb{E}\left[F(\boldsymbol{\theta}_{k+1})-F(\boldsymbol{\theta}_k)\right]\leq-s_1\sum_{m=0}^{k}\beta^{k-m}\mathbb{E}\left\|\nabla F(\boldsymbol{\theta}_m)\right\|^2+s_2\delta, \tag{80}$$

where s_1 and s_2 are defined in Theorem 3. The stated hyperparameter conditions ensure $s_1>0$, and we verify this at the end of the proof. Summing (80) over k and using $F(\boldsymbol{\theta})\geq F(\boldsymbol{\theta}^\star)$ gives

$$s_1\sum_{n=0}^{k}\sum_{m=0}^{n}\beta^{n-m}\mathbb{E}\left\|\nabla F(\boldsymbol{\theta}_m)\right\|^2\leq F(\boldsymbol{\theta}_0)-F(\boldsymbol{\theta}^\star)+(k+1)s_2\delta. \tag{81}$$

Retaining the terms with weights 1 and β yields

$$(1+\beta)s_1 \sum_{m=0}^{k-1} \mathbb{E}\left\|\nabla F(\boldsymbol{\theta}_m)\right\|^2 \leq F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^\star) + (k+1)s_2\delta. \tag{82}$$

Therefore,

$$\min_{0\leq m\leq k-1} \mathbb{E}\left\|\nabla F(\boldsymbol{\theta}_m)\right\|^2 \leq \frac{F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^\star)}{k(1+\beta)s_1} + \frac{k+1}{k}\frac{s_2\delta}{(1+\beta)s_1}, \tag{83}$$

which gives the desired result.

It remains to verify that $s_1 > 0$ under the stated hyperparameter conditions. For SHB-ALS, we consider the two possible cases. If $\gamma_{\max} \leq (1-\sqrt{\beta})^2\omega(1-c)/L$, then $\gamma_{\min} = \gamma_{\max}$ and

$$s_1 = \gamma_{\max} - \rho\frac{L\gamma_{\max}^2}{(1-\sqrt{\beta})^2} = \gamma_{\max}\left(1 - \frac{\rho L\gamma_{\max}}{(1-\sqrt{\beta})^2}\right). \tag{84}$$

Thus $c > 1 - \frac{1}{\omega\rho}$ implies $s_1 > 0$. Otherwise, $\gamma_{\min} = (1-\sqrt{\beta})^2\omega(1-c)/L$. Substituting it into the definition of s_1 gives

$$2s_1 = -\frac{2\rho L}{(1-\sqrt{\beta})^2}\gamma_{\max}^2 + (1-\rho)\gamma_{\max} + (1+\rho)(1-\sqrt{\beta})^2\frac{\omega(1-c)}{L}. \tag{85}$$

The right-hand side is positive whenever $\gamma_{\max}$ is smaller than the positive root of the corresponding quadratic equation, namely,

$$\gamma_{\max} < \frac{(1-\sqrt{\beta})^2}{4\rho L}\left(1-\rho+\sqrt{(1-\rho)^2 + 8\rho(1+\rho)\omega(1-c)}\right), \tag{86}$$

which is precisely the assumed upper bound for SHB-ALS.

For SHB-PS, similarly, if $\gamma_{\max} \leq (1-\sqrt{\beta})^2/(4cL)$, then again $\gamma_{\min} = \gamma_{\max}$ and

$$s_1 = \gamma_{\max}\left(1 - \frac{\rho L\gamma_{\max}}{(1-\sqrt{\beta})^2}\right). \tag{87}$$

Moreover, $c \geq \rho/4$ implies

$$\frac{(1-\sqrt{\beta})^2}{4cL} \leq \frac{(1-\sqrt{\beta})^2}{\rho L}. \tag{88}$$

Together with the strict upper bound on $\gamma_{\max}$ in the theorem, this gives $s_1 > 0$. Otherwise, $\gamma_{\min} = (1-\sqrt{\beta})^2/(4cL)$, and hence

$$2s_1 = -\frac{2\rho L}{(1-\sqrt{\beta})^2}\gamma_{\max}^2 + (1-\rho)\gamma_{\max} + (1+\rho)\frac{(1-\sqrt{\beta})^2}{4cL}. \tag{89}$$

Therefore, $s_1 > 0$ whenever

$$\gamma_{\max} < \frac{(1-\sqrt{\beta})^2}{4\rho L}\left(1-\rho+\sqrt{(1-\rho)^2 + \frac{2\rho(1+\rho)}{c}}\right), \tag{90}$$

which is the assumed upper bound for SHB-PS.

B.4 Proof of Theorem 4

Under interpolation, $\sigma^2 = 0$. Then following (74) in the proof of Theorem 2, we have

$$(1+\beta)\sum_{m=0}^{k-1} \mathbb{E}\left[F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)\right] \leq \frac{\left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\right\|^2}{\left(2-\frac{1}{c}\right)\gamma_{\min}}. \tag{91}$$

Since $F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star) \geq 0$, letting $k \to \infty$ and applying the monotone convergence theorem gives

$$\mathbb{E}\left[\sum_{m=0}^{\infty}\left(F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)\right)\right] < \infty. \tag{92}$$

Hence,

$$\sum_{m=0}^{\infty} (F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)) < \infty \quad \text{a.s.}, \tag{93}$$

which implies $F(\boldsymbol{\theta}_k) \to F(\boldsymbol{\theta}^\star)$ almost surely. Moreover, by convexity,

$$F(\bar{\boldsymbol{\theta}}_k) - F(\boldsymbol{\theta}^\star) \le \frac{1}{k} \sum_{m=0}^{k-1} (F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)) = \mathcal{O}(1/k) \quad \text{a.s.}, \tag{94}$$

which completes the proof.

B.5 Proof of Theorem 5

When $\delta = 0$, using (82) in the proof of Theorem 3 yields

$$\sum_{m=0}^{k-1} \mathbb{E} \left\| \nabla F(\boldsymbol{\theta}_m) \right\|^2 \le \frac{F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^\star)}{(1+\beta)s_1}. \tag{95}$$

Letting $k \to \infty$ and applying the monotone convergence theorem yields

$$\mathbb{E} \left[\sum_{m=0}^{\infty} \left\| \nabla F(\boldsymbol{\theta}_m) \right\|^2 \right] < \infty. \tag{96}$$

Therefore,

$$\sum_{m=0}^{\infty} \left\| \nabla F(\boldsymbol{\theta}_m) \right\|^2 < \infty \quad \text{a.s.}, \tag{97}$$

and consequently $\|\nabla F(\boldsymbol{\theta}_k)\| \to 0$ almost surely. Furthermore,

$$\min_{0 \le m \le k-1} \left\| \nabla F(\boldsymbol{\theta}_m) \right\|^2 \le \frac{1}{k} \sum_{m=0}^{k-1} \left\| \nabla F(\boldsymbol{\theta}_m) \right\|^2 = \mathcal{O}(1/k) \quad \text{a.s.}, \tag{98}$$

which proves the result.

C Proofs for diminishing step sizes in Section 4

We first collect the auxiliary lemmas used in the diminishing-step-size analysis in Appendix C.1. Then we prove Theorems 6 and 7 in Appendices C.2 and C.3, respectively. Finally, Appendices C.4 and C.5 establish the corresponding little-o refinements.

Throughout this section, let $\mathcal{F}_k = \sigma(\boldsymbol{\theta}_0, x_0, \ldots, x_{k-1})$ denote the natural filtration, i.e., the information before sampling x_k. The schedule hyperparameters are deterministic, with $\eta_k, \gamma_{\max,k} > 0$. All almost sure bounds below allow finite constants that depend on the sample path.

C.1 Auxiliary lemmas

The following is the diminishing-step-size counterpart of Lemma 1.

Lemma 6 *For either SHB-PS$_{dec}$ with $c > 0$ or SHB-ALS$_{dec}$ with $c \in (0, 1)$, we have*

$$\gamma_k \|\nabla f(\boldsymbol{\theta}_k; x_k)\|^2 \le \frac{\eta_k}{c} \big(f(\boldsymbol{\theta}_k; x_k) - f^\star(x_k) \big). \tag{99}$$

We also use the almost-supermartingale convergence theorem of Robbins & Siegmund (1971).

Lemma 7 (Robbins–Siegmund) *Let $\{V_k\}$, $\{U_k\}$, and $\{Z_k\}$ be nonnegative, integrable processes adapted to $\{\mathcal{F}_k\}$. Suppose that $\{\alpha_k\}$ is a deterministic nonnegative sequence with $\sum_k \alpha_k < \infty$, $\sum_k Z_k < \infty$ almost surely, and*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] + U_k \le (1 + \alpha_k) V_k + Z_k \quad \text{a.s.} \tag{100}$$

Then V_k converges and $\sum_k U_k < \infty$ almost surely.

C.2 Proof of Theorem 6

By definition,

$$\begin{aligned}
&\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2\\
=&\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star-\gamma_k\nabla f(\boldsymbol{\theta}_k;x_k)\|^2+\beta_k^2\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+2\beta_k\langle\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1},\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star-\gamma_k\nabla f(\boldsymbol{\theta}_k;x_k)\rangle\\
\le&(1+\beta_k)\left(\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2+\gamma_k^2\|\nabla f(\boldsymbol{\theta}_k;x_k)\|^2-2\gamma_k\langle\nabla f(\boldsymbol{\theta}_k;x_k),\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\rangle+\beta_k\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2\right). \quad (101)
\end{aligned}$$

Combining with convexity and Lemma 6 gives

$$\begin{aligned}
&\frac{\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2}{1+\beta_k}+\gamma_k\left(2-\frac{\eta_k}{c}\right)\big(f(\boldsymbol{\theta}_k;x_k)-f^\star(x_k)\big)\\
&\le\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2+\beta_k\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+2\gamma_k\big(f(\boldsymbol{\theta}^\star;x_k)-f^\star(x_k)\big), \quad (102)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k\|^2+\gamma_k\left(2\beta_k-\frac{\eta_k}{c}\right)\big(f(\boldsymbol{\theta}_k;x_k)-f^\star(x_k)\big)\\
&\le\beta_k^2\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+2\beta_k\gamma_k\big(f(\boldsymbol{\theta}_{k-1};x_k)-f^\star(x_k)\big). \quad (103)
\end{aligned}$$

For all sufficiently large k, condition (33) implies

$$\eta_k\gamma_{\min,k}\left(1+\beta_k-\frac{\eta_k}{c}\right)\ge\frac{\eta_k\gamma_{\max,k}\beta_k}{1+\beta_k}. \quad (104)$$

Adding (102) and (103), applying Lemma 5, and taking conditional expectation yields

$$\begin{aligned}
&\frac{\mathbb{E}[\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2\mid\mathcal{F}_k]}{1+\beta_k}+\mathbb{E}[\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k\|^2\mid\mathcal{F}_k]+\frac{2\eta_k\gamma_{\max,k}\beta_k}{1+\beta_k}\big(F(\boldsymbol{\theta}_k)-F(\boldsymbol{\theta}^\star)\big)\\
&\le\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2+\beta_k(1+\beta_k)\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+2\eta_k\gamma_{\max,k}\beta_k\big(F(\boldsymbol{\theta}_{k-1})-F(\boldsymbol{\theta}^\star)\big)\\
&\quad+\left(2\eta_k(1+\beta_k)(\gamma_{\max,k}-\gamma_{\min,k})+\frac{2\eta_k^2\gamma_{\min,k}}{c}\right)\sigma^2. \quad (105)
\end{aligned}$$

Consider the nonnegative Lyapunov sequence

$$V_k=\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2+(1+\beta_{k-1})\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+2\eta_{k-1}\gamma_{\max,k-1}\beta_{k-1}\big(F(\boldsymbol{\theta}_{k-1})-F(\boldsymbol{\theta}^\star)\big). \quad (106)$$

The monotonicity of $\eta_k\gamma_{\max,k}$ and β_k, together with (105), gives

$$\begin{aligned}
&\mathbb{E}[V_{k+1}\mid\mathcal{F}_k]+(1-\beta_k)\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2\\
&\le(1+\beta_k)V_k+(1+\beta_k)\left(2\eta_k(1+\beta_k)(\gamma_{\max,k}-\gamma_{\min,k})+\frac{2\eta_k^2\gamma_{\min,k}}{c}\right)\sigma^2. \quad (107)
\end{aligned}$$

Indeed, the available coefficient of $\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2$ is at least $(1+\beta_k)(1-\beta_k^2)\ge1-\beta_k$. By Lemma 5 and $\gamma_{\max,k}\to0$, $\gamma_{\max,k}-\gamma_{\min,k}$ vanishes for all sufficiently large k. Moreover, $\sum_k\eta_k^2\gamma_{\min,k}\le\sum_k\eta_k^2\gamma_{\max,k}<\infty$. Thus the error term in (107) is summable. Since $\sum_k\beta_k<\infty$ and $1-\beta_k$ is bounded away from zero, Lemma 7 implies

$$\sum_{k=1}^{\infty}\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2<\infty\quad\text{a.s.} \quad (108)$$

Returning to (102), the step-size bounds and conditional expectation give

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}^\star\|^2\mid\mathcal{F}_k]+\eta_k\gamma_{\min,k}(1+\beta_k)\left(2-\frac{\eta_k}{c}\right)\big(F(\boldsymbol{\theta}_k)-F(\boldsymbol{\theta}^\star)\big)\\
&\le(1+\beta_k)\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star\|^2+\beta_k(1+\beta_k)\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2\\
&\quad+(1+\beta_k)\left(2\eta_k(\gamma_{\max,k}-\gamma_{\min,k})+\frac{\eta_k^2\gamma_{\min,k}}{c}\right)\sigma^2. \quad (109)
\end{aligned}$$

Condition (104) implies $\eta_k/c \le 1+\beta_k$, and hence $(1+\beta_k)(2-\eta_k/c) \ge 1-\beta_k^2$ for all sufficiently large k. The error terms in (109) are summable by (108) and the same bounds used above. A second application of Lemma 7 shows that $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|^2$ converges and

$$\sum_{k=1}^{\infty} \eta_k \gamma_{\min,k} \big(F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^\star)\big) < \infty \quad \text{a.s.} \tag{110}$$

Finally, Jensen's inequality yields

$$F(\bar{\boldsymbol{\theta}}_k^{\mathrm{w}}) - F(\boldsymbol{\theta}^\star) \le \frac{\sum_{m=1}^k \eta_m \gamma_{\min,m} \big(F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)\big)}{\sum_{m=1}^k \eta_m \gamma_{\min,m}} = \mathcal{O}\left(\frac{1}{\sum_{m=1}^k \eta_m \gamma_{\max,m}}\right) \quad \text{a.s.,} \tag{111}$$

where $\sum_{m=1}^k \eta_m \gamma_{\min,m} \sim \sum_{m=1}^k \eta_m \gamma_{\max,m}$ by the step-size bounds, and

$$\bar{\boldsymbol{\theta}}_k^{\mathrm{w}} = \frac{\sum_{m=1}^k \eta_m \gamma_{\min,m} \boldsymbol{\theta}_m}{\sum_{m=1}^k \eta_m \gamma_{\min,m}}. \tag{112}$$

This proves the theorem.

C.3 Proof of Theorem 7

By smoothness,

$$\begin{aligned}
&F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_k) \\
\le& \langle \nabla F(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2 \\
=& -\gamma_k \langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k) \rangle + \langle \nabla F(\boldsymbol{\theta}_k), \beta_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}) \rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2.
\end{aligned} \tag{113}$$

We have

$$\langle \nabla F(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1} \rangle \le F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_{k-1}) + \frac{L}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|^2. \tag{114}$$

and

$$\begin{aligned}
&-\gamma_k \langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k) \rangle \\
=& -\frac{\gamma_k}{2}\|\nabla F(\boldsymbol{\theta}_k)\|^2 - \frac{\gamma_k}{2}\|\nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k)\|^2 + \frac{\gamma_k}{2}\|\nabla F(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k)\|^2 \\
\le& \frac{\eta_k \gamma_{\max,k} - \eta_k \gamma_{\min,k}}{2} \left(\|\nabla F(\boldsymbol{\theta}_k)\|^2 + \|\nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k)\|^2\right) - \eta_k \gamma_{\max,k} \langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k) \rangle.
\end{aligned} \tag{115}$$

Insert (114) and (115) to (113) and note that $F(\boldsymbol{\theta}_{k-1}) \ge F(\boldsymbol{\theta}^*)$,

$$\begin{aligned}
&F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}^*) \\
\le& (1+\beta_k)(F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^*)) + \frac{L\beta_k}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|^2 + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2 \\
&+ \frac{\eta_k \gamma_{\max,k} - \eta_k \gamma_{\min,k}}{2} \left(\|\nabla F(\boldsymbol{\theta}_k)\|^2 + \|\nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k)\|^2\right) - \eta_k \gamma_{\max,k} \langle \nabla F(\boldsymbol{\theta}_k), \nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k) \rangle.
\end{aligned} \tag{116}$$

Then taking expectation condition on $\mathcal{F}_k$ and applying Assumption 4 yields

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}^\star) \mid \mathcal{F}_k] \le& (1+\beta_k)\big(F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^\star)\big) + \frac{L\beta_k}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|^2 \\
&+ \frac{L}{2}\mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2 \mid \mathcal{F}_k] + \frac{\eta_k(\gamma_{\max,k} - \gamma_{\min,k})}{2}\delta \\
&- \left(\eta_k \gamma_{\max,k} - \frac{(1+\rho)\eta_k(\gamma_{\max,k} - \gamma_{\min,k})}{2}\right) \|\nabla F(\boldsymbol{\theta}_k)\|^2.
\end{aligned} \tag{117}$$

We also note that

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2 \le (\gamma_k^2 + \gamma_k\beta_k)\|\nabla f(\boldsymbol{\theta}_k; \mathbf{x}_k)\|^2 + (\beta_k^2 + \gamma_k\beta_k)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|^2 \tag{118}$$

Thus, again by Assumption 4,

$$\mathbb{E}[\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k\|^2 \mid \mathcal{F}_k] \le (\eta_k^2\gamma_{\max,k}^2+\eta_k\gamma_{\max,k}\beta_k)\rho\|\nabla F(\boldsymbol{\theta}_k)\|^2 + (\beta_k^2+\beta_k\eta_k\gamma_{\max,k})\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2+(\eta_k^2\gamma_{\max,k}^2+\eta_k\gamma_{\max,k}\beta_k)\delta. \quad (119)$$

Denote

$$C_k=\frac{\frac{L\beta_k}{2}(1+\beta_k+\eta_k\gamma_{\max,k})}{1+\beta_k(1-\beta_k-\eta_k\gamma_{\max,k})},$$

$$D_k=\eta_k\gamma_{\max,k}-(1+\rho)\eta_k\frac{\gamma_{\max,k}-\gamma_{\min,k}}{2}-\left(\frac{L}{2}+C_k\right)(\eta_k\gamma_{\max,k}+\beta_k)\eta_k\gamma_{\max,k}\rho,$$

$$E_k=\frac{\eta_k\gamma_{\max,k}-\eta_k\gamma_{\min,k}}{2}+\left(\frac{L}{2}+C_k\right)(\eta_k\gamma_{\max,k}+\beta_k)\eta_k\gamma_{\max,k}.$$

One may verify that C_k is decreasing under Assumption 5 and

$$\mathbb{E}[F(\boldsymbol{\theta}_{k+1})-F(\boldsymbol{\theta}^*)+C_k\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k\|^2 \mid \mathcal{F}_k]+D_k\|\nabla F(\boldsymbol{\theta}_k)\|^2 \le (1+\beta_k)\left(F(\boldsymbol{\theta}_k)-F(\boldsymbol{\theta}^*)+C_{k-1}\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|^2\right)+E_k\delta. \quad (120)$$

The error term $E_k\delta$ is summable because

$$\sum_k \eta_k^2\gamma_{\max,k}^2 \le \left(\sup_k \gamma_{\max,k}\right)\sum_k \eta_k^2\gamma_{\max,k}<\infty,$$

$$\sum_k \eta_k\gamma_{\max,k}\beta_k \le \eta_1\gamma_{\max,1}\sum_k \beta_k<\infty,$$

and $\gamma_{\max,k}-\gamma_{\min,k}$ vanishes eventually by Assumption 5 and Lemma 5. And similarly, $D_k \sim \eta_k\gamma_{\max,k}$ and $C_k, D_k>0$ for a sufficiently large k. Hence, applying Lemma 7 gives

$$\sum_{k=1}^{\infty}\eta_k\gamma_{\max,k}\|\nabla F(\boldsymbol{\theta}_k)\|^2<\infty \quad \text{a.s.} \quad (121)$$

which completes the proof.

C.4 Little-o refinement for convex objectives

Theorem 8 *Under the same conditions as Theorem 6, we have*

$$F(\bar{\boldsymbol{\theta}}_k^{\mathrm{w}})-F(\boldsymbol{\theta}^\star)=o\left(\frac{1}{\sum_{m=1}^k \eta_m\gamma_{\min,m}}\right) \quad \text{a.s.} \quad (122)$$

Proof. The proof of Theorem 6 has shown that $\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^*\|^2$ converges almost surely and $F(\bar{\boldsymbol{\theta}})\to F(\boldsymbol{\theta}^\star)$ almost surely, which leads to almost sure boundedness of $\{\boldsymbol{\theta}_k\}$. Define the tail average

$$\boldsymbol{y}_{n,k}=\frac{\sum_{m=n+1}^k \eta_m\gamma_{\min,m}\boldsymbol{\theta}_m}{\sum_{m=n+1}^k \eta_m\gamma_{\min,m}}, \qquad k>n. \quad (123)$$

Then

$$\bar{\boldsymbol{\theta}}_k^{\mathrm{w}}=\frac{\sum_{m=1}^n \eta_m\gamma_{\min,m}}{\sum_{m=1}^k \eta_m\gamma_{\min,m}}\bar{\boldsymbol{\theta}}_n^{\mathrm{w}}+\frac{\sum_{m=n+1}^k \eta_m\gamma_{\min,m}}{\sum_{m=1}^k \eta_m\gamma_{\min,m}}\boldsymbol{y}_{n,k}. \quad (124)$$

Boundedness provides a finite constant M such that $\|\bar{\boldsymbol{\theta}}_n^{\mathrm{w}}-\boldsymbol{y}_{n,k}\|\le M$ for every $k>n$. Then by smoothness and (124),

$$\left(\sum_{m=1}^k \eta_m\gamma_{\min,m}\right)\left(F(\bar{\boldsymbol{\theta}}_k^{\mathrm{w}})-F(\boldsymbol{\theta}^\star)\right) \le \left(\sum_{m=1}^k \eta_m\gamma_{\min,m}\right)\left(F(\boldsymbol{y}_{n,k})-F(\boldsymbol{\theta}^\star)\right) + M\left(\sum_{m=1}^n \eta_m\gamma_{\min,m}\right)\|\nabla F(\boldsymbol{y}_{n,k})\|+\frac{LM^2\left(\sum_{m=1}^n \eta_m\gamma_{\min,m}\right)^2}{2\sum_{m=1}^k \eta_m\gamma_{\min,m}}. \quad (125)$$

Smoothness also yields

$$\|\nabla F(\boldsymbol{y}_{n,k})\|^2 \leq 2L\left(F(\boldsymbol{y}_{n,k}) - F(\boldsymbol{\theta}^\star)\right). \tag{126}$$

Thus $\lim_{k\to\infty}\|\nabla F(\boldsymbol{y}_{n,k})\| = 0$. By convexity,

$$F(\boldsymbol{y}_{n,k}) - F(\boldsymbol{\theta}^\star) \leq \frac{\sum_{m=n+1}^{\infty}\eta_m\gamma_{\min,m}\big(F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)\big)}{\sum_{m=n+1}^{k}\eta_m\gamma_{\min,m}}. \tag{127}$$

Consequently,

$$\limsup_{k\to\infty}\left(\sum_{m=1}^{k}\eta_m\gamma_{\min,m}\right)\big(F(\bar{\boldsymbol{\theta}}_k^{\mathrm{w}}) - F(\boldsymbol{\theta}^\star)\big) \leq \sum_{m=n+1}^{\infty}\eta_m\gamma_{\min,m}\big(F(\boldsymbol{\theta}_m) - F(\boldsymbol{\theta}^\star)\big). \tag{128}$$

Letting $n\to\infty$ and by Theorem 6, the right-hand side goes to zero. Finally, since $\gamma_{\min,m} = \gamma_{\max,m}$ for all sufficiently large m and $\sum_{m=1}^{k}\eta_m\gamma_{\min,m}\to\infty$, Theorem 8 follows.

C.5 Little-o refinement for non-convex objectives

Theorem 9 *Under the same conditions as Theorem 7, we have*

$$\min_{1\leq m\leq k}\|\nabla F(\boldsymbol{\theta}_m)\|^2 = o\left(\frac{1}{\sum_{m=1}^{k}\eta_m\gamma_{\max,m}}\right) \quad \text{a.s.} \tag{129}$$

Proof. For any fixed n and $k\geq n$,

$$\begin{aligned}\left(\min_{1\leq j\leq k}\|\nabla F(\boldsymbol{\theta}_j)\|^2\right)\sum_{m=1}^{k}\eta_m\gamma_{\max,m} \leq{}& \sum_{m=n}^{\infty}\eta_m\gamma_{\max,m}\|\nabla F(\boldsymbol{\theta}_m)\|^2 \\ &+\left(\min_{1\leq j\leq k}\|\nabla F(\boldsymbol{\theta}_j)\|^2\right)\sum_{m=1}^{n}\eta_m\gamma_{\max,m}.\end{aligned} \tag{130}$$

By Theorem 7, we obtain

$$\limsup_{k\to\infty}\left(\sum_{m=1}^{k}\eta_m\gamma_{\max,m}\right)\min_{1\leq j\leq k}\|\nabla F(\boldsymbol{\theta}_j)\|^2 \leq \sum_{m=n}^{\infty}\eta_m\gamma_{\max,m}\|\nabla F(\boldsymbol{\theta}_m)\|^2 \quad \text{a.s.} \tag{131}$$

The right-hand side tends to zero as $n\to\infty$ again by Theorem 7. Thus Theorem 9 follows.

D More related works

D.1 Stochastic Polyak step sizes with momentum

Within the literature on Polyak-type step sizes for stochastic optimization, several works have explored their combination with momentum motivated by the practical benefits of momentum acceleration. Early approaches such as L4Mom (Rolinek & Martius, 2018) and ALI-G (Berrada et al., 2020) incorporated momentum empirically, while their convergence analyses do not cover the corresponding momentum variants.

More recent works have started to investigate the convergence theory of Polyak-type step sizes with momentum. Among the earliest such analyses, ALR-SMAG (Wang et al., 2023) considers stochastic moving averaged gradients (SMAG), whose update takes the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_k\sum_{m=0}^{k}\beta^{k-m}\nabla f(\boldsymbol{\theta}_m; x_m). \tag{132}$$

In SMAG, the current step size γ_k is applied to all historical gradients accumulated in the momentum direction. Consequently, SMAG and SHB are equivalent under a constant step size but generally

differ when adaptive step sizes are used, as also noted in (Wang et al., 2023). Therefore, the convergence guarantees developed for ALR-SMAG do not directly apply to SHB. Notably, Wang et al. (2023) also introduce Polyak step-size variants for SHB and evaluate them empirically, but do not establish corresponding convergence guarantees. SGDM-APS (Zhang et al., 2025) further develops a momentum-aware Polyak step size for SMAG and establishes convergence rates for both convex and non-convex objectives. Our results complement SGDM-APS (Zhang et al., 2025) by establishing convergence guarantees for Polyak step size in SHB, alongside their analysis for SMAG.

Among the existing works, MomSPS (Oikonomou & Loizou, 2025) is most closely related to ours, as it directly studies Polyak step sizes for SHB. Convergence analysis of $\text{MomSPS}_{\max}$ (Oikonomou & Loizou, 2025) focuses on convex objectives and imposes restrictions on the momentum parameter β. In comparison, our results for standard SHB-PS allow any $\beta \in [0, 1)$ in the convex setting, cover general non-convex objectives, and further establish almost sure convergence under interpolation.

D.2 Stochastic line search with momentum

SLS (Vaswani et al., 2019b) already incorporated momentum into stochastic line search as a practical acceleration technique. In particular, it considered both heavy-ball momentum and Nesterov acceleration, and demonstrated empirically that momentum can substantially accelerate convergence, while its convergence analysis is restricted to the momentum-free setting. MSL (Fan et al., 2023) further explored stochastic line search with momentum-based search directions. It introduced momentum correction and restart mechanisms to address the problem that a momentum direction may not be a descent direction for the current stochastic objective, with its main focus on empirical performance.

More recently, Lapucci and Pucci (Lapucci & Pucci, 2025; 2026) have studied convergence guarantees for stochastic line searches with more general search directions. Lapucci & Pucci (2025) characterize conditions on the search direction that ensure a well-defined stochastic line search and establish fast convergence. The subsequent work (Lapucci & Pucci, 2026) explicitly incorporates momentum-like directions into the line-search framework, using a dynamically selected momentum parameter β_k based on conjugate gradient rules, and establishes convergence guarantees under interpolation and the PL condition. In comparison, for the non-diminishing variants, our guarantees allow any fixed $\beta \in [0, 1)$ and cover strongly convex, convex, and general non-convex objectives without requiring interpolation or the PL condition. This setting is also closer to the standard use of momentum in practice, where β is typically kept fixed throughout optimization. Furthermore, under suitable step-size and momentum decay conditions, we establish almost sure convergence to the optimum or stationary beyond interpolation.

D.3 Diminishing variants and exact convergence

A number of works have studied diminishing or adaptive variants of PS and ALS to recover exact convergence beyond the interpolation regime. Most of this literature focuses on momentum-free stochastic optimization. For stochastic line-search methods, Bellavia et al. (2026) study a subsampled line-search spectral-gradient scheme with diminishing step sizes and establish almost sure convergence guarantees. For Polyak-type methods, Orvieto et al. (2022) introduce DecSPS, a diminishing variant of SPS that converges to the exact solution beyond interpolation, with convergence guarantees given in expectation. Jiang & Stich (2023) further propose AdaSPS and AdaSLS, which adapt to both interpolation and non-interpolation regimes, and additionally combine these methods with variance reduction to obtain improved expected convergence rates. Most closely related to our setting, Sebbouh et al. (2021) analyze decreasing variants of both ALS and PS for SGD and establish almost sure convergence rates for convex objectives.

For momentum methods, Oikonomou & Loizou (2025) extend decreasing Polyak step sizes to SHB through MomDecSPS and MomAdaSPS, and establish expected convergence to the exact minimizer for convex objectives without interpolation. Our results complement these works in two aspects. First, we establish almost sure convergence for diminishing PS and ALS. Second, our analysis covers not only convex objectives, but also general non-convex objectives, for which we prove almost sure convergence to stationarity. In particular, while the diminishing PS and ALS considered in our analysis are closely related to Sebbouh et al. (2021) for SGD, we extend their almost-sure perspective to the stochastic heavy ball setting and further provide guarantees for general non-convex objectives.

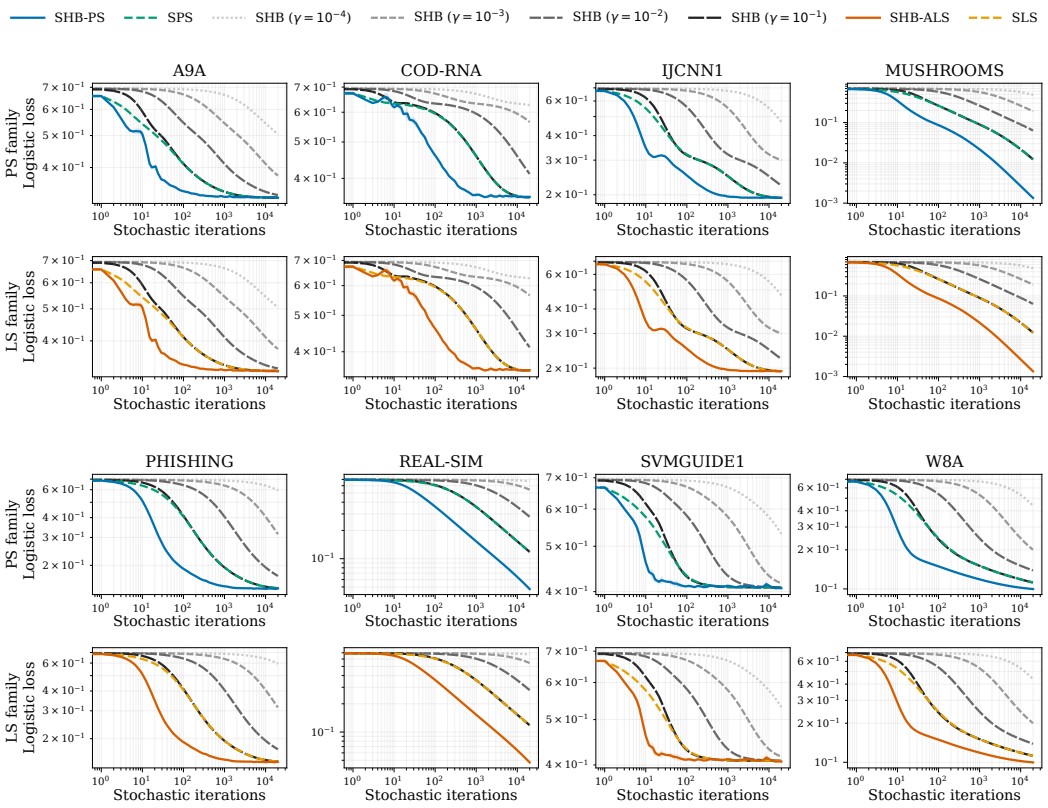


Figure 2: Full logistic training loss on eight LIBSVM datasets.

E ADDITIONAL NUMERICAL RESULTS

Here we present more results on practical machine learning tasks. We emphasize that the goal of this work is not to propose new algorithms, but to provide more general convergence guarantees for existing adaptive step-size methods with momentum. Accordingly, these experiments are intended to complement our theoretical analysis by illustrating the practical behavior of the studied methods on representative machine learning tasks, rather than to establish their empirical advantage. The curves for logistic-regression experiments on LIBSVM datasets report the mean and 95% confidence interval over five random seeds, while the deep learning experiments use three seeds.

Practical step-size rescaling. In the literature on PS and ALS, numerical evaluations often employ larger effective step sizes than those directly suggested by the theoretical analysis (Vaswani et al., 2019b; Loizou et al., 2021; Wang et al., 2023; Oikonomou & Loizou, 2025; Zhang et al., 2025). For Polyak-type methods, this can be conveniently achieved by choosing a smaller value of c (Loizou et al., 2021; Zhang et al., 2025), whereas the same adjustment is not directly applicable to ALS. We follow this practical convention in our experiments. To obtain comparable rescaling for both SHB-PS and SHB-ALS, we multiply the adaptive step size by a β-dependent factor r_β.

Hyperparameters. We use $\beta = 0.9$ and set $c = 1$ for SHB-PS and $c = 0.1$ for SHB-ALS. Both methods use the common rescaling factor $r_\beta = (1 - \sqrt{\beta})^{-2}$. We cap the step size after rescaling:

$$\gamma_k^{\text{eff}} = \min\{r_\beta \gamma_k^{\text{raw}}, \gamma_{\max}\}, \qquad \gamma_{\max} = 1, \tag{133}$$

where γ_k^{raw} denotes the adaptive step size before rescaling. The same rescaling rule and effective step-size cap are used throughout this appendix. We adopt the smooth reset strategy (Vaswani et al., 2019b; Loizou et al., 2021) with $q = 2^{2/n_b}$ for SHB-PS and $q = 2^{1/n_b}$ for SHB-ALS, where n_b denotes the number of minibatches per epoch. The initial raw step-size proposal is $\gamma_{\max}/r_\beta$, and subsequent proposals are $\min\{q\gamma_{k-1}^{\text{raw}}, \gamma_{\max}/r_\beta\}$. SHB-PS uses this proposal as an upper bound on its raw Polyak step size, while SHB-ALS uses it to initialize backtracking with decay factor $\omega = 0.9$.

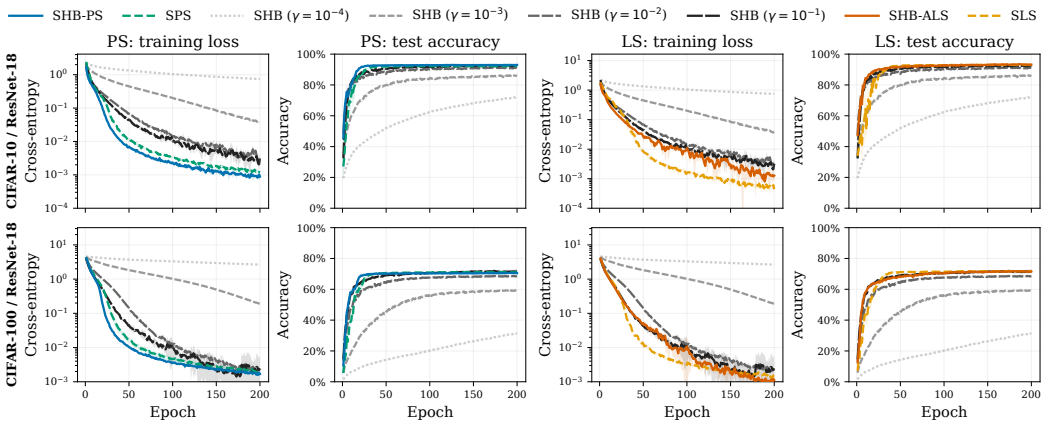


Figure 3: ResNet-18 training on CIFAR-10 (top) and CIFAR-100 (bottom). The first two columns compare the PS methods, and the last two compare the LS methods. We report training cross-entropy and test accuracy versus epochs.

Baseline methods. We mainly compare against SPS (Loizou et al., 2021) and SLS (Vaswani et al., 2019b) without momentum, as well as fixed-step SHB. For both SPS and SLS, we use $c = 0.1$ and $\gamma_{\max} = 1$. Both baselines use the same smooth reset strategy. For fixed-step SHB, we set $\beta = 0.9$ and sweep $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

E.1 Convex logistic regression on LIBSVM

Following the unregularized logistic-regression protocol of Sebbouh et al. (2021), we evaluate the methods on eight binary classification datasets from LIBSVM (Chang & Lin, 2011): `a9a`, `Cod-RNA`, `IJCNN1`, `Mushrooms`, `Phishing`, `real-sim`, `SVMguide1`, and `w8a`. Each example is normalized to have unit Euclidean norm. We use a mini-batch size of 128 and run each method for 20,000 stochastic iterations. For fixed-step SHB, we use the same four candidate step sizes across all datasets.

As shown in Figure 2, SHB-PS and SHB-ALS generally reduce the training loss faster than their momentum-free counterparts. The advantage is particularly pronounced on `Mushrooms` and `real-sim`, where the momentum variants achieve substantially lower losses within the given iteration budget. These results demonstrate the empirical benefit of incorporating momentum into adaptive step-size methods on these logistic-regression tasks.

E.2 Non-convex deep-network training

We evaluate SHB-PS and SHB-ALS on image-classification tasks using ResNet-18 (He et al., 2016) with BatchNorm on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). We train each model for 200 epochs with minibatches of size 256 and cross-entropy loss. Training images receive random crops with four pixels of padding and random horizontal flips. For these experiments, we report the mean and 95% confidence interval over three random seeds.

Figure 3 shows that SHB-PS reduces the training loss faster than SPS on both datasets and reaches high test accuracy earlier. For the line-search methods, SLS generally achieves lower training loss, particularly on CIFAR-10, while SHB-ALS attains comparable final test accuracy. Both momentum variants also achieve lower training losses than SHB with fixed step sizes. Overall, these results demonstrate the practical effectiveness of SHB-PS and SHB-ALS on deep-network training tasks, while indicating that the benefits of momentum depend on the step-size rules.