OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Noise-adaptive (Accelerated) Stochastic Heavy-Ball Momentum

Anh Dang ANH_DANG@SFU.CA
Simon Fraser University

Reza Babanezhad REZA.BH@SAMSUNG.COM
Samsung AI, Montreal

Sharan Vaswani VASWANI.SHARAN@GMAIL.COM
Simon Fraser University

Abstract

We analyze the convergence of stochastic heavy ball (SHB) momentum in the smooth, strongly-convex setting. Kidambi et al. [8] show that SHB (with small mini-batches) cannot attain an accelerated rate of convergence even for quadratics, and conjecture that the practical gain of SHB is a by-product of mini-batching. We substantiate this claim by showing that SHB can obtain an accelerated rate when the mini-batch size is larger than some threshold. In particular, for strongly-convex quadratics with condition number κ, we prove that SHB with the standard step-size and momentum parameters results in an $O\left(\exp(-T/\sqrt{\kappa}) + \sigma\right)$ convergence rate, where T is the number of iterations and σ^2 is the variance in the stochastic gradients. To ensure convergence to the minimizer, we propose a multi-stage approach that results in a noise-adaptive $O\left(\exp\left(-T/\sqrt{\kappa}\right) + \frac{\sigma}{T}\right)$ rate. For general strongly-convex functions, we use the averaging interpretation of SHB along with exponential step-sizes to prove an $O\left(\exp\left(-T/\kappa\right) + \sigma^2/T\right)$ convergence to the minimizer. Finally, we empirically demonstrate the effectiveness of the proposed algorithms.

1. Introduction

We consider the unconstrained minimization of a finite-sum objective $f : \mathbb{R}^d \to \mathbb{R}$, $\min_{w\in\mathbb{R}^d} f(w) := \frac{1}{n}\sum_{i=1}^n f_i(w)$. For supervised learning, n represents the number of training examples and f_i is the loss of example i. We denote w^* to be the unique minimizer of the above problem. We exclusively consider f to be a smooth, strongly-convex function and pay special attention to when f is a strongly-convex quadratic.

Smooth, strongly-convex quadratics: Heavy Ball (HB) momentum [16] has been extensively studied for minimizing smooth, strongly-convex quadratics in the deterministic setting. In this setting, HB results in an accelerated linear rate [16, 22]. In the stochastic setting, Kidambi et al. [8] show that SHB (with small mini-batches) cannot attain an accelerated rate of convergence even for quadratics. They conjecture that the practical gains of SHB is a by-product of mini-batching. Similarly, Paquette and Paquette [15] demonstrate that SHB with small batch-sizes cannot obtain a faster rate than stochastic gradient descent (SGD). While Loizou and Richtárik [12] prove an accelerated rate for SHB in the "$L1$ sense", this does not imply acceleration according to the standard metrics of measuring sub-optimality. Bollapragada et al. [2], Lee et al. [9] use results from random matrix theory to prove that SHB with a constant step-size and momentum can achieve an accelerated rate when the mini-batch size is sufficiently large. Compared to these works, we will use the non-asymptotic analysis standard in the optimization literature, and prove stronger worst-case results.

Contribution: Our result in Theorem 1 substantiates the claim by Kidambi et al. [8]. Specifically, we prove that for SHB with a batch-size b larger than a certain threshold b^*, using the standard constant step-size and momentum parameter can achieve an $O\left(\exp(-T/\sqrt{\kappa}) + \sigma\right)$ non-asymptotic convergence rate up to a neighborhood of the solution, where T is the number of iterations, κ is the condition number and σ^2 is the variance in the stochastic gradients. In the deterministic setting where $\sigma = 0$, we recover the optimal accelerated rate in [16, 22]. In contrast, Bollapragada et al. [2, Theorem 3.1] achieve a convergence rate of $O\left(T\, \exp(-T/\sqrt{\kappa}) + \sigma\, \log(d)\right)$ where d is the problem dimension. Hence, we obtain a faster convergence rate without an additional T dependence in the bias term, nor an additional $\log(d)$ dependence in the variance term. Our results require the batch-size to scale as $O\left(1/(1/n + 1/\kappa^2)\right)$, and hence, when $n >> O(\kappa^2)$, our results imply that SHB with a relatively small batch-size can attain an accelerated rate of convergence to a neighbourhood of the minimizer; while Bollapragada et al. [2] require a batch-size larger than $O(d\,\kappa^{3/2})$ in the worst-case. This condition is vacuous in the over-parameterized regime when $d > n$ hence Bollapragada et al. [2] require a more stringent condition on the batch-size when $d > \sqrt{\kappa}$. Lee et al. [9] provide an average-case analysis of SHB as $d, n \to \infty$, and prove an accelerated rate when $b \geq n\,\frac{\bar{\kappa}}{\sqrt{\kappa}}$ where $\bar{\kappa}$ is the average condition number. In the worst-case, $\bar{\kappa} = \kappa$, Lee et al. [9] require $b = n$.

In order to counteract the noise, SGD requires decreasing the step-size at an appropriate rate. For example, Gower et al. [6] assume the knowledge of σ^2 to set the step-size resulting in convergence to the minimizer. On the other hand, Aybat et al. [1] propose a multi-stage approach for accelerated SGD that adapts the choice of the parameters of Nesterov's accelerated gradient [14] at each stage in order to achieve the optimal rate. Importantly, the proposed algorithm is *noise-adaptive* and unlike [6], does not require the knowledge of σ^2.

Contribution: In Theorem 6, we first assume knowledge of σ^2 and prove that with a sufficiently large mini-batch, the step-size and momentum parameter of SHB can be adjusted to achieve an ϵ sub-optimality (for some $\epsilon > 0$) at an $O\left(\sqrt{\kappa}\log(1/\epsilon) + 1/\epsilon\right)$ rate. We also propose a multi-stage SHB method in Theorem 2. This SHB variant achieves a noise-adaptive accelerated convergence rate of $O\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right) + \frac{\sigma}{T}\right)$ to the minimizer.

Contribution: In Section 5, we empirically validate the effectiveness of the proposed algorithms on simple benchmarks. In particular, for strongly-convex quadratics, we consider solving a synthetic feasible linear system such that interpolation [13, 18] is satisfied. We demonstrate that SHB attains an accelerated rate when the mini-batch size is larger than a threshold, and that this threshold depends on the condition number of the problem, thus validating our theoretical results.

Smooth, strongly-convex functions: In the deterministic setting, Wang et al. [23] show an accelerated linear convergence rate for 1 dimensional problems. More recently, Goujaud et al. [5] prove that HB momentum (with any step-size or momentum parameter) cannot achieve an accelerated convergence rate on general (non-quadratic and with dimension greater than 1) smooth, strongly-convex problems. Hence, for this class of functions, HB and thus SHB can only achieve a non-accelerated linear convergence rate. For smooth, strongly-convex functions, Ghadimi et al. [4] prove a non-accelerated linear convergence rate in the deterministic setting. In the stochastic setting, Sebbouh et al. [17] use a constant step-size and momentum parameter to prove linear convergence rate of SHB to the neighborhood of the minimizer. They also show that by keeping the step-size constant for some fixed number of iterations, and then switching to a decreasing sequence, SHB can converge to the minimizer but with the non-optimal $O\left(\frac{\kappa^2}{T^2} + \frac{\sigma^2}{T}\right)$ rate. On the other hand, for SGD, Khaled and Richtárik [7], Li et al. [10], Vaswani et al. [21] propose noise-adaptive variants that result in an

$O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$ convergence rate. In particular, Li et al. [10], Vaswani et al. [21] use exponential step-sizes and the knowledge of the problem smoothness to prove the desired rate.

Contribution: In Section 4, we propose a SHB method which combines the averaging perspective of SHB [17] and the exponentially decreasing step-sizes to achieve a noise-adaptive, non-accelerated convergence rate of $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$. Importantly, the proposed algorithm provides an adaptive way to set the momentum parameters, alleviating the need to tune this additional hyper-parameter. In Appendix E, we study the effect of misestimating the smoothness on the convergence of SHB, and prove convergence to the minimizer, albeit at a slower rate.

Contribution: Our experimental results using the standard benchmarks with both the squared and logistic loss in Appendix F demonstrate that SHB with exponential step-sizes results in stable convergence to the minimizer at the same rate as SGD, and that it can be easily combined with SLS [20] or other alternative methods to estimate the smoothness constant.

2. Problem Formulation

Throughout the paper, we assume that f and each f_i are differentiable and lower-bounded by f^* and f_i^*, respectively. We also assume that each function f_i is L_i-smooth, implying that f is L-smooth with $L := \max_i L_i$. Furthermore, f is considered to be μ-strongly convex while each f_i is convex. We include definitions of these properties in Appendix A. In order to prove accelerated results, we will focus on strongly-convex quadratic objectives where $f_i(w) := \frac{1}{2}w^T A_i w - \langle d_i, w\rangle$ and $f(w) := w^T A w - \langle d, w\rangle = \frac{1}{n}\sum_1^n f_i(w)$, where A, A_i are symmetric positive definite matrices. In this case, $L = \lambda_{\max}[A]$ and $\mu = \lambda_{\min}[A]$, where $\lambda_{\max}$ and $\lambda_{\min}$ refer to the maximum and minimum eigenvalues. We define the condition number $\kappa := \frac{L}{\mu}$ and $[T] := \{0, 1, .., T\}$.

In each iteration $k \in [T]$, SHB selects a function f_{ik} (typically uniformly) at random, computes its gradient, and takes a descent step in that direction together with a momentum term. Specifically, the SHB update is given as:

$$w_{k+1} = w_k - \alpha_k \nabla f_{ik}(w_k) + \beta_k (w_k - w_{k-1}) \quad ; \quad w_{-1} = w_0 \tag{1}$$

where w_{k+1}, w_k, and w_{k-1} are the SHB iterates, and $\nabla f_{ik}(\cdot)$ is the gradient of the loss function chosen at iteration k, and $\{\alpha_k\}_{k=0}^{T-1}$ and $\{\beta_k\}_{k=0}^{T-1}$ are sequences of step-sizes and momentum parameters respectively. Each stochastic gradient $\nabla f_{ik}(w)$ is unbiased, implying that $\mathbb{E}_i\left[\nabla f_i(w)|w_k\right] = \nabla f(w)$. We will also make use of the mini-batch variant of SHB that samples a batch B of examples ($|B| = b$) in every iteration and uses it to compute the stochastic gradient. In this case, we abuse the notation $\nabla f_{ik}(w_k)$ to refer to the average stochastic gradient for the batch B_k sampled at iteration k, meaning that $\nabla f_{ik}(w_k) = \frac{1}{b}\sum_{i\in B_k} \nabla f_i(w_k)$. In the next section, we analyze the convergence of SHB for strongly-convex quadratics.

3. Strongly-convex Quadratics

Our first result substantiates the claim in Kidambi et al. [8]. Specifically, in Appendix B, we prove that for SHB with a batch-size b larger than a certain threshold b^*, using the standard constant step-size and momentum parameter can achieve an $O\left(\exp(-T/\sqrt{\kappa}) + \chi\right)$ non-asymptotic convergence rate up to a neighborhood of the solution. The proof heavily relies on the non-asymptotic result for HB in the deterministic setting, coupled with an inductive argument over the iterations.

Theorem 1 *For L-smooth, μ strongly-convex quadratics, SHB (Eq. (1)) with $\alpha_k = \alpha = \frac{a}{L}$ for $a \leq 1$, $\beta_k = \beta = \left(1 - \frac{1}{2}\sqrt{\alpha\mu}\right)^2$, batch-size b s.t. $b \geq b^* := n \max\left\{\frac{1}{1+\frac{n}{C\,\kappa^2}}, \frac{1}{1+\frac{n\,a}{3}}\right\}$ has the following convergence rate,*

$$\mathbb{E}\,\|w_T - w^*\| \leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}$$

where $\chi = \sqrt{\mathbb{E}\,\|\nabla f_i(w^*)\|^2}$ is the noise in the stochastic gradients, $\zeta = \sqrt{3\,\frac{n-b}{n\,b}}$ captures the dependence on the batch-size and $C := 3^5 2^6$ is the constant in the batch-size constraint.

We conjecture that the dependence on the constant C is loose, whereas the dependence on κ^2 in the definition of b^* can be improved to κ. The above result only implies convergence to a neighbourhood of the solution. In Theorem 6, we first assume knowledge of χ^2 and prove that with a sufficiently large mini-batch, the step-size and momentum parameter of SHB can be adjusted to achieve an ϵ sub-optimality (for some $\epsilon > 0$) at an $O\left(\sqrt{\kappa}\log(1/\epsilon) + 1/\epsilon\right)$ rate.

Multi-stage SHB: When the noise is unknown, we propose to use a multi-stage approach (Algorithm 1) similar to [1] in order to achieve convergence to the minimizer.

Algorithm 1 Multi-stage SHB

Input: T (iteration budget), b (batch-size)
  Initialization: w_0, $w_{-1} = w_0$, $k = 0$
  Set $I = \left\lfloor \frac{1}{\ln(\sqrt{2})}\,\mathcal{W}\left(\frac{T\,\ln(\sqrt{2})}{384\,\sqrt{\kappa}}\right)\right\rfloor$
  /* $\mathcal{W}(.)$ is the Lambert W function[1] */
$T_0 = \frac{T}{2}$
$T_i = \left\lceil \frac{4\,2^{i/2}\sqrt{\kappa}}{(2-\sqrt{2})}\left((i/2 + 5)\ln(2) + \ln(\sqrt{\kappa})\right)\right\rceil \;\forall i \in [1, I]$
for $i = 0$; $i \leq I$; $i = i + 1$ **do**
  Set $a_i = 2^{-i}$, $\alpha_i = \frac{a_i}{L}$, $\beta_i = \left(1 - \frac{1}{2}\sqrt{\alpha_i\mu}\right)^2$
  for $t = 1$; $t \leq T_i$; $t = t + 1$ **do**
    $w_{k+1} = w_k - \alpha_i \nabla f_{ik}(w_k) + \beta_i\,(w_k - w_{k-1})$
    Update $k = k + 1$
  end
end

In Appendix C, we theoretically analyze Algorithm 1 and prove Theorem 2. We see that Algorithm 1 achieves a convergence rate of $O\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right) + \frac{\chi}{T}\right)$ to the minimizer. In the deterministic setting when $\chi = 0$, we recover the optimal accelerated rate. This method does not require knowledge of σ and is thus noise-adaptive. We note that the choice of T_i is similar to that for Accelerated SGD using Nesterov's method [1, Theorem 3.4]. In the next section, we consider the convergence of SHB for general smooth, strongly-convex objectives.

1. The principal branch of the Lambert W function can be defined as: for $x, y \in \mathbb{R}$, $y = \mathcal{W}(x) \implies y\exp(y) = x$.

Theorem 2 *For L-smooth, μ strongly-convex quadratics with $\kappa > 1$, for $T \geq \max\left\{\frac{3 \cdot 2^{10} \kappa \sqrt{\kappa}}{\ln(2)}, \frac{3 \cdot 2^8 \, e^2 \sqrt{\kappa}}{\ln(2)}\right\}$, Algorithm 1 with batch-size b such that $b \geq b^* := n \, \max\left\{\frac{1}{1+\frac{n}{C\,\kappa^2}}, \frac{1}{1+\frac{n\,a_I}{3}}\right\}$ results in the following convergence,*

$$\mathbb{E}\,\|w_T - w^*\| \leq 6\sqrt{2}\sqrt{C_1}\sqrt{\kappa}\,\frac{1}{\sqrt{T}}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{\sqrt{C_1}}{\sqrt{T}}\,.$$

where $C_1 := \frac{2^9\,3\sqrt{\kappa\left(1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\,\sqrt{\kappa}}\right)\right)}}{\ln(2)}$ and $C := 3^5 2^6$.

4. Strongly-convex Functions

For general strongly-convex functions, we develop an SHB method that (i) converges to the minimizer at an $O\left(\exp\left(-T/\kappa\right) + \sigma^2/T\right)$ (ii) is noise-adaptive that it does not require knowledge of σ^2 and (iii) does not require manually tuning the momentum parameter. We use an equivalent form of the SHB update as Sebbouh et al. [17], showing that SHB can be interpreted as a moving average of the iterates w_k, i.e. for $z_0 = w_0$, $w_{k+1} := \frac{\lambda_{k+1}}{\lambda_{k+1}+1} w_k + \frac{1}{\lambda_{k+1}+1} z_k$, $z_k := z_{k-1} - \eta_k \nabla f_{ik}(w_k)$. In particular, for any $\{\eta_k, \lambda_k\}$ sequence, and $\alpha_k = \frac{\eta_k}{1+\lambda_{k+1}}$, $\beta_k = \frac{\lambda_k}{1+\lambda_{k+1}}$, the above update is equivalent to the SHB update in Eq. (1) [17]. The proposed SHB method combines the above interpretation of SHB [17] and the exponentially decreasing step-sizes [10, 21].

Theorem 3 *For $\tau \geq 1$, set $\eta_k = \upsilon_k \gamma_k$ where $\upsilon = \upsilon_k = \frac{1}{2L}$, $\gamma = \left(\frac{\tau}{T}\right)^{1/T}$ and $\gamma_k = \gamma^{k+1}$. Define $\lambda_k := \frac{1-2\eta L}{\eta_k \mu}\left(1 - (1-\eta_k\mu)^k\right)$. Assuming each f_i to be convex and L-smooth and f to be μ strongly-convex, SHB with the above choices of $\{\eta_k, \lambda_k\}$ results in the following convergence:*

$$\mathbb{E}\,\|w_{T-1} - w^*\|^2 \leq \frac{c_2}{c_L}\,\|w_0 - w^*\|^2 \exp\left(-\frac{T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) + \frac{32 L \sigma^2 c_2 \zeta^2 \kappa^3}{e^2\,c_L}\frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

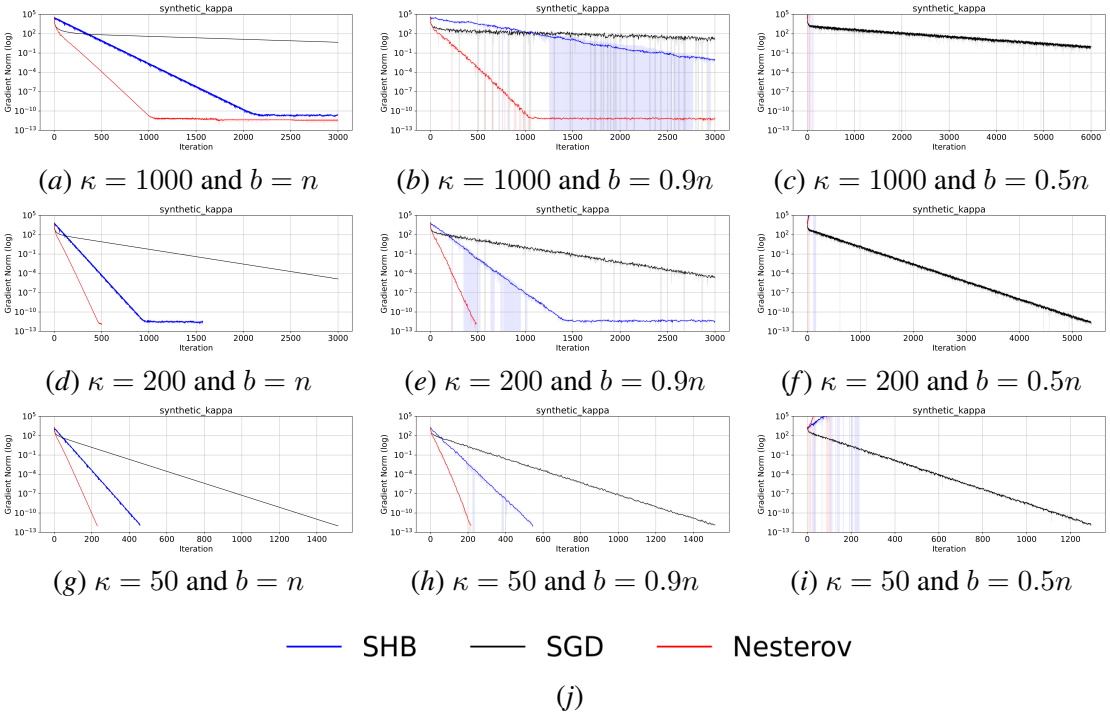
where $\zeta = \sqrt{\frac{n-b}{n\,b}}$ captures the dependence on the batch-size, $c_2 := \exp\left(\frac{1}{2\kappa}\frac{2\tau}{\ln(T/\tau)}\right)$ and $c_L := \frac{4(1-\gamma)}{\mu^2}\left[1 - \exp\left(-\frac{\mu\gamma}{2L}\right)\right]$.

The above theorem implies that SHB with exponentially decreasing step-sizes achieves an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$ convergence rate. This rate is optimal since Goujaud et al [5] show that for general strongly-convex functions, we cannot get an accelerated $O\left(\exp\left(\frac{-T}{\sqrt{\kappa}}\right)\right)$ convergence rate. We note that the momentum parameter β_k does not require tuning as it is automatically and adaptively inferred from η_k and λ_k. Hence, we effectively eliminate the need to tune the momentum parameter, one of the key hyper-parameters of SHB. Finally, we reiterate that the proposed method does not require knowledge of σ^2 and is hence noise-adaptive.

5. Experiments

For strongly-convex quadratics, we consider solving a synthetic feasible linear system such that interpolation [13, 18] is satisfied. Our experiments will be conducted on randomly generated synthetic

datasets of shape $(10000, 20)$. With the generated datasets, we can control the L-smoothness and μ-strong-convexity hence we are able to manually set different κ for each experiment. The baselines used for comparison include: SGD with constant step-size (`SGD`), Accelerated SGD using Nesterov acceleration with constant step-sizes [18] (`Nesterov`), SHB with constant step-sizes (`SHB`). For each dataset, we will run the experiments with a set of distinct batch-sizes. Each experiment will be run 5 times independently, and the average result and standard deviation will be plotted. We will use the full gradient norm as the performance measure and plot it against the number of iterations.



(a) $\kappa = 1000$ and $b = n$ (b) $\kappa = 1000$ and $b = 0.9n$ (c) $\kappa = 1000$ and $b = 0.5n$

(d) $\kappa = 200$ and $b = n$ (e) $\kappa = 200$ and $b = 0.9n$ (f) $\kappa = 200$ and $b = 0.5n$

(g) $\kappa = 50$ and $b = n$ (h) $\kappa = 50$ and $b = 0.9n$ (i) $\kappa = 50$ and $b = 0.5n$

(j)

Figure 1: Comparision of the above methods for square loss on synthetic datasets of shape $(10000, 20)$ with different κ and batch-size b. Observe that SHB attains an accelerated rate when the mini-batch size is sufficiently large.

When employing sufficiently large batch-sizes, SHB and ASGD demonstrate an accelerated convergence rate, surpassing that of the conventional SGD method. ASGD, which incorporates Nesterov acceleration parameters, performs slightly better compared to SHB. Our experiments for strongly-convex setting is deferred to Appendix F.

6. Conclusion

We showed that for strongly-convex quadratic objective functions, if the mini-batch size is above a certain threshold, then SHB can achieve an accelerated convergence rate upto a neighborhood of the minimizer. We proposed a multi-stage SHB approach that can achieve a noise-adaptive accelerated convergence rate for quadratics. For general smooth, strongly-convex functions, we developed a novel SHB algorithm that uses exponentially decreasing step-sizes and achieves an optimal noise-adaptive convergence rate.

References

[1] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32, 2019.

[2] Raghu Bollapragada, Tyler Chen, and Rachel Ward. On the fast convergence of minibatch heavy ball momentum. *arXiv preprint arXiv:2206.07553*, 2022.

[3] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[4] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[5] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method, 2023.

[6] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

[7] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world, 2020.

[8] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[9] Kiwon Lee, Andrew N. Cheng, Courtney Paquette, and Elliot Paquette. Trajectory of minibatch momentum: Batch size saturation and convergence in high dimensions, 2022.

[10] Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pages 6553–6564. PMLR, 2021.

[11] Sharon L Lohr. *Sampling: design and analysis*. CRC press, 2021.

[12] 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.

[13] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

[14] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[15] Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. *Advances in Neural Information Processing Systems*, 34:9229–9240, 2021.

[16] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[17] Othmane Sebbouh, Robert Mansel Gower, and Aaron Defazio. On the convergence of the stochastic heavy ball method. *ArXiv*, abs/2006.07867, 2020. URL https://api.semanticscholar.org/CorpusID:219687308.

[18] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[19] Sharan Vaswani, Issam Laradji, Frederik Kunstner, Si Yi Meng, Mark Schmidt, and Simon Lacoste-Julien. Adaptive gradient methods converge faster with over-parameterization (but you should do a line-search), 2021.

[20] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates, 2021.

[21] Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pages 22015–22059. PMLR, 2022.

[22] Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network. In *International Conference on Machine Learning*, pages 10816–10827. PMLR, 2021.

[23] Jun-Kun Wang, Chi-Heng Lin, Andre Wibisono, and Bin Hu. Provable acceleration of heavy ball beyond quadratics for a class of polyak-lojasiewicz functions when the non-convexity is averaged-out. In *International Conference on Machine Learning*, pages 22839–22864. PMLR, 2022.

Supplementary Material

Organization of the Appendix

Appendix A. Definitions

Our main assumptions are that each individual function f_i is differentiable, has a finite minimum f_i^*, and is L_i-smooth, meaning that for all v and w,

$$f_i(v) \leq f_i(w) + \langle \nabla f_i(w), v - w \rangle + \frac{L_i}{2} \|v - w\|^2, \qquad \text{(Individual Smoothness)}$$

which also implies that f is L-smooth, where L is the maximum smoothness constant of the individual functions. A consequence of smoothness is the following bound on the norm of the stochastic gradients,

$$\|\nabla f_i(w)\|^2 \leq 2L(f_i(w) - f_i^*).$$

We also assume that each f_i is convex, meaning that for all v and w,

$$f_i(v) \geq f_i(w) - \langle \nabla f_i(w), w - v \rangle, \qquad \text{(Convexity)}$$

Depending on the setting, we will also assume that f is μ strongly-convex, meaning that for all v and w,

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2} \|v - w\|^2, \qquad \text{(Strong Convexity)}$$

Appendix B. Proof of SHB for quadratics

Lemma 4 *For L-smooth and μ strongly-convex quadratics, SHB (Eq. (1)) with $\alpha_k = \alpha = \frac{a}{L}$ and $a \leq 1$, $\beta_k = \beta = \left(1 - \frac{1}{2}\sqrt{\alpha\mu}\right)^2$, batch-size b satisfies the following recurrence relation,*

$$\mathbb{E}[\|\Delta_T\|] \leq C_0\,\rho^T\,\|\Delta_0\| + 2aC_0\,\zeta(b)\,\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\,\mathbb{E}\,\|\Delta_k\|\right] + \frac{aC_0\,\chi\,\zeta(b)}{L}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\right],$$

where $\Delta_k := \begin{bmatrix} w_k - w^* \\ w_{k-1} - w^* \end{bmatrix}$, $C_0 \leq 3\sqrt{\frac{\kappa}{a}}$, $\zeta(b) = \sqrt{3\,\frac{n-b}{nb}}$ *and* $\rho = 1 - \frac{\sqrt{a}}{2\sqrt{\kappa}}$

Proof With the definition of SHB (1), if $\nabla f_{ik}(w)$ is the mini-batch gradient at iteration k, then, for quadratics,

$$\underbrace{\begin{bmatrix} w_{k+1} - w^* \\ w_k - w^* \end{bmatrix}}_{\Delta_{k+1}} = \underbrace{\begin{bmatrix} (1+\beta)I_d - \alpha A & -\beta I_d \\ I_d & 0 \end{bmatrix}}_{\mathcal{H}}\underbrace{\begin{bmatrix} w_k - w^* \\ w_{k-1} - w^* \end{bmatrix}}_{\Delta_k} + \alpha\underbrace{\begin{bmatrix} \nabla f(w_k) - \nabla f_{ik}(w_k) \\ 0 \end{bmatrix}}_{\delta_k}$$

$$\Delta_{k+1} = \mathcal{H}\Delta_k + \alpha\delta_k$$

Recursing from $k = 0$ to $T-1$, taking norm and expectation w.r.t to the randomness in all iterations.

$$\mathbb{E}[\|\Delta_T\|] \leq \left\|\mathcal{H}^T\Delta_0\right\| + \alpha\mathbb{E}\left[\left\|\sum_{k=0}^{T-1}\mathcal{H}^{T-1-k}\delta_k\right\|\right]$$

Using Theorem 7 and Corollary 8, for any vector v, $\left\|\mathcal{H}^k v\right\| \leq C_0\,\rho^k\,\|v\|$ where $\rho = \sqrt{\beta}$. Hence,

$$\mathbb{E}[\|\Delta_T\|] \leq C_0\,\rho^T\,\|\Delta_0\| + \frac{C_0\,a}{L}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\,\mathbb{E}\,\|\delta_k\|\right] \qquad (\alpha = \tfrac{a}{L})$$

In order to simplify δ_k, we will use the result for sampling with replacement from Lohr [11],

$$\mathbb{E}_k[\|\delta_k\|^2] = \mathbb{E}_k[\|\nabla f(w_k) - \nabla f_{ik}(w_k)\|^2] = \frac{n-b}{n\,b}\,\mathbb{E}_i\,\|\nabla f(w_k) - \nabla f_i(w_k)\|^2$$

(Sampling with replacement where b is the batch-size and n is the total number of examples)

$$= \frac{n-b}{n\,b}\,\mathbb{E}_i\,\|\nabla f(w_k) - \nabla f(w^*) - \nabla f_i(w_k) + \nabla f_i(w^*) - \nabla f_i(w^*)\|^2$$

$$(\nabla f(w^*) = 0)$$

$$\leq 3\,\frac{n-b}{n\,b}\left[\mathbb{E}_i\,\|\nabla f(w_k) - \nabla f(w^*)\|^2 + \mathbb{E}_i\,\|\nabla f_i(w_k) - \nabla f_i(w^*)\|^2 + \mathbb{E}_i\,\|\nabla f_i(w^*)\|^2\right]$$

$$(\,(a+b+c)^2 \leq 3[a^2+b^2+c^2])$$

$$\leq 3\,\frac{n-b}{n\,b}\left[L^2\,\mathbb{E}_i\,\|w_k - w^*\|^2 + L^2\,\mathbb{E}_i\,\|w_k - w^*\|^2 + \mathbb{E}_i\,\|\nabla f_i(w^*)\|^2\right]$$

(Using the L smoothness of f and f_i)

$$\leq 3\frac{n-b}{n\,b}\left[2L^2\,\|w_k - w^*\|^2 + \chi^2\right]$$

(w_k is independent of the randomness and by definition $\chi^2 = \mathbb{E}_i\,\|\nabla f_i(w^*)\|^2$)

$$\leq 3\frac{n-b}{n\,b}\left[2L^2[\|w_k - w^*\|^2 + \|w_{k-1} - w^*\|^2] + \chi^2\right] \qquad (\|w_{k-1} - w^*\|^2 \geq 0)$$

$$\implies \mathbb{E}_k[\|\delta_k\|^2] \leq 3\frac{n-b}{n\,b}\left[2L^2\,\|\Delta_k\|^2 + \chi^2\right] \qquad \text{(Definition of } \Delta_k)$$

$$\implies \mathbb{E}_k[\|\delta_k\|] \leq \underbrace{\sqrt{3\,\frac{n-b}{n\,b}}}_{:=\zeta(b)}\left[\sqrt{2L^2}\,\|\Delta_k\| + \chi\right]$$

(Taking square-roots, using Jensen's inequality on the LHS and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ on the RHS)

$$\implies \mathbb{E}_k[\|\delta_k\|] \leq \sqrt{2}L\,\zeta(b)\;\|\Delta_k\| + \zeta(b)\,\chi$$

Putting everything together,

$$\mathbb{E}[\|\Delta_T\|] \leq C_0\,\rho^T\,\|\Delta_0\| + \sqrt{2}aC_0\,\zeta(b)\,\mathbb{E}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\,\|\Delta_k\|\right] + \frac{aC_0\,\chi\,\zeta(b)}{L}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\right]$$

■

Theorem 1 *For L-smooth, μ strongly-convex quadratics, SHB (Eq. (1)) with $\alpha_k = \alpha = \frac{a}{L}$ for $a \leq 1$, $\beta_k = \beta = \left(1 - \frac{1}{2}\sqrt{\alpha\mu}\right)^2$, batch-size b s.t. $b \geq b^* := n \max\left\{\frac{1}{1+\frac{n}{C\,\kappa^2}}, \frac{1}{1+\frac{n\,a}{3}}\right\}$ has the following convergence rate,*

$$\mathbb{E}\,\|w_T - w^*\| \leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}$$

where $\chi = \sqrt{\mathbb{E}\,\|\nabla f_i(w^*)\|^2}$ is the noise in the stochastic gradients, $\zeta = \sqrt{3\,\frac{n-b}{n\,b}}$ captures the dependence on the batch-size and $C := 3^5 2^6$ is the constant in the batch-size constraint.

Proof Using Theorem 4, we have that,

$$\mathbb{E}\,\|\Delta_T\| \leq C_0\,\rho^T\,\|\Delta_0\| + \sqrt{2}aC_0\,\zeta\,\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\mathbb{E}\,\|\Delta_k\|\right] + \frac{aC_0\,\zeta\,\chi}{L}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\right]$$

We use induction to prove that for all $T \geq 1$,

$$\mathbb{E}\,\|\Delta_T\| \leq 2C_0\,\left[\rho + \sqrt{\zeta}\sqrt{a}\right]^T\,\|\Delta_0\| + \frac{2C_0\,\zeta\,a\,\chi}{L(1-\rho)}$$

where $\rho + \sqrt{\zeta}\sqrt{a} < 1$.
Base case: By Theorem 7, $C_0 \geq 1$ hence $\|\Delta_0\| \leq 2C_0\,\|\Delta_0\| + \frac{2C_0\,a\zeta\,\chi}{L(1-\rho)}$
Inductive hypothesis: For all $k \in \{0, 1, \ldots, T-1\}$, $\|\Delta_k\| \leq 2C_0\,\left[\rho + \sqrt{\zeta}\sqrt{a}\right]^k\,\|\Delta_0\| + \frac{2C_0\,a\zeta\,\chi}{L(1-\rho)}$
Inductive step: Using the above inequality,

$$\begin{aligned}
\mathbb{E}\,\|\Delta_T\| \leq& C_0\,\rho^T\,\|\Delta_0\| + \sqrt{2}aC_0\,\zeta\,\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\mathbb{E}\,\|\Delta_k\|\right] + \frac{aC_0\,\zeta\,\chi}{L}\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\right] \\
\leq& C_0\,[\rho + \sqrt{\zeta}\,\sqrt{a}]^T\,\|\Delta_0\| + \sqrt{2}aC_0\,\zeta\,\left[\sum_{k=0}^{T-1}\rho^{T-1-k}\mathbb{E}\,\|\Delta_k\|\right] + \frac{aC_0\,\zeta\,\chi}{L}\left[\sum_{k=0}^{T-1}\rho^{k}\right] \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since } \zeta, a > 0\text{)} \\
\leq& C_0\,[\rho + \sqrt{\zeta}\,\sqrt{a}]^T\,\|\Delta_0\| + \frac{\sqrt{2}aC_0\,\zeta}{\rho}\rho^T\,\left[\sum_{k=0}^{T-1}\rho^{-k}\left(2C_0\,\left[\rho + \sqrt{\zeta}\sqrt{a}\right]^k\,\|\Delta_0\| + \frac{2C_0\,a\zeta\,\chi}{L(1-\rho)}\right)\right] \\
& + \frac{aC_0\,\zeta\,\chi}{L}\frac{1-\rho^T}{1-\rho} \qquad \text{(Sum of geometric series and using the inductive hypothesis)} \\
=& C_0\,[\rho + \sqrt{\zeta}\,\sqrt{a}]^T\,\|\Delta_0\| + \frac{2\sqrt{2}\,aC_0^2\,\zeta}{\rho}\,\rho^T\,\left[\sum_{k=0}^{T-1}\left(\frac{\rho + \sqrt{\zeta}\sqrt{a}}{\rho}\right)^k\right]\,\|\Delta_0\| \\
& + \frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{\rho L(1-\rho)}\rho^T\,\left[\sum_{k=0}^{T-1}\left(\frac{1}{\rho}\right)^k\right] + \frac{aC_0\,\zeta\,\chi}{L}\frac{1-\rho^T}{1-\rho}
\end{aligned}$$

First, we need to prove that $\frac{2\sqrt{2}\,aC_0^2\,\zeta}{\rho}\,\rho^T\,\left[\sum_{k=0}^{T-1}\left(\frac{\rho+\sqrt{\zeta}\sqrt{a}}{\rho}\right)^k\right]\,\|\Delta_0\| \leq C_0\,\left[\rho+\sqrt{\zeta}\sqrt{a}\right]^T\|\Delta_0\|$.

$$\begin{aligned}\frac{2\sqrt{2}\,aC_0^2\,\zeta}{\rho}\,\rho^T\,\left[\sum_{k=0}^{T-1}\left(\frac{\rho+\sqrt{\zeta}\sqrt{a}}{\rho}\right)^k\right]\,\|\Delta_0\| &= \frac{2\sqrt{2}\,aC_0^2\,\zeta}{\rho}\,\rho^T\,\frac{\left(\frac{\rho+\sqrt{\zeta}\sqrt{a}}{\rho}\right)^T-1}{\left(\frac{\rho+\sqrt{\zeta}\sqrt{a}}{\rho}\right)-1}\,\|\Delta_0\| \\ &\qquad\text{(Sum of geometric series)} \\ &\leq 2\sqrt{2}\,\sqrt{a}C_0^2\,\sqrt{\zeta}\,\left(\rho+\sqrt{\zeta}\sqrt{a}\right)^T\,\|\Delta_0\|\end{aligned}$$

Hence, we require that,

$$2\sqrt{2}\,\sqrt{a}C_0^2\,\sqrt{\zeta} \leq C_0 \implies \zeta \leq \frac{1}{8\,C_0^2}\frac{1}{a}$$

Hence it suffices to choose ζ s.t.

$$\begin{aligned}&\implies \zeta \leq \frac{a}{3^2 2^3 \kappa}\frac{1}{a} && \text{(Since } C_0 \leq 3\sqrt{\tfrac{\kappa}{a}}\text{)} \\ &\implies \zeta \leq \frac{1}{3^2 2^3 \kappa} \\ &\implies \frac{n-b}{n\,b} \leq \frac{1}{3^5 2^6\,\kappa^2} \implies \frac{b}{n} \geq \frac{1}{1+\frac{n}{3^5 2^6\,\kappa^2}} && \text{(Using the definition of } \zeta\text{)}\end{aligned}$$

Since the batch-size b satisfies the condition that: $\frac{b}{n} \geq \frac{1}{1+\frac{n}{C\,\kappa^2}}$ for $C := 15552 = 3^5 2^6$, the above requirement is satisfied, and $\zeta \leq \frac{1}{3^2 2^3 \kappa}$.

Next, we need to show $D := \frac{2\sqrt{2}a^2C_0^2\,\zeta^2\chi}{\rho L(1-\rho)}\rho^T\left[\sum_{k=0}^{T-1}\left(\frac{1}{\rho}\right)^k\right] + \frac{aC_0\,\zeta\,\chi}{L}\frac{1-\rho^T}{1-\rho} \leq \frac{2C_0\,a\zeta\,\chi}{L(1-\rho)}$

$$\begin{aligned}D &= \frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{\rho L(1-\rho)}\rho^T\left[\sum_{k=0}^{T-1}\left(\frac{1}{\rho}\right)^k\right] + \frac{aC_0\,\zeta\,\chi}{L}\frac{1-\rho^T}{1-\rho} \\ &= \frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{\rho L(1-\rho)}\rho^T\frac{\left(\frac{1}{\rho}\right)^T-1}{\left(\frac{1}{\rho}\right)-1} + \frac{aC_0\,\zeta\,\chi}{L}\frac{1-\rho^T}{1-\rho} \\ &\qquad\text{(Sum of geometric series)} \\ &< \frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{\rho L(1-\rho)}\rho^T\frac{1-\rho^T}{1-\rho}\frac{\rho}{\rho^T} + \frac{aC_0\,\zeta\,\chi}{L(1-\rho)} \\ &< \frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{L(1-\rho)^2} + \frac{aC_0\,\zeta\,\chi}{L(1-\rho)}\end{aligned}$$

Since we want $D \leq \frac{2aC_0\,\zeta\,\chi}{L(1-\rho)}$, we require that

$$\begin{aligned}&\frac{2\sqrt{2}\,a^2C_0^2\,\zeta^2\chi}{L(1-\rho)^2} \leq \frac{aC_0\,\zeta\,\chi}{L(1-\rho)} \\ \implies &\frac{2\sqrt{2}\,C_0 a\,\zeta}{1-\rho} \leq 1\end{aligned}$$

Ensuring this imposes an additional constraint on ζ. We require ζ such that,

$$\zeta \leq \frac{1-\rho}{2\sqrt{2}\,C_0\,a} \implies \zeta \leq \frac{1}{4\sqrt{2}\,\sqrt{a}\,\sqrt{\kappa}}\frac{1}{C_0} \qquad \text{(Since } \rho = 1-\tfrac{\sqrt{a}}{2\sqrt{\kappa}}\text{)}$$

Hence it suffices to choose ζ such that,

$$\zeta \leq \frac{1}{12\sqrt{2}\kappa} \qquad \text{(Since } C_0 \leq 3\sqrt{\tfrac{\kappa}{a}}\text{)}$$

Since the condition on the batch-size ensures that $\zeta \leq \frac{1}{3^2 2^3 \kappa}$, this condition is satisfied. Hence,

$$\mathbb{E}\,\|\Delta_T\| \leq 2C_0\,\left[\rho + \sqrt{\zeta}\sqrt{a}\right]^T \|\Delta_0\| + \frac{2C_0\,a\,\zeta\,\chi}{L(1-\rho)}$$

This completes the induction.

In order to bound the noise term as $\frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}$, we will require an additional constraint on the batch-size that ensures $\zeta \leq \sqrt{a}$. Using the definition of ζ, we require that,

$$\begin{aligned}\sqrt{3\,\frac{n-b}{nb}} \leq& \sqrt{a}\\ \implies \frac{b}{n} \geq& \frac{1}{1+\frac{na}{3}}\end{aligned}$$

which is satisfied by the condition on the batch-size. From the result of the induction,

$$\begin{aligned}\mathbb{E}\,\|\Delta_T\| &\leq 2C_0\,\left[\rho + \sqrt{\zeta}\sqrt{a}\right]^T \|\Delta_0\| + \frac{2C_0\,a\,\zeta\,\chi}{L(1-\rho)}\\
&= 2C_0\,\left[1 - \frac{\sqrt{a}}{2\sqrt{\kappa}} + \sqrt{\zeta}\sqrt{a}\right]^T \|\Delta_0\| + \frac{2C_0\,a\,\zeta\,\chi}{L}\frac{2\sqrt{\kappa}}{\sqrt{a}} && (\rho = 1-\tfrac{\sqrt{a}}{2\sqrt{\kappa}})\\
&\leq 2C_0\left[1 - \frac{\sqrt{a}}{2\sqrt{\kappa}} + \frac{\sqrt{a}}{\sqrt{3^2 2^3\kappa}}\right]^T \|\Delta_0\| + \frac{2C_0\,a\,\zeta\,\chi}{L}\frac{2\sqrt{\kappa}}{\sqrt{a}} && (\zeta \leq \tfrac{1}{3^2 2^3\kappa})\\
&= 6\sqrt{\frac{\kappa}{a}}\left[1 - \frac{\sqrt{a}}{2\sqrt{\kappa}} + \frac{\sqrt{a}}{6\sqrt{2}\,\sqrt{\kappa}}\right]^T \|\Delta_0\| + \frac{2a\,\zeta\,\chi}{L}3\sqrt{\frac{\kappa}{a}}\frac{2\sqrt{\kappa}}{\sqrt{a}} && (C_0 \leq 3\sqrt{\tfrac{\kappa}{a}})\\
&\leq \frac{6}{\sqrt{a}}\sqrt{\kappa}\,\left[1 - \frac{\sqrt{a}}{4\sqrt{\kappa}}\right]^T \|\Delta_0\| + \frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}\\
& && (\tfrac{1}{6\sqrt{2}} < \tfrac{1}{4} \text{ and } \zeta \leq \sqrt{a})\\
\implies \mathbb{E}\,\|w_T - w^*\| &\leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}\\
& && (\text{for all } x,\, 1-x \leq \exp(-x))\end{aligned}$$

■

Corollary 5 *For L-smooth, μ strongly-convex quadratics, under interpolation, SHB (Eq. (1)) with $\alpha_k = \alpha = \frac{1}{L}$, $\beta_k = \beta = \left(1 - \frac{1}{2}\sqrt{\alpha\mu}\right)^2$, batch-size b s.t. $b \geq b^* := n\frac{1}{1+\frac{n}{C\,\kappa^2}}$ where $C := 3^5 2^6$ has the following convergence rate,*

$$\mathbb{E}\,\|w_T - w^*\| \leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\|$$

Proof Under interpolation $\chi = 0$. This removes the additional constraint on b^* that depends on the constant a, finishing the proof. ■

Corollary 6 *Under the same conditions of Theorem 1, for a target error $\epsilon > 0$, setting $a := \min\left\{1, (\frac{\mu}{24\,\chi})^2\epsilon\right\}$ and $T \geq \frac{4\sqrt{\kappa}}{\sqrt{a}}\,\log\left(\frac{12\sqrt{2}\,\sqrt{\kappa}\,\|w_0 - w^*\|}{\sqrt{a\epsilon}}\right)$ ensures that $\|w_T - w^*\| \leq \sqrt{\epsilon}$.*

Proof Using Theorem 1, we have that,

$$\mathbb{E}\,\|w_T - w^*\| \leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\sqrt{a}\chi}{\mu}$$

Using the step-size similar to that for SGD in [6, Theorem 3.1], we see that to get $\sqrt{\epsilon}$ accuracy first we consider $\frac{12\sqrt{a}\chi}{\mu} \leq \frac{\sqrt{\epsilon}}{2}$ that implies $a \leq (\frac{\mu}{24\,\chi})^2\epsilon$. We also need $\frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| \leq \frac{\sqrt{\epsilon}}{2}$. Taking log on both sides,

$$\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right) \leq \log\left(\frac{\sqrt{\epsilon}}{2}\,\frac{\sqrt{a}}{6\sqrt{2}\,\sqrt{\kappa}}\,\frac{1}{\|w_0 - w^*\|}\right)$$

$$\implies T \geq \frac{4\sqrt{\kappa}}{\sqrt{a}}\,\log\left(\frac{12\sqrt{2}\,\sqrt{\kappa}\,\|w_0 - w^*\|}{\sqrt{a\epsilon}}\right)$$

■

B.1. Helper Lemmas

We restate [22, Theorem 5] that we used in our proof.

Theorem 7 *Let* $H := \begin{bmatrix}(1+\beta)I_d - \alpha A & \beta I_d \\ I_d & 0\end{bmatrix} \in \mathbb{R}^{2n\times 2n}$. *Suppose that* $A \in \mathbb{R}^{n\times n}$ *is a positive semi-definite matrix. Fix a vector* $v_0 \in \mathbb{R}^n$. *If* β *is chosen to satisfy* $1 \geq \beta \geq \max\left\{\left(1-\sqrt{\alpha\lambda_{\min}(A)}\right)^2, \left(1-\sqrt{\alpha\lambda_{\max}(A)}\right)^2\right\}$ *then*

$$\left\|H^k v_0\right\| \leq \left(\sqrt{\beta}\right)^k C_0 \left\|v_0\right\|$$

where the constant

$$C_0 := \frac{\sqrt{2}(\beta+1)}{\sqrt{\min\left\{h\left(\beta, \alpha\lambda_{\min}(A)\right), h\left(\beta, \alpha\lambda_{\max}(A)\right)\right\}}} \geq 1$$

and $h(\beta, z) := -\left(\beta - (1-\sqrt{z})^2\right)\left(\beta - (1+\sqrt{z})^2\right)$

Lemma 8 *For a positive definite matrix A, denote* $\kappa := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{L}{\mu}$. *Set* $\alpha = \frac{a}{\lambda_{\max}(A)} = \frac{a}{L}$ *for* $a \leq 1$ *and* $\beta = \left(1 - \frac{1}{2}\sqrt{\alpha\lambda_{\min}(A)}\right)^2 = \left(1 - \frac{\sqrt{a}}{2\sqrt{\kappa}}\right)^2$. *Then,* $C_0 := \frac{\sqrt{2}(\beta+1)}{\sqrt{\min\{h(\beta,\alpha\lambda_{\min}(A)), h(\beta,\alpha\lambda_{\max}(A))\}}} \leq 3\sqrt{\frac{\kappa}{a}}$ *where* $h(\beta, z) := -\left(\beta - (1-\sqrt{z})^2\right)\left(\beta - (1+\sqrt{z})^2\right)$.

Proof Using the definition of $h(\beta, z)$ with the above setting for β and simplifying,

$$\begin{aligned}
h(\beta, \alpha\mu) &= 3\alpha\mu\left(1 - \frac{1}{2}\sqrt{\alpha\mu} - \frac{3}{16}\alpha\mu\right) \\
&= 3\frac{a}{\kappa}\left(1 - \frac{\sqrt{a}}{2\sqrt{\kappa}} - \frac{3a}{16\kappa}\right) && (\alpha = \tfrac{a}{L}) \\
&\geq 3\frac{a}{\kappa}\left(1 - \frac{1}{2\sqrt{\kappa}} - \frac{3}{16\kappa}\right) && (a \leq 1) \\
&\geq 3\frac{a}{\kappa}\left(1 - \frac{1}{2} - \frac{3}{16}\right) && (\kappa \geq 1) \\
&= \frac{15}{16}\frac{a}{\kappa} \\
\implies \frac{\sqrt{2}(1+\beta)}{\sqrt{h(\beta,\alpha\mu)}} &\leq \frac{2\sqrt{2}}{\sqrt{\frac{15}{16}\frac{a}{\kappa}}} = \frac{8\sqrt{2}\sqrt{\kappa}}{\sqrt{15a}} \leq 3\sqrt{\frac{\kappa}{a}} && (\beta \leq 1)
\end{aligned}$$

Now we need to bound $\frac{\sqrt{2}(1+\beta)}{\sqrt{h(\beta,\alpha L)}}$. Using the definition of $h\left(\beta, z\right)$ and simplifying,

$$\begin{aligned}
h(\beta,\alpha L) &= (2\sqrt{\alpha L} - \sqrt{\alpha\mu} - \alpha L + \frac{1}{4}\alpha\mu)(\sqrt{\alpha\mu} + 2\sqrt{\alpha L} + \alpha L - \frac{1}{4}\alpha\mu) \\
&= 4a - \frac{a}{\kappa} - 2\frac{a^{3/2}}{\sqrt{\kappa}} + \frac{1}{2}\frac{a^{3/2}}{\kappa^{3/2}} - a^2\left[1 - \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right] \\
&\qquad\qquad \text{(setting } \alpha = a/L \text{ and expanding above)} \\
&= a\left[4 - \frac{1}{\kappa} - 2\frac{a^{1/2}}{\sqrt{\kappa}} + \frac{1}{2}\frac{a^{1/2}}{\kappa^{3/2}} - a\left(1 - \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right)\right] \\
&= a\left[4 - \frac{1}{\kappa} - \sqrt{a}\left(\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}}\right) - a\left(1 - \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right)\right]
\end{aligned}$$

Since $\kappa \geq 1$, $\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}} > 0$ and $1 - \frac{1}{2\kappa} + \frac{1}{16\kappa^2} > 0$, hence

$$\begin{aligned}
h(\beta,\alpha L) &\geq a\left[4 - \frac{1}{\kappa} - \left(\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}}\right) - \left(1 - \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right)\right] \qquad (a \leq \sqrt{a} \leq 1) \\
&= a\left[4 - \left(\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}}\right) - \left(1 + \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right)\right]
\end{aligned}$$

Both $\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}}$ and $1 + \frac{1}{2\kappa} + \frac{1}{16\kappa^2}$ are decreasing functions of κ for $\kappa \geq 1$.
Hence, RHS(κ) $:= \left[4 - \left(\frac{2}{\sqrt{\kappa}} - \frac{1}{2\kappa^{3/2}}\right) - \left(1 + \frac{1}{2\kappa} + \frac{1}{16\kappa^2}\right)\right]$ is an increasing function of κ. Since, $h(\beta,\alpha L) \geq$ RHS(κ) $\geq$ RHS(1) for all $\kappa \geq 1$,

$$h(\beta,\alpha L) \geq a\left[4 - 2 + \frac{1}{2} - 1 - \frac{1}{2} - \frac{1}{16}\right] = \frac{15a}{16} \qquad (\beta \leq 1)$$

Using the above lower-bound for $\frac{\sqrt{2}(1+\beta)}{\sqrt{h(\beta,\alpha L)}}$ we have

$$\frac{\sqrt{2}(1+\beta)}{\sqrt{h(\beta,\alpha L)}} \leq \frac{8\sqrt{2}}{\sqrt{15a}} \leq \frac{3}{\sqrt{a}}$$

Putting everything together we get,

$$C_0 \leq \max\left\{3\sqrt{\frac{\kappa}{a}}, \frac{3}{\sqrt{a}}\right\} \implies C_0 \leq 3\sqrt{\frac{\kappa}{a}}$$

■

Appendix C. Proofs for multi-stage SHB

Theorem 2 *For L-smooth, μ strongly-convex quadratics with $\kappa > 1$, for $T \geq \max\left\{\frac{3\cdot 2^{10}\kappa\sqrt{\kappa}}{\ln(2)}, \frac{3\cdot 2^8\, e^2\sqrt{\kappa}}{\ln(2)}\right\}$, Algorithm 1 with batch-size b such that $b \geq b^* := n\,\max\left\{\frac{1}{1+\frac{n}{C\,\kappa^2}}, \frac{1}{1+\frac{n\,a_I}{3}}\right\}$ results in the following convergence,*

$$\mathbb{E}\,\|w_T - w^*\| \leq 6\sqrt{2}\sqrt{C_1}\sqrt{\kappa}\,\frac{1}{\sqrt{T}}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{\sqrt{C_1}}{\sqrt{T}}\,.$$

where $C_1 := \frac{2^9\,3\sqrt{\kappa\left(1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\,\sqrt{\kappa}}\right)\right)}}{\ln(2)}$ and $C := 3^5 2^6$.

Proof Stage zero consists of $T_0 = \frac{T}{2}$ iterations with $\alpha = \frac{1}{L}$ and $\beta = \left(1 - \frac{1}{2\sqrt{\kappa}}\right)^2$. Let T_i be the last iteration in stage i, $T_I = T$. Using the result of Theorem 1 with $a = 1$ for T_0 iterations in stage zero and defining $\Delta_t := w_t - w^*$,

$$\begin{aligned}
\mathbb{E}\,\|w_T - w^*\| &\leq \frac{6\sqrt{2}}{\sqrt{a}}\sqrt{\kappa}\,\exp\left(-\frac{\sqrt{a}}{4}\frac{T}{\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\sqrt{a}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}\\
\mathbb{E}\,\|\Delta_{T_0}\| &\leq 6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\chi}{\mu}\min\left\{1, \frac{\zeta}{\sqrt{a}}\right\}\\
&\leq 6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\chi}{\mu}
\end{aligned}$$

We split the remaining $\frac{T}{2}$ iterations into I stages. For stage $i \in [1, I]$, we set $\alpha_i = \frac{a_i}{L}$ and choose $a_i = 2^{-i}$. Using Theorem 1 for stage i,

$$\begin{aligned}
\mathbb{E}\,\|\Delta_{T_i}\| &\leq 6\sqrt{2}\sqrt{\frac{\kappa}{a_i}}\,\exp\left(-\frac{\sqrt{a_i}}{4}\frac{T_i}{\sqrt{\kappa}}\right)\mathbb{E}\,\|\Delta_{T_i-1}\| + \frac{12\sqrt{a_i}\chi}{\mu}\,\min\left\{1, \frac{\zeta}{\sqrt{a_i}}\right\}\\
&\leq 6\sqrt{2}\,2^{i/2}\sqrt{\kappa}\,\exp\left(-\frac{1}{4\,2^{i/2}}\frac{T_i}{\sqrt{\kappa}}\right)\mathbb{E}\,\|\Delta_{T_i-1}\| + \frac{12\,\chi}{\mu\,2^{i/2}}\\
&\leq \exp\left((i/2+5)\ln(2) + \ln(\sqrt{\kappa}) - \frac{T_i}{2^{i/2}4\,\sqrt{\kappa}}\right)\mathbb{E}\,\left\|\Delta_{T_{i-1}}\right\| + \frac{12\,\chi}{\mu\,2^{i/2}}
\end{aligned}$$

Now we want to find T_i such that $(i/2+5)\ln(2) + \ln(\sqrt{\kappa}) - \frac{T_i}{2^{i/2}4\,\sqrt{\kappa}} \leq -\frac{T_i}{2^{(i+1)/2}4\,\sqrt{\kappa}}$.

$$\begin{aligned}
&\implies T_i \geq \frac{4}{(2-\sqrt{2})}2^{i/2}\sqrt{\kappa}((i/2+5)\ln(2) + \ln(\sqrt{\kappa}))\\
&\implies T_i = \left\lceil\frac{4}{(2-\sqrt{2})}2^{i/2}\sqrt{\kappa}((i/2+5)\ln(2) + \ln(\sqrt{\kappa}))\right\rceil\\
\implies \frac{T_i}{2^{(i+1)/2}4\,\sqrt{\kappa}} &\geq \frac{1}{\sqrt{2}-1}((i/2+5)\ln(2) + \ln(\sqrt{\kappa})) \geq 2\ln(2)(i/2+5) + 2\ln(\sqrt{\kappa}) \geq (i/2+5) + \ln(\kappa)\\
&\implies \exp\left(-\frac{T_i}{2^{(i+1)/2}4\,\sqrt{\kappa}}\right) \leq \frac{1}{\kappa}\exp(-(i/2+5))
\end{aligned}$$

Define $\rho_i := \frac{1}{\kappa}\exp(-(i/2+5))$. If we unroll the above for I stages we have:

$$\begin{aligned}
\mathbb{E}[\|\Delta_{T_I}\|] &\le \prod_{i=1}^{I} \rho_i \mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{-i/2} \prod_{j=i+1}^{I} \rho_j \\
&= \exp\left(-\sum_{i=1}^{I}(i/2+5) - I\ln\kappa)\right)\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{-i/2}\exp\left(-\sum_{j=i+1}^{I}(j/2+5) - i\ln\kappa\right) \\
&\le \exp\left(-I^2/4 - I\ln\kappa\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{-i/2}\exp\left(-\sum_{j=i+1}^{I}(j/2) - i\ln\kappa\right) \\
&\le \exp\left(-I^2/4 - I\ln\kappa\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{-i/2}\exp\left(-\frac{(I-i)(I+i+1)}{4} - i\ln\kappa\right) \\
&\le \exp\left(-I^2/4 - I\ln\kappa\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{-i/2}\, 2^{\left(-\frac{(I^2-i^2+I-i)}{4}\right)}\,\exp\left(-i\ln\kappa\right) \\
&\qquad\qquad \text{(since } 2\le e\text{)} \\
&= \exp\left(-I^2/4 - I\ln\kappa\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\chi}{\mu}\sum_{i=1}^{I} 2^{\left(-\frac{I}{4}\right)}\,\exp\left(-i\ln\kappa\right) \qquad \text{(since } I^2\ge i^2\text{)} \\
&\le \exp\left(-I^2/4 - I\ln\kappa\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}} \qquad \text{(Simplifying } \textstyle\sum \frac{1}{\kappa^i}\text{)} \\
&\le \exp\left(-I^2/4\right)\,\mathbb{E}\,\|\Delta_{T_0}\| + \frac{12\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}}
\end{aligned}$$

Putting together the convergence from stage 0 and stages $[1, I]$,

$$\begin{aligned}
\mathbb{E}[\|\Delta_T\|] &\le \exp\left(-I^2/4\right)\left(6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\chi}{\mu}\right) + \frac{12\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}} \\
&\le \frac{1}{2^{\left(\frac{I}{4}\right)}}\left(6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{12\chi}{\mu}\right) + \frac{12\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}} \\
&\le \frac{1}{2^{\left(\frac{I}{4}\right)}}\left(6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\|\right) + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}} \qquad (2)
\end{aligned}$$

Now we need to bound the number of iterations in $\sum_{i=1}^{I} T_i$.

$$\begin{aligned}
\sum_{i=1}^{I} T_i &\leq \sum_{i=1}^{I} \left[\frac{4}{(2-\sqrt{2})} 2^{i/2} \sqrt{\kappa} ((i/2 + 5)\ln(2) + \ln(\sqrt{\kappa})) + 1 \right] \\
&\leq 8 \sum_{i=1}^{I} 2^{i/2} \sqrt{\kappa} ((i/2 + 5)\ln(2) + \ln(\sqrt{\kappa})) + I \\
&\leq 8\sqrt{\kappa} \sum_{i=1}^{I} 2^{i/2} \left\{ 4i + \ln(\sqrt{\kappa}) \right\} + I && \text{(For } i \geq 1\text{)} \\
&\leq 8\sqrt{\kappa} \left\{ 4I + \ln(\sqrt{\kappa}) \right\} \sum_{i=1}^{I} 2^{i/2} + I \\
&\leq 8\sqrt{\kappa} \left\{ 4I + \ln(\sqrt{\kappa}) \right\} \frac{2^{(I+1)/2}}{\sqrt{2}-1} + I \\
&\leq 16\sqrt{\kappa} \left[5I + \ln(\sqrt{\kappa}) \right] 2^{(I+1)/2}
\end{aligned}$$

Assume that $I \geq \ln(\sqrt{\kappa})$. In this case,

$$\sum_{i=1}^{I} T_i \leq 192\sqrt{\kappa}\, I\, 2^{(I/2)} \tag{3}$$

We need to set I s.t. the upper-bound on the total number of iterations in the I stages is smaller than the available budget on the iterations which is equal to $T/2$. Hence,

$$\begin{aligned}
&\frac{T}{2} \geq \frac{192}{\ln(\sqrt{2})} \sqrt{\kappa} \exp\left(\ln(\sqrt{2})\, I\right) \left(I \ln(\sqrt{2})\right) \\
\implies &\exp\left(\ln(\sqrt{2})\, I\right) \left(I \ln(\sqrt{2})\right) \leq \frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}} \implies I \ln(\sqrt{2}) \leq \mathcal{W}\left(\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}\right) \\
&&\text{(where } \mathcal{W} \text{ is the Lambert function)}
\end{aligned}$$

Hence, it suffices to set $I = \left\lfloor \frac{1}{\ln(\sqrt{2})} \mathcal{W}\left(\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}\right) \right\rfloor$. We know that,

$$\begin{aligned}
\frac{I}{2} &\geq \frac{\frac{1}{\ln(\sqrt{2})} \mathcal{W}\left(\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}\right) - 1}{2} \\
\implies \exp\left(\frac{I}{2}\right) &\geq \sqrt{1/e} \left(\exp\left(\mathcal{W}\left(\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}\right)\right)\right)^{1/(2\ln(\sqrt{2}))} \\
&= \sqrt{1/e} \left(\frac{\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}}{\mathcal{W}\left(\frac{T \ln(\sqrt{2})}{384\sqrt{\kappa}}\right)} \right)^{1/(2\ln(\sqrt{2}))} && \text{(since } \exp(\mathcal{W}(x)) = \tfrac{x}{\mathcal{W}(x)}\text{)}
\end{aligned}$$

For $x \geq e^2$, $W(x) \leq \sqrt{1+2\log^2(x)}$. Assuming $T \geq \frac{384\sqrt{\kappa}}{\ln(\sqrt{2})}e^2$ so $\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}} \geq e^2$,

$$\exp\left(\frac{I}{2}\right) \geq \sqrt{1/e}\left(\frac{\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}}{\sqrt{1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)}}\right)^{1/\ln(2)}$$

Since $2^x = (\exp(x))^{\ln(2)}$,

$$\begin{aligned}
2^{\left(\frac{I}{2}\right)} &= (\exp(I/2))^{\ln(2)} \geq (\sqrt{1/e})^{\ln(2)}\left(\frac{\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}}{\sqrt{1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)}}\right)^{\ln(2)/\ln(2)} \\
&= (\exp(I/2))^{\ln(2)} \geq (\sqrt{1/e})^{\ln(2)}\left(\frac{\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}}{\sqrt{1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)}}\right) \\
\implies \frac{1}{2^{I/2}} &\leq 2\,\frac{\sqrt{1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)}\,384\sqrt{\kappa}}{T\ln(\sqrt{2})} \\
&= \frac{2^9\,3\sqrt{\kappa\left(1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)\right)}}{\ln(2)}\frac{1}{T}
\end{aligned}$$

Define $C_1 := \frac{2^9\,3\sqrt{\kappa\left(1+2\log^2\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)\right)}}{\ln(2)}$, meaning that $\frac{1}{2^{I/2}} \leq \frac{C_1}{T}$ Using the overall convergence rate,

$$\begin{aligned}
\mathbb{E}[\|w_T - w^*\|] &\leq \frac{1}{2^{\left(\frac{I}{4}\right)}}\left(6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\|\right) + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{1}{2^{\left(\frac{I}{4}\right)}} \\
&\leq \frac{\sqrt{C_1}}{\sqrt{T}}\left(6\sqrt{2}\sqrt{\kappa}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\|\right) + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{\sqrt{C_1}}{\sqrt{T}} \\
\implies \mathbb{E}\,\|w_T - w^*\| &\leq 6\sqrt{2}\sqrt{C_1}\sqrt{\kappa}\,\frac{1}{\sqrt{T}}\,\exp\left(-\frac{T}{8\sqrt{\kappa}}\right)\|w_0 - w^*\| + \frac{24\,\chi\kappa}{\mu(\kappa-1)}\,\frac{\sqrt{C_1}}{\sqrt{T}}
\end{aligned}$$

We assumed that $I \geq \ln(\sqrt{\kappa})$ meaning that we want T s.t.

$$\left\lfloor \frac{1}{\ln(\sqrt{2})}\,\mathcal{W}\left(\frac{T\,\ln(\sqrt{2})}{384\sqrt{\kappa}}\right)\right\rfloor \geq \ln(\sqrt{\kappa})$$

Since $\mathcal{W}(x) \geq \log\left(\frac{\sqrt{4x+1}+1}{2}\right)$ for $x > 0$ we need to have:

$$\begin{aligned}
&\frac{1}{\ln(\sqrt{2})}\,\mathcal{W}\left(\frac{T\,\ln(\sqrt{2})}{384\,\sqrt{\kappa}}\right) - 1 \geq \ln\sqrt{\kappa} \\
&\implies \log\left(\frac{\sqrt{4\frac{T\,\ln(\sqrt{2})}{384\,\sqrt{\kappa}}+1}+1}{2}\right) \geq \ln\left(\sqrt{\kappa}\right)+1 \\
&\implies T \geq \frac{384\sqrt{\kappa}}{4\ln(\sqrt{2})}\left((2^{\ln(\sqrt{\kappa})+2}-2)^2-1\right)
\end{aligned}$$

Hence, it suffices to choose

$$\implies T > \frac{96\sqrt{\kappa}}{\ln(\sqrt{2})}16\cdot 2^{\ln(\kappa)}$$

Since $e > 2$

$$\implies T > \frac{96\sqrt{\kappa}}{\ln(\sqrt{2})}16\cdot e^{\ln(\kappa)} = \frac{3\cdot 2^9\kappa\sqrt{\kappa}}{\ln(\sqrt{2})}$$

Therefore to satisfy all the assumptions we need that

$$\begin{aligned}
T &\geq \max\left\{\frac{3\cdot 2^9\kappa\sqrt{\kappa}}{\ln(\sqrt{2})}, \frac{384\,\sqrt{\kappa}}{\ln(\sqrt{2})}e^2\right\} \\
&= \max\left\{\frac{3\cdot 2^{10}\kappa\sqrt{\kappa}}{\ln(2)}, \frac{3\cdot 2^8\,e^2\sqrt{\kappa}}{\ln(2)}\right\}
\end{aligned}$$

■

Appendix D. Proofs for non-accelerated rates

We will require [17, Theorem H.1]. We include its proof for completeness.

Theorem 9 *Assuming convexity and smoothness of each f_i, strongly-convex f. Suppose $(\eta_k)_k$ is a decreasing sequence such that $\eta_0 = \eta$ and $0 < \eta_k < \frac{1}{2L}$. Define $\lambda_k := \frac{1-2\eta L}{\eta_k \mu}\left(1 - (1-\eta_k\mu)^k\right)$, $A_k := \|w_k - w^* + \lambda_k(w_k - w_{k-1})\|^2$, $\mathcal{E}_k := A_k + 2\eta_k\lambda_k(f(w_{k-1}) - f(w^*))$, $\alpha_k := \frac{\eta_k}{1+\lambda_{k+1}}$, $\beta_k := \lambda_k \frac{1-\eta_k\mu}{1+\lambda_{k+1}}$, $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] \geq 0$ and i_k is a mini-batch of size b. Then SHB Eq. (1) converges as*

$$\mathbb{E}[\mathcal{E}_{k+1}] \leq (1 - \eta_k\mu)\mathbb{E}[\mathcal{E}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 \tag{4}$$

where $\zeta = \sqrt{\frac{n-b}{nb}}$.

Proof

$$\begin{aligned}
A_{k+1} &= \|w_{k+1} - w^* + \lambda_{k+1}(w_{k+1} - w_k)\|^2 \\
&= \|w_k - w^* - \alpha_k \nabla f_{ik}(w_k) + \beta_k(w_k - w_{k-1}) + \lambda_{k+1}\left[-\alpha_k\nabla f_{ik}(w_k) + \beta_k(w_k - w_{k-1})\right]\|^2 \\
&\qquad\qquad \text{(SHB step)} \\
&= \|w_k - w^* - \alpha_k(1+\lambda_{k+1})\nabla f_{ik}(w_k) + \beta_k(1+\lambda_{k+1})(w_k - w_{k-1})\|^2 \\
&= \|w_k - w^* - \eta_k \nabla f_{ik}(w_k) + \lambda_k(1-\eta_k\mu)(w_k - w_{k-1})\|^2 \\
&\qquad\qquad \text{(definition of } \alpha_k \text{ and } \beta_k) \\
&= \|w_k - w^* + \lambda_k(w_k - w_{k-1}) - \eta_k\left[\mu\lambda_k(w_k - w_{k-1}) + \nabla f_{ik}(w_k)\right]\|^2 \\
&= A_k + \eta_k^2\|\mu\lambda_k(w_k - w_{k-1}) + \nabla f_{ik}(w_k)\|^2 \\
&\quad - 2\eta_k\langle w_k - w^* + \lambda_k(w_k - w_{k-1}), \mu\lambda_k(w_k - w_{k-1}) + \nabla f_{ik}(w_k)\rangle \\
&= A_k + \eta_k^2\|\nabla f_{ik}(w_k)\|^2 + \underbrace{\eta_k^2\mu^2}_{\leq \eta_k\mu}\lambda_k^2\|w_k - w_{k-1}\|^2 \\
&\quad + 2\eta_k^2\mu\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle - 2\eta_k\mu\lambda_k\langle w_k - w^*, w_k - w_{k-1}\rangle \\
&\quad - 2\eta_k\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle - 2\eta_k\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle - 2\eta_k\mu\lambda_k^2\|w_k - w_{k-1}\|^2 \\
&\leq A_k - \eta_k\mu\left(\lambda_k^2\|w_k - w_{k-1}\|^2 + 2\lambda_k\langle w_k - w^*, w_k - w_{k-1}\rangle\right) - 2\eta_k\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle \\
&\quad + \underbrace{\eta_k^2\|\nabla f_{ik}(w_k)\|^2}_{\leq 2L\eta_k^2[f_{ik}(w_k) - f_{ik}^*]} + 2\eta_k^2\mu\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle - 2\eta_k\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle \\
&\qquad\qquad \text{(by } L\text{-smoothness of } f_{ik}) \\
\end{aligned}$$

Add $B_{k+1} = 2\eta_{k+1}\lambda_{k+1}\left(f(w_k) - f^*\right)$ on both sides

$$\begin{aligned}
A_{k+1} + B_{k+1} &\leq A_k - \eta_k\mu\left(\lambda_k^2\|w_k - w_{k-1}\|^2 + 2\lambda_k\langle w_k - w^*, w_k - w_{k-1}\rangle\right) - 2\eta_k\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle \\
&\quad + 2L\eta_k^2[f_{ik}(w_k) - f_{ik}^*] + 2\eta_k^2\mu\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle \\
&\quad - 2\eta_k\lambda_k\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle + 2\eta_{k+1}\lambda_{k+1}\left(f(w_k) - f^*\right) \\
&\leq A_k - \eta_k\mu\left(\lambda_k^2\|w_k - w_{k-1}\|^2 + 2\lambda_k\langle w_k - w^*, w_k - w_{k-1}\rangle\right) - 2\eta_k\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle \\
&\quad + 2L\eta_k^2[f_{ik}(w_k) - f_{ik}^*] - 2\eta_k\lambda_k(1-\eta_k\mu)\langle w_k - w_{k-1}, \nabla f_{ik}(w_k)\rangle + 2\eta_{k+1}\lambda_{k+1}[f(w_k) - f^*]
\end{aligned}$$

Taking expectation w.r.t i_k, $f_{ik}(w_k) - f_{ik}^* = [f_{ik}(w_k) - f_{ik}(w^*)] + [f_{ik}(w^*) - f_{ik}^*]$ then

$$\begin{aligned}\mathbb{E}[A_{k+1} + B_{k+1}] \leq & \mathbb{E}[A_k] - \mathbb{E}\left[\eta_k \mu \left(\lambda_k^2 \left\|w_k - w_{k-1}\right\|^2 + 2\lambda_k \langle w_k - w^*, w_k - w_{k-1}\rangle\right)\right] \\ & - 2\eta_k \mathbb{E}[\langle w_k - w^*, \nabla f(w_k)\rangle] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 2L\eta_k^2 \mathbb{E}[f(w_k) - f^*] \\ & - 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}[\langle w_k - w_{k-1}, \nabla f(w_k)\rangle] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}[f(w_k) - f^*]\end{aligned}$$

(Using Lemma 11)

Since f is strongly-convex, $-2\eta_k\langle w_k - w^*, \nabla f(w_k)\rangle \leq -\eta_k\mu \left\|w_k - w^*\right\|^2 - 2\eta_k[f(w_k) - f^*]$, then

$$\begin{aligned}\mathbb{E}[\mathcal{E}_{k+1}] \leq & \mathbb{E}[A_k] - \eta_k\mu\mathbb{E}[\underbrace{\left\|w_k - w^*\right\|^2 + \lambda_k^2 \left\|w_k - w_{k-1}\right\|^2 + 2\lambda_k\langle w_k - w^*, w_k - w_{k-1}\rangle}_{A_k}] + 2L\kappa\zeta^2\eta_k^2\sigma^2 \\ & + 2L\eta_k^2\mathbb{E}[f(w_k) - f^*] - 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}\left[\langle w_k - w_{k-1}, \nabla f(w_k)\rangle\right] \\ & - 2\eta_k\mathbb{E}[f(w_k) - f^*] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}[f(w_k) - f^*] \\ \leq & (1 - \eta_k\mu)\mathbb{E}[A_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 2L\eta_k^2\mathbb{E}[f(w_k) - f^*] - 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}\left[\langle w_k - w_{k-1}, \nabla f(w_k)\rangle\right] \\ & - 2\eta_k\mathbb{E}[f(w_k) - f^*] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}[f(w_k) - f^*]\end{aligned}$$

By convexity, $-\langle \nabla f(w_k), w_k - w_{k-1}\rangle \leq f(w_{k-1}) - f(w_k) = [f(w_{k-1}) - f^*] - [f(w_k) - f^*]$

$$\begin{aligned}\mathbb{E}[\mathcal{E}_{k+1}] \leq & (1 - \eta_k\mu)\mathbb{E}[A_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + \underbrace{2L\eta_k^2\mathbb{E}[f(w_k) - f^*]}_{\leq 4L\eta_k^2\mathbb{E}[f(w_k) - f^*]} + 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}[f(w_{k-1}) - f^*] \\ & - 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}[f(w_k) - f^*] - 2\eta_k\mathbb{E}[f(w_k) - f^*] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}[f(w_k) - f^*] \\ \leq & (1 - \eta_k\mu)\mathbb{E}[A_k + \underbrace{2\eta_k\lambda_k[f(w_{k-1}) - f^*]}_{B_k}] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 4L\eta_k^2\mathbb{E}[f(w_k) - f^*] \\ & - 2\eta_k\lambda_k(1 - \eta_k\mu)\mathbb{E}[f(w_k) - f^*] - 2\eta_k\mathbb{E}[f(w_k) - f^*] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}[f(w_k) - f^*] \\ \leq & (1 - \eta_k\mu)\mathbb{E}[\mathcal{E}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 2\mathbb{E}[f(w_k) - f^*]\left(2L\eta_k^2 - \eta_k\lambda_k(1 - \eta_k\mu) - \eta_k + \eta_{k+1}\lambda_{k+1}\right)\end{aligned}$$

(Theorem 9 first part)

We want to show that $2L\eta_k^2 - \eta_k\lambda_k(1-\eta_k\mu) - \eta_k + \eta_{k+1}\lambda_{k+1} \leq 0$ which is equivalent to $\eta_{k+1}\lambda_{k+1} \leq \eta_k\left(1 - 2L\eta_k + \lambda_k(1-\eta_k\mu)\right)$.

$$
\begin{aligned}
\text{RHS} &= \eta_k\left(1 - 2L\eta_k + \lambda_k(1-\eta_k\mu)\right) \\
&= \eta_k(1-2L\eta_k) + \eta_k\lambda_k(1-\eta_k\mu) \\
&= \eta_k(1-2L\eta_k) + \frac{1-2\eta L}{\mu}\left(1-(1-\eta_k\mu)^k\right)(1-\eta_k\mu) && \text{(definition of } \lambda_k\text{)} \\
&= \eta_k(1-2L\eta_k) - \frac{1-2\eta L}{\mu}\eta_k\mu + \frac{1-2\eta L}{\mu}\left(1-(1-\eta_k\mu)^{k+1}\right) \\
&= \frac{1-2\eta L}{\mu}\left(1-(1-\eta_k\mu)^{k+1}\right) + 2L\eta_k\underbrace{(\eta-\eta_k)}_{\geq 0} && \text{(since } \eta \geq \eta_k\text{)} \\
&\geq \frac{1-2\eta L}{\mu}\left(1-(1-\eta_k\mu)^{k+1}\right) \\
&\geq \frac{1-2\eta L}{\mu}\left(1-(1-\eta_{k+1}\mu)^{k+1}\right) && \text{(since } \eta_k \geq \eta_{k+1}\text{)} \\
&= \eta_{k+1}\lambda_{k+1} = \text{LHS}
\end{aligned}
$$

Hence,

$$\mathbb{E}[\mathcal{E}_{k+1}] \leq (1-\eta_k\mu)\mathbb{E}[\mathcal{E}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2$$

■

Theorem 3 *For $\tau \geq 1$, set $\eta_k = \upsilon_k\,\gamma_k$ where $\upsilon = \upsilon_k = \frac{1}{2L}$, $\gamma = \left(\frac{\tau}{T}\right)^{1/T}$ and $\gamma_k = \gamma^{k+1}$. Define $\lambda_k := \frac{1-2\eta L}{\eta_k\mu}\left(1-(1-\eta_k\mu)^k\right)$. Assuming each f_i to be convex and L-smooth and f to be μ strongly-convex, SHB with the above choices of $\{\eta_k, \lambda_k\}$ results in the following convergence:*

$$\mathbb{E}\left\|w_{T-1} - w^*\right\|^2 \leq \frac{c_2}{c_L}\left\|w_0 - w^*\right\|^2 \exp\left(-\frac{T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) + \frac{32L\sigma^2 c_2\zeta^2\kappa^3}{e^2\,c_L}\frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

where $\zeta = \sqrt{\frac{n-b}{n\,b}}$ captures the dependence on the batch-size, $c_2 := \exp\left(\frac{1}{2\kappa}\frac{2\tau}{\ln(T/\tau)}\right)$ and $c_L := \frac{4(1-\gamma)}{\mu^2}\left[1-\exp\left(-\frac{\mu\gamma}{2L}\right)\right]$.

Proof From the result of Theorem 9 we have

$$\mathbb{E}[\mathcal{E}_k] \leq (1-\eta_k\mu)\mathbb{E}[\mathcal{E}_{k-1}] + 2L\kappa\zeta^2\eta_k^2\sigma^2$$

Unrolling the recursion starting from w_0 and using the exponential step-sizes γ_k

$$\begin{aligned}\mathbb{E}[\mathcal{E}_T] &\leq \mathbb{E}[\mathcal{E}_0] \prod_{k=1}^{T} \left(1 - \frac{\mu\gamma^k}{2L}\right) + 2L\kappa\zeta^2\sigma^2 \sum_{k=1}^{T} \left[\prod_{i=k+1}^{T} \gamma^{2k} \left(1 - \frac{\mu\gamma^i}{2L}\right)\right] \\ &\leq \|w_0 - w^*\|^2 \exp\left(\frac{-\mu}{2L} \underbrace{\sum_{k=1}^{T} \gamma^k}_{:=C}\right) + 2L\kappa\zeta^2\sigma^2 \underbrace{\sum_{k=1}^{T} \gamma^{2k} \exp\left(-\frac{\mu}{2L} \sum_{i=k+1}^{T} \gamma^i\right)}_{:=D} \\ &\qquad\qquad (\lambda_0 = 0 \text{ and } 1 - x < \exp(-x))\end{aligned}$$

Using Lemma 12 to lower-bound C then the first term can be bounded as

$$\|w_0 - w^*\|^2 \exp\left(\frac{-\mu}{2L} C\right) \leq \|w_0 - w^*\|^2 c_2 \exp\left(-\frac{T}{2\kappa} \frac{\gamma}{\ln(T/\tau)}\right)$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{2\kappa} \frac{2\tau}{\ln(T/\tau)}\right)$. Using Lemma 13 to upper-bound D, we have $D \leq \frac{16\kappa^2 c_2 (\ln(T/\tau))^2}{e^2\gamma^2 T}$ then the second term can be bounded as

$$2L\kappa\zeta^2\sigma^2 D \leq \frac{32L\sigma^2 c_2 \zeta^2 \kappa^3}{e^2} \frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

Hence

$$\mathbb{E}[\mathcal{E}_T] \leq \|w_0 - w^*\|^2 c_2 \exp\left(-\frac{T}{2\kappa} \frac{\gamma}{\ln(T/\tau)}\right) + \frac{32L\sigma^2 c_2 \zeta^2 \kappa^3}{e^2} \frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

By Theorem 10, then

$$\mathbb{E}\,\|w_{T-1} - w^*\|^2 \leq \frac{c_2}{c_L}\, \|w_0 - w^*\|^2 \exp\left(-\frac{T}{2\kappa} \frac{\gamma}{\ln(T/\tau)}\right) + \frac{32L\sigma^2 c_2 \zeta^2 \kappa^3}{e^2\, c_L} \frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

■

D.1. Helper Lemmas

Lemma 10 *For* $\mathcal{E}_T := \|w_T - w^* + \lambda_T(w_T - w_{T-1})\|^2 + 2\eta_T\lambda_T(f(w_{t-1}) - f(w^*))$, $\mathcal{E}_T \geq c_L \|w_{T-1} - w^*\|^2$ *where* $c_L = \frac{4(1-\gamma)}{\mu^2}\left[1 - \exp\left(-\frac{\mu\gamma}{2L}\right)\right]$

Proof

$$\mathbb{E}[\mathcal{E}_T] = \mathbb{E}[A_T] + \mathbb{E}[B_T] \geq \mathbb{E}[B_T] = 2\lambda_T\eta_T\, \mathbb{E}[f(w_{T-1}) - f^*]$$

Hence, we want to lower-bound $\lambda_T\,\eta_T$ and we do this next

$$\begin{aligned}
\lambda_T\,\eta_T &= \frac{1-\gamma}{\mu}\,\left[1-\left(1-\frac{\mu\,\gamma^{T+1}}{2L}\right)^T\right] && \text{(Using the definition of } \eta_k \text{ and } \lambda_k\text{)}\\
&\geq \frac{1-\gamma}{\mu}\,\left[1-\exp\left(-T\,\gamma^T\,\frac{\mu\,\gamma}{2L}\right)\right] && \text{(Since } 1-x\leq\exp(-x)\text{)}\\
&= \frac{1-\gamma}{\mu}\,\left[1-\exp\left(-\frac{\mu\,\gamma}{2L}\right)\right] && \text{(Since } \gamma=\left(\tfrac{1}{T}\right)^{1/T}\text{)}
\end{aligned}$$

Putting everything together, and using strong-convexity of f

$$\mathbb{E}[\mathcal{E}_T] \geq \underbrace{\frac{4(1-\gamma)}{\mu^2}\,\left[1-\exp\left(-\frac{\mu\,\gamma}{2L}\right)\right]}_{:=c_L}\,\mathbb{E}\,\|w_{T-1}-w^*\|^2$$

■

We restate [21, Lemma 2, Lemma 5, and Lemma 6] that we used in our proof.

Lemma 11 *If*

$$\sigma^2 := \mathbb{E}[f_i(w^*)-f_i^*],$$

and each function f_i is μ strongly-convex and L-smooth, then

$$\sigma_B^2 := \mathbb{E}_{\mathcal{B}}[f_{\mathcal{B}}(w^*)-f_{\mathcal{B}}^*] \leq \kappa\,\underbrace{\frac{n-b}{nb}}_{:=\zeta^2}\sigma^2.$$

Lemma 12

$$A := \sum_{t=1}^{T}\gamma^t \geq \frac{\gamma T}{\ln(T/\tau)} - \frac{2\tau}{\ln(T/\tau)}$$

Lemma 13 *For $\gamma=\left(\frac{\tau}{T}\right)^{1/T}$ and any $\kappa>0$,*

$$\sum_{k=1}^{T}\gamma^{2k}\exp\left(-\frac{1}{\kappa}\sum_{i=k+1}^{T}\gamma^i\right) \leq \frac{4\kappa^2 c_2(\ln(T/\tau))^2}{e^2\gamma^2 T}$$

where $c_2=\exp\left(\frac{1}{\kappa}\frac{2\tau}{\ln(T/\tau)}\right)$

Appendix E. Proof of SHB with smoothness mis-estimation

Similar to the dependence of SGD on smoothness mis-estimation obtained by [21], Theorem 14 shows that with any mis-estimation on L we can still recover the convergence rate of $O\left(\exp\left(\frac{-T}{\kappa}\right)+\frac{\sigma^2}{T}\right)$ to the minimizer w^*. When $\nu_L \leq 1$, $\ln(\nu_L) \leq 0$, the last term will be zero which implies that the rate matches Theorem 3 up to constant that depends on ν_L. When $\nu_L > 1$, we pay a price of mis-estimation of the unknown smoothness as $\ln(\nu_L) > 0$ so the last term is non-zero and the convergence rate slows down by a factor that depends on ν_L.

Theorem 14 *Under the same settings as Theorem 3, SHB with the estimated $\hat{L} = \frac{L}{\nu_L}$ results in the following convergence,*

$$\begin{aligned}
&\mathbb{E}\left\|w_{T-1}-w^*\right\|^2 \\
&\leq \left\|w_0-w^*\right\|^2 \frac{c_2}{c_L}\exp\left(-\frac{\min\{\nu_L,1\}T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) \\
&+\frac{c_2}{c_L}\frac{32L\kappa^3\zeta^2\ln(T/\tau)}{e^2\gamma^2 T}\times\left[\left(\max\left\{1,\frac{\nu_L^2}{4L}\right\}\ln(T/\tau)\sigma^2\right)\right. \\
&\left.+\left(\max\{0,\ln(\nu_L)\}\left(\sigma^2+2\Delta_f\frac{\nu_L-1}{\nu_L\kappa}\right)\right)\right]
\end{aligned}$$

where $c_2 = \exp\left(\frac{1}{2\kappa}\frac{2\tau}{ln(T/\tau)}\right)$, $k_0 = T\frac{\ln(\nu_L)}{\ln(T/\tau)}$, *and* $\Delta_f = \max_{i\in[k_0]}\mathbb{E}[f(w_i)-f^*]$ *and* $c_L = \frac{4(1-\gamma)}{\mu^2}\left[1-\exp\left(-\frac{\mu\gamma}{2L}\right)\right]$

Proof Suppose we estimate L to be $\hat{L}$. Now redefine

$$\begin{aligned}
\eta_k &= \frac{1}{2\hat{L}}\gamma_k \\
\hat{\lambda_k} &= \frac{1-2\eta\hat{L}}{\eta_k\mu}\left(1-(1-\eta_k\mu)^k\right) \\
\hat{A}_k &= \left\|w_k-w^*+\hat{\lambda_k}(w_k-w_{k-1})\right\|^2 \\
\hat{B}_k &= 2\eta_k\hat{\lambda_k}(f(w_{k-1})-f(w^*)) \\
\hat{\mathcal{E}}_k &= \hat{A}_k+\hat{B}_k
\end{aligned}$$

Follow the proof of Theorem 9 until Theorem 9 first part step with the new definition,

$$\mathbb{E}[\hat{\mathcal{E}}_{k+1}] \leq (1-\eta_k\mu)\mathbb{E}[\hat{\mathcal{E}}_k]+2L\kappa\zeta^2\eta_k^2\sigma^2+2\mathbb{E}[f(w_k)-f^*]\underbrace{\left(2L\eta_k^2-\eta_k\hat{\lambda_k}(1-\eta_k\mu)-\eta_k+\eta_{k+1}\hat{\lambda_{k+1}}\right)}_{G} \tag{5}$$

G can be bound as

$$
\begin{aligned}
G &= 2L\eta_k^2 - \eta_k\hat{\lambda_k}(1-\eta_k\mu) - \eta_k + \eta_{k+1}\hat{\lambda_{k+1}} \\
&= \eta_k(2L\eta_k - 1) - \eta_k\hat{\lambda_k}(1-\eta_k\mu) + \eta_{k+1}\hat{\lambda_{k+1}} \\
&= \eta_k(2L\eta_k - 1) + \eta_k(1-2\hat{L}\eta) - \frac{1-2\eta\hat{L}}{\mu}\left(1-(1-\eta_k\mu)^{k+1}\right) + \eta_{k+1}\hat{\lambda_{k+1}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{definition of } \hat{\lambda_k}) \\
&\leq 2\eta_k(L\eta_k - \hat{L}\eta) - \frac{1-2\eta\hat{L}}{\mu}\left(1-(1-\eta_{k+1}\mu)^{k+1}\right) + \eta_{k+1}\hat{\lambda_{k+1}} \qquad (\eta_{k+1} \leq \eta_k) \\
&= 2\eta_k(L\eta_k - \hat{L}\eta) - \eta_{k+1}\hat{\lambda_{k+1}} + \eta_{k+1}\hat{\lambda_{k+1}} \\
&= 2\eta_k(L\eta_k - \hat{L}\eta)
\end{aligned}
$$

Hence Eq. (5) can be written as

$$\mathbb{E}[\hat{\mathcal{E}}_{k+1}] \leq (1-\eta_k\mu)\mathbb{E}[\hat{\mathcal{E}}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 4\mathbb{E}[f(w_k)-f^*]\eta_k(L\eta_k - \hat{L}\eta)$$

First case if $\nu_L \leq 1$ then $L\eta_k - \hat{L}\eta \leq 0$ and we will recover the proof of Theorem 3 with a slight difference including ν_L.

$$\mathbb{E}[\hat{\mathcal{E}}_k] \leq \|w_0 - w^*\|^2 \, c_2 \exp\left(-\frac{\nu_L T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) + \frac{32L\kappa\zeta^2\sigma^2 c_2\kappa^2}{e^2}\frac{(\ln(T/\tau))^2}{\gamma^2 T}$$

Second case if $\nu_L > 1$
Let $k_0 = T\frac{\ln(\nu_L)}{\ln(T/\tau)}$ then for $k < k_0$ regime, $L\eta_k - \hat{L}\eta > 0$

$$\mathbb{E}[\hat{\mathcal{E}}_{k+1}] \leq (1-\eta_k\mu)\mathbb{E}[\hat{\mathcal{E}}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 4\mathbb{E}[f(w_k)-f^*]\eta_k(L\eta_k - \hat{L}\eta)$$

Let $\Delta_f = \max_{i\in[k_0]}\mathbb{E}[f(w_i)-f^*]$ and observe that $L\eta_k - \hat{L}\eta \leq L\eta_k\frac{\nu_L - 1}{\nu_L}$ then

$$
\begin{aligned}
\mathbb{E}[\hat{\mathcal{E}}_{k+1}] &\leq (1-\eta_k\mu)\mathbb{E}[\hat{\mathcal{E}}_k] + 2L\kappa\zeta^2\eta_k^2\sigma^2 + 4L\eta_k^2\Delta_f\frac{\nu_L-1}{\nu_L} \\
&= (1-\frac{\mu\nu_L}{2L}\gamma^k)\mathbb{E}[\hat{\mathcal{E}}_k] + \underbrace{2L(\kappa\zeta^2\sigma^2 + 2\Delta_f\frac{\nu_L-1}{\nu_L})}_{c_5}\eta_k^2
\end{aligned}
$$

Since $\nu_L > 1$

$$\mathbb{E}[\hat{\mathcal{E}}_{k+1}] \leq (1-\frac{\mu}{2L}\gamma^k)\mathbb{E}[\hat{\mathcal{E}}_{k+1}] + c_5\eta_k^2$$

Unrolling the recursion for the first k_0 iterations we get

$$\mathbb{E}[\hat{\mathcal{E}}_{k_0}] \leq \mathbb{E}[\hat{\mathcal{E}}_0]\prod_{k=1}^{k_0-1}\left(1-\frac{\mu}{2L}\gamma^k\right) + c_5\sum_{k=1}^{k_0-1}\gamma_k^2\prod_{i=k+1}^{k_0-1}\left(1-\frac{\mu}{2L}\gamma_i\right)$$

Bounding the first term using Lemma 12,

$$\prod_{k=1}^{k_0-1}\left(1-\frac{\mu}{2L}\gamma^k\right) \le \exp\left(-\frac{\mu}{2L}\frac{\gamma-\gamma^{k_0}}{1-\gamma}\right)$$

Bounding the second term using Lemma 13 similar to [21, Section C3]

$$\sum_{k=1}^{k_0-1}\gamma_k^2\prod_{i=k+1}^{k_0-1}\left(1-\frac{\mu}{2L}\gamma_i\right) \le \exp\left(\frac{\gamma^{k_0}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{k_0\ln(T/\tau)^2}{T^2}$$

Put everything together,

$$\mathbb{E}[\hat{\mathcal{E}}_{k_0}] \le \|w_0-w^*\|^2\exp\left(-\frac{\mu}{2L}\frac{\gamma-\gamma^{k_0}}{1-\gamma}\right) + c_5\exp\left(\frac{\gamma^{k_0}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{k_0\ln(T/\tau)^2}{T^2}$$

Now consider the regime $k \ge k_0$ where $L\eta_k - \hat{L}\eta \le 0$

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{E}}_{k+1}] &\le (1-\frac{\mu}{2L}\gamma^k)\mathbb{E}[\hat{\mathcal{E}}_k] + 2L\kappa\zeta^2\sigma^2\frac{\nu_L^2}{4L}\gamma_k^2\\ &\le (1-\frac{\mu}{2L}\gamma^k)\mathbb{E}[\hat{\mathcal{E}}_k] + \frac{\nu_L^2\sigma^2}{2L}\gamma_k^2\end{aligned}$$

Unrolling the recursion from $k=k_0$ to T

$$\mathbb{E}[\hat{\mathcal{E}}_T] \le \mathbb{E}[\hat{\mathcal{E}}_{k_0}]\prod_{k=k_0}^{T}(1-\frac{\mu}{2L}\gamma_k) + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{2L}\sum_{k=k_0}^{T}\gamma_k^2\prod_{i=k+1}^{T}(1-\frac{\mu}{L}\gamma_i)$$

Bounding the first term using Lemma 12,

$$\prod_{k=k_0}^{T}\left(1-\frac{\mu}{2L}\gamma^k\right) \le \exp\left(-\frac{\mu}{2L}\frac{\gamma^{k_0}-\gamma^{T+1}}{1-\gamma}\right)$$

Bounding the second term using Lemma 13 similar to [21, Section C3]

$$\sum_{k=k_0}^{T}\gamma_k^2\prod_{i=k+1}^{T}\left(1-\frac{\mu}{2L}\gamma_i\right) \le \exp\left(\frac{\gamma^{T+1}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{(T-k_0+1)\ln(T/\tau)^2}{T^2}$$

Hence, put everything together

$$\mathbb{E}[\hat{\mathcal{E}}_T] \le \mathbb{E}[\hat{\mathcal{E}}_{k_0}]\exp\left(-\frac{\mu}{2L}\frac{\gamma^{k_0}-\gamma^{T+1}}{1-\gamma}\right) + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{2L}\exp\left(\frac{\gamma^{T+1}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{(T-k_0+1)\ln(T/\tau)^2}{T^2}$$

Combining the bounds for two regimes

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{E}}_T] &\le \exp\left(-\frac{\mu}{2L}\frac{\gamma^{k_0}-\gamma^{T+1}}{1-\gamma}\right)\left(\|w_0-w^*\|^2\exp\left(-\frac{\mu}{2L}\frac{\gamma-\gamma^{k_0}}{1-\gamma}\right) + c_5\exp\left(\frac{\gamma^{k_0}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{k_0\ln(T/\tau)^2}{T^2}\right)\\ &\quad + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{2L}\exp\left(\frac{\gamma^{T+1}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{(T-k_0+1)\ln(T/\tau)^2}{T^2}\\ &= \|w_0-w^*\|^2\exp\left(-\frac{\mu}{2L}\frac{\gamma-\gamma^{T+1}}{1-\gamma}\right) + c_5\exp\left(\frac{\gamma^{T+1}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{k_0\ln(T/\tau)^2}{T^2}\\ &\quad + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{2L}\exp\left(\frac{\gamma^{T+1}}{2\kappa(1-\gamma)}\right)\frac{16\kappa^2}{e^2\gamma^2}\frac{(T-k_0+1)\ln(T/\tau)^2}{T^2}\end{aligned}$$

Using Lemma 12 to bound the first term and noting that $\frac{\gamma^{T+1}}{1-\gamma} \leq \frac{2\tau}{\ln(T/\tau)}$, let $c_2 = \exp\left(\frac{1}{2\kappa}\frac{2\tau}{ln(T/\tau)}\right)$

$$\mathbb{E}[\hat{\mathcal{E}}_T] \leq \|w_0 - w^*\|^2 \exp\left(-\frac{T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) + c_5 \frac{16c_2\kappa^2}{e^2\gamma^2}\frac{k_0 \ln(T/\tau)^2}{T^2} + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{2L}\frac{16c_2\kappa^2}{e^2\gamma^2}\frac{(T-k_0+1)\ln(T/\tau)^2}{T^2}$$

Substitute the value of c_5 and k_0 we have

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{E}}_T] \leq \|w_0 - w^*\|^2 \exp\left(-\frac{T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) + \frac{\nu_L^2\kappa\zeta^2\sigma^2}{LT}\frac{8c_2\kappa^2\ln(T/\tau)^2}{e^2\gamma^2} \\ + 32\left(\kappa\zeta^2\sigma^2 + 2\Delta_f\frac{\nu_L - 1}{\nu_L}\right)\frac{L}{T}\frac{c_2\kappa^2\ln(\nu_L)\ln(T/\tau)}{e^2\gamma^2}\end{aligned}$$

Combining the statements from $\nu_L \leq 1$ and $\nu_L > 1$ gives us

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{E}}_T] \leq \|w_0 - w^*\|^2 c_2 \exp\left(-\frac{\min\{\nu_L, 1\}T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) \\ + \frac{32Lc_2\kappa^2\ln(T/\tau)}{e^2\gamma^2 T}\left(\max\left\{1, \frac{\nu_L^2}{4L}\right\}\ln(T/\tau)\kappa\zeta^2\sigma^2 + \max\{0, \ln(\nu_L)\}\left(\kappa\zeta^2\sigma^2 + 2\Delta_f\frac{\nu_L - 1}{\nu_L}\right)\right)\end{aligned}$$

The next step is to remove the $\hat{L}$ from the LHS, and obtain a better measure of sub-optimality. By Theorem 10,

$$\mathbb{E}[\hat{\mathcal{E}}_T] \geq \underbrace{\frac{4(1-\gamma)}{\mu^2}\left[1 - \exp\left(-\frac{\mu\gamma}{2L}\right)\right]}_{:=c_L}\|w_{T-1} - w^*\|^2$$

Note that $c_L > 0$ is constant w.r.t T. Hence,

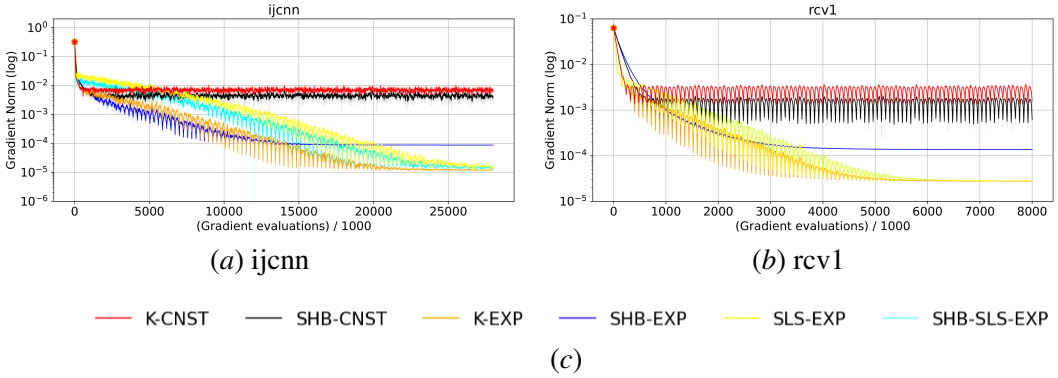
$$\begin{aligned}\mathbb{E}\|w_{T-1} - w^*\|^2 \leq \|w_0 - w^*\|^2 \frac{c_2}{c_L}\exp\left(-\frac{\min\{\nu_L, 1\}T}{2\kappa}\frac{\gamma}{\ln(T/\tau)}\right) \\ + \frac{c_2}{c_L}\frac{32L\kappa^2\ln(T/\tau)}{e^2\gamma^2 T}\left(\max\left\{1, \frac{\nu_L^2}{4L}\right\}\ln(T/\tau)\kappa\zeta^2\sigma^2 + \max\{0, \ln(\nu_L)\}\left(\kappa\zeta^2\sigma^2 + 2\Delta_f\frac{\nu_L - 1}{\nu_L}\right)\right)\end{aligned}$$

■

Appendix F. Additional experiments

To conduct experiments for smooth, strongly-convex functions, we adopt the settings from [21]. Our experiment involves the SHB variant and other commonly used optimization methods. The comparison will be based on two common supervised learning losses, squared loss for regression tasks and logistic loss for classification. We will utilize a linear model with ℓ_2-regularization $\frac{\lambda}{2}\|w\|^2$ in which $\lambda = 0.01$. To assess the performance of the optimization methods, we use *ijcnn* and *rcv1* data sets from LIBSVM [3]. For each dataset, the training iterations will be fixed at $T = 100n$, where n is the number of samples in the training dataset, and we will use a batch-size of 100. To ensure statistical significance, each experiment will be run 5 times independently, and the average result and standard deviation will be plotted. We will use the full gradient norm as the performance measure and plot it against the number of gradient evaluations.

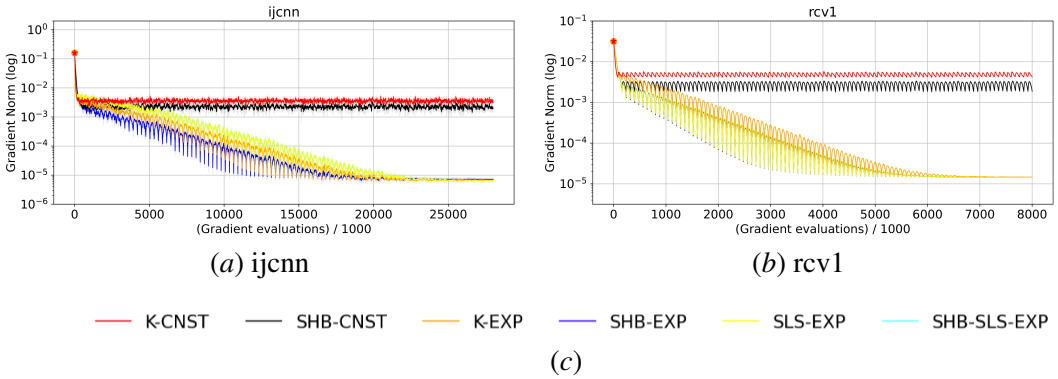
The methods for comparison are: SGD with constant step-sizes (`K-CNST`), SGD with exponentially decreasing step-sizes [21] (`K-EXP`), SGD with exponentially decreasing step-sizes and SLS [19, 21] (`SLS-EXP`), SHB with constant step-sizes [17] (`SHB-CNST`), SHB with exponentially decreasing step-sizes (`SHB-EXP`), SHB with exponentially decreasing step-sizes and SLS (`SHB-SLS-EXP`).



(*a*) ijcnn (*b*) rcv1

(*c*)

Figure 2: Squared loss on ijcnn and rcv1 datasets



(*a*) ijcnn (*b*) rcv1

(*c*)

Figure 3: Logistic loss on ijcnn and rcv1 datasets

We observe that exponentially decreasing step-sizes for both SHB and SGD have close performance and they both outperform their constant step-sizes variants. We also note that using stochastic line-search by [19], SHB-SLS-EXP matches the performance of the variant with known smoothness.