Compute Efficiency and Serial Runtime Tradeoffs for Stochastic Momentum Methods

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Stochastic momentum methods such as heavy ball (HB), Nesterov momentum, and variants of Accelerated SGD (ASGD) [12] are widely used in modern training, but their stochastic benefits depend on two distinct quantities: serial runtime, the number of iterations needed to reach a target accuracy, and compute efficiency (CE), the inverse total gradient-query or FLOP cost. Larger batches reduce serial runtime without hurting CE only when the contraction gap grows linearly with batch size. We study stochastic HB and ASGD for consistent linear regression with Gaussian covariates and prove finite-dimensional, discrete-time *lower* bounds on their batch-size tradeoffs. Our first result shows that HB does *not* improve the CE frontier over SGD for arbitrary spectra; rather, it preserves SGD-level CE over a larger batch-size window, allowing larger batches to reduce serial runtime until HB reaches its deterministic accelerated scale. This window can be a factor $\sqrt{\kappa}$ larger than the SGD critical batch size. For ASGD, the picture is more spectrum-dependent: for rapidly decaying power-law spectra, ASGD improves small-batch CE over HB/SGD, but as batch size grows it trades this CE advantage for improved serial runtime. Synthetic linear-regression experiments verify these qualitative regimes, including near-overlap of ASGD and HB for slowly decaying spectra and the predicted CE–serial tradeoff for rapidly decaying spectra.

1. Introduction

Existing momentum acceleration does not extend to the stochastic regime [12], limiting their practicality for current training regimes. Empirically, several works have found that momentum primarily accelerates in the large batch size regime, while in the small batch case its effects vanish [9, 13, 25, 26, 29]. Most recently, Marek et al. [18] conducted a large scale empirical study in large language model (LLM) pretraining, and have found that at small batch sizes momentum does not bring any benefits over SGD. These works indicate a subtle interplay between compute efficiency and serial runtime in stochastic acceleration. Theoretically, there has been partial progress in understanding stochastic momentum at batch size 1, through the lower bounds established by Kidambi et al. [12] and the asymptotic characterization of Ferbach et al. [8], Lee et al. [14]. We refer the reader to Section C where we discuss these works further. In this paper we provide a nearly full characterization of the serial runtime tradeoffs by providing lower bounds for *arbitrary* spectra for heavy ball, and power law spectra for ASGD.

Contributions. In the deterministic, full-batch regime, the classical picture is clear: on a κ-conditioned quadratic, GD has contraction gap $\alpha_{\mathrm{GD}}(\infty) \asymp \kappa^{-1}$, while HB achieves the accelerated gap $\alpha_{\mathrm{HB}}(\infty) \asymp \kappa^{-1/2}$. Our goal is to understand what remains of this acceleration in the stochastic mini-batch regime, where both serial runtime and compute efficiency matter.

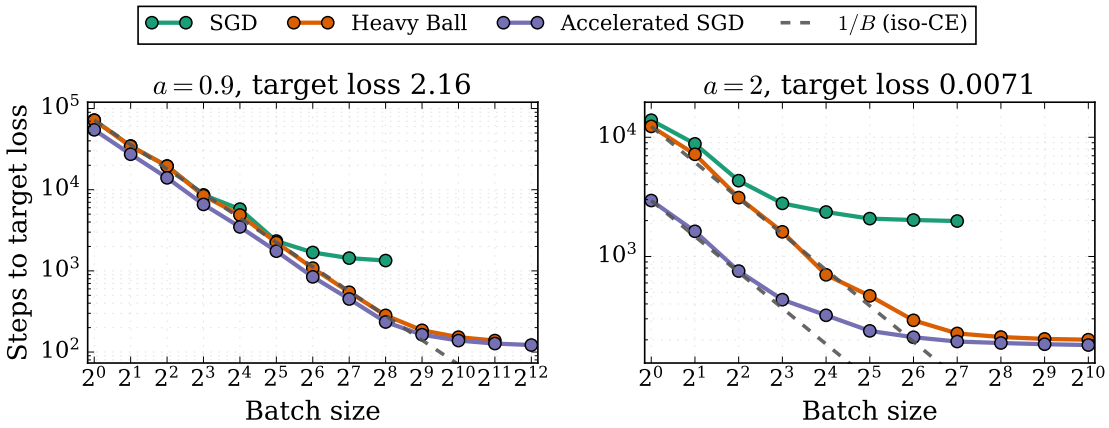

Figure 1: Batch-size tradeoffs for stochastic momentum methods in synthetic linear regression with power-law spectra. Plots show the number of serial steps needed to reach a fixed target loss as the batch size varies, after tuning hyperparameters for each method. (Left) For slow decaying power law spectra, HB and ASGD do not yield better CE, but they decrease the serial runtime. (Right) For fast decaying spectra, ASGD has better CE than HB, but has a shorter linear scaling window before it begins trading off CE for serial runtime. Experimental details are provided in Section D.

For a given algorithm, let $s^*(H, B)$ denote the best-achievable spectral radius at batch size B, and define the contraction gap

$$\alpha(B) := 1 - s^*(H, B).$$

The serial steps needed to reach a fixed target scale as $1/\alpha(B)$, while the total gradient-query cost scales as $B/\alpha(B)$. Thus the compute efficiency (CE) scales as $\frac{\alpha(B)}{B}$. Consequently, if $\alpha(B)$ grows linearly with B, larger batches reduce serial runtime without hurting CE; if it grows sublinearly, larger batches still reduce serial runtime but only by sacrificing CE. Our first main result, summarized in Table 1, holds for *arbitrary* spectra.

HB improves serial runtime over SGD, but it does not improve the CE frontier.

More precisely, HB cannot improve over the optimal CE scaling of SGD at any batch size. Instead, HB preserves SGD-level CE over a larger batch-size window: above the SGD critical batch size, it can continue converting larger batches into fewer serial steps until it reaches the deterministic accelerated scale.

Our second main result concerns ASGD-type methods, which are known to improve small-batch CE over SGD.

ASGD improves small-batch CE, but its CE-preserving linear-scaling window is shorter.

For rapidly decaying spectra, ASGD initially has better CE than HB/SGD. As the batch size grows, its contraction gap continues to improve, but only sublinearly, so ASGD trades part of its small-batch CE advantage for improved serial runtime. Eventually it reaches the deterministic

Algorithm	Batch-size regime	Serial runtime	Compute efficiency
SGD	$1 \le B \lesssim \frac{\mathrm{Tr}(H)}{\lambda_{\max}}$	$\frac{\mathrm{Tr}(H)}{B\lambda_{\min}}$	$\frac{\lambda_{\min}}{\mathrm{Tr}(H)}$
SGD	$B \gtrsim \frac{\mathrm{Tr}(H)}{\lambda_{\max}}$	κ	$\frac{1}{B\kappa}$
HB	$1 \le B \lesssim \frac{\mathrm{Tr}(H)}{\lambda_{\max}}\sqrt{\kappa}$	$\frac{\mathrm{Tr}(H)}{B\lambda_{\min}}$	$\frac{\lambda_{\min}}{\mathrm{Tr}(H)}$
HB	$B \gtrsim \frac{\mathrm{Tr}(H)}{\lambda_{\max}}\sqrt{\kappa}$	$\sqrt{\kappa}$	$\frac{1}{B\sqrt{\kappa}}$

Table 1: Serial-runtime and CE implications of our lower bounds for SGD and HB under arbitrary spectra. Here $\kappa = \lambda_{\max}/\lambda_{\min}$, serial runtime scales as $\alpha(B)^{-1}$, and CE scales as $\alpha(B)/B$. HB preserves the SGD CE frontier over a larger batch-size window, up to $B \asymp (\mathrm{Tr}(H)/\lambda_{\max})\sqrt{\kappa}$, reducing serial runtime from κ to $\sqrt{\kappa}$ before saturating.

Algorithm	Batch-size regime	Serial runtime	Compute efficiency
SGD	$B \gtrsim 1$	d^a	$B^{-1}d^{-a}$
HB	$1 \le B \lesssim d^{a/2}$	$\frac{d^a}{B}$	d^{-a}
HB	$B \gtrsim d^{a/2}$	$d^{a/2}$	$B^{-1}d^{-a/2}$
ASGD	$B \asymp 1$	$d^{\frac{a^2}{2a-1}}$	$d^{-\frac{a^2}{2a-1}}$
ASGD	$1 \lesssim B \lesssim d^{1/2}$	$d^{\frac{a^2}{2a-1}}B^{-\frac{a}{2a-1}}$	$d^{-\frac{a^2}{2a-1}}B^{-\frac{a-1}{2a-1}}$
ASGD	$B \gtrsim d^{1/2}$	$d^{a/2}$	$B^{-1}d^{-a/2}$

Table 2: Power-law consequences of our lower bounds for spectra $\lambda_i \asymp i^{-a}$ with $a > 1$. SGD has only a constant-size linear-scaling window; HB preserves SGD-level CE up to $B \lesssim d^{a/2}$; ASGD improves small-batch CE but loses CE as B grows through the intermediate regime.

accelerated scale; past that point, larger batches no longer improve serial runtime and only decrease CE at the usual $1/B$ rate. For more slowly decaying spectra, ASGD and HB are comparable.

Figure 1 illustrates these regimes in synthetic linear regression. For slowly decaying spectra, ASGD and HB nearly overlap while HB improves serial runtime over SGD at essentially the same CE. For rapidly decaying spectra, ASGD improves small-batch CE over HB/SGD and then spends this advantage to reduce serial runtime as B grows. Thus, HB mainly enlarges the batch-size window over which SGD-level CE can be converted into serial speedup, whereas ASGD improves the small-batch CE frontier but has a shorter CE-preserving window.

2. Background and Preliminaries

Definitions. We study online stochastic optimization for noiseless linear regression with Gaussian covariates. Compute is measured by the total number of gradient queries. For the covariance dynamics of the iterate, let $\mathcal{T}$ denote the linear operator such that $\Sigma_{t+1} = \mathcal{T}(\Sigma_t)$. Following prior work on consistent linear systems [19, 27, 30], the error contracts geometrically at a rate governed by the spectral radius of $\mathcal{T}$. At each iteration, we sample a minibatch of size B from the population $x_i \sim \mathcal{N}(0, H)$ and $y_i = \langle w^*, x_i \rangle$ where $w^* \in \mathbb{R}^d$ is the minimizer and $H \succ 0$ is the covariance matrix.Denote the minibatch empirical covariance $\bar{X} = \frac{1}{B}\sum_{i=1}^{B} x_i x_i^\top$. We write where $H = Q\Lambda Q^\top$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_{\max} = \lambda_1 \ge \cdots \ge \lambda_d = \lambda_{\min} > 0$. and $\kappa = \lambda_{\max}/\lambda_{\min}$. When stated,

we specialize to power-law spectra $\lambda_i \eqsim i^{-a}$ with $a > 1$, so that $\lambda_{\max} \eqsim 1$ and $\lambda_{\min} \eqsim d^{-a}$. The population risk is $\mathcal{R}(w) := \frac{1}{2}\mathbb{E}\big[(\langle w, x\rangle - y)^2\big]$.

Background. We analyze ASGD [11, 12, 20],

$$\begin{aligned} \mu_t &= \beta\mu_{t-1} + (1-\beta)g_t \\ w_{t+1} &= w_t - \eta(\mu_t + \zeta g_t), \end{aligned} \tag{1}$$

where μ_t is the momentum buffer, w_t is the iterate, g_t is the minibatch gradient, and the hyperparameters are the learning rate η, momentum EMA parameter β, and direct-gradient weight ζ. Setting $\zeta = 0$ recovers HB.

3. Compute Efficiency Lower Bounds

We analyze these algorithms through the covariance dynamics of an augmented state vector z_t dependent on the distance to the optimizer $w_t - w^\star$, and includes all variables needed to make the dynamics first-order linear in the covariance $\Sigma_t = \mathbb{E}[z_t z_t^\top]$. For each method, the error covariance satisfies a linear recursion $\Sigma_{t+1} = \mathcal{T}(\Sigma_t)$. where $\mathcal{T}$ is a positive operator defined on the space of PSD matrices that depends on the data covariance H, batch size B and the algorithm hyperparameters. Since the excess risk is a linear function of Σ_t [19, 27, 28, 30], the risk contraction is mainly governed by the *spectral radius* $\rho(\mathcal{T}) = \max_i |t_i|$ where t_i are the eigenvalues of $\mathcal{T}$. Thus, establishing lower bounds on $\rho(\mathcal{T})$, or equivalently upper bounds on the *spectral gap* $s = 1 - \rho(\mathcal{T})$ would directly imply lower bounds on the error rate for a consistent linear system. We leave the formal extension of our spectral bounds to risk bounds to future work.

3.1. Heavy Ball

The operator setup for HB is deferred to Appendix A.2. This brings us to our first main result, and we defer the full proofs to Appendix A.

Theorem 1 (HB Compute Efficiency Lower Bound) *For any data covariance matrix H and minibatch size $B \geq 1$, $\beta, \eta > 0$, the optimal spectral radius of HB satisfies:*

$$s^*(H, B) \;\gtrsim\; 1 - \min\left\{ \frac{B\,\lambda_{\min}}{\mathrm{tr}(H)},\; \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \right\}.$$

where $\gtrsim$ absorbs universal constants. In particular, the transition to the accelerated regime occurs at batch size $B^{\mathrm{crit}}_{\mathrm{HB}} = \frac{\mathrm{Tr}(H)\sqrt{\kappa}}{\lambda_{\max}}$.

Theorem 1 establishes a lower bound on the optimal spectral radius of HB, which in the case of consistent linear systems governs the error rate. From the theorem, we can see that there are, in effect, 2 regimes for HB: for batch size $B < B^{\mathrm{crit}}_{\mathrm{HB}}$, the algorithm cannot improve over the SGD scaling, whereas for batch size $B \geq B^{\mathrm{crit}}_{\mathrm{HB}}$ the best-achievable spectral gap is bounded above by the asymptotic contraction rate.

Corollary 2 (HB on power-law spectra) *Assume the eigenvalues of H satisfy $\lambda_i \eqsim i^{-a}$ for some $a > 1$. Then, the optimal spectral radius of HB satisfies:*

$$s^*(H, B) \;\gtrsim\; 1 - \min\left\{ Bd^{-a},\; d^{-a/2} \right\},$$

where $\gtrsim$ absorbs universal constants. In particular, the transition to the accelerated regime occurs at batch size $B_{\mathrm{HB}}^{\mathrm{crit}} \eqsim d^{a/2}$

3.2. Accelerated SGD

The operator setup for ASGD is deferred to Appendix B.1. Then, we have the following statement.

Theorem 3 (ASGD Lower Bound under power-law spectra) *Assume the eigenvalues of H satisfy $\lambda_i \eqsim i^{-a}$ for some $a > 1$. Then the optimal spectral radius of ASGD satisfies:*

$$s^*(H,B) \;\gtrsim\; 1 - \begin{cases} B\, d^{-\frac{a^2}{2a-1}}, & B \lesssim 1, \\ B^{\frac{a}{2a-1}}\, d^{-\frac{a^2}{2a-1}}, & 1 \lesssim B \lesssim d^{1/2}, \\ d^{-a/2}, & B \gtrsim d^{1/2}, \end{cases}$$

where $\gtrsim$ absorbs universal constants. In particular, the lower bound saturates at the accelerated scale at batch size $B_{\mathrm{ASGD}}^{\mathrm{crit}} \eqsim d^{1/2}$.

Theorem 3 gives three batch-size regimes. First, in the *linear regime*, increasing B linearly decreases the serial steps needed to reach a fixed target. Next, in the *diminishing returns* regime, serial runtime still improves with B, but only sublinearly, since $a/(2a-1) < 1$ for $a > 1$ (roughly a $\sqrt{B}$ gain for large a). Finally, in the *saturation* regime, the lower bound reaches the accelerated scale, so larger batches no longer improve serial runtime and only waste compute. However, note that in the power law setting, the ASGD lower bound reaches the accelerated scale at a *smaller* batch size than the HB lower bound. We expand upon this finding in Corollary 4.

Corollary 4 (ASGD Accelerated Regime.) *Assume the eigenvalues of H satisfy $\lambda_i \eqsim i^{-a}$ for some $a > 1$. Let $B_{\mathrm{HB}}^{\mathrm{crit}}$ and $B_{\mathrm{ASGD}}^{\mathrm{crit}}$ denote the smallest batch sizes at which the HB and ASGD lower bounds reach the optimal accelerated spectral gap $\Theta(d^{-a/2})$. Then*

$$B_{\mathrm{HB}}^{\mathrm{crit}} \eqsim d^{a/2}, \qquad B_{\mathrm{ASGD}}^{\mathrm{crit}} \eqsim d^{1/2}, \qquad \frac{B_{\mathrm{HB}}^{\mathrm{crit}}}{B_{\mathrm{ASGD}}^{\mathrm{crit}}} \eqsim d^{(a-1)/2}.$$

Thus, for every $a > 1$, the ASGD lower bound reaches the accelerated scale at a strictly smaller batch size than the HB lower bound. Equivalently, throughout the interval $d^{1/2} \lesssim B \lesssim d^{a/2}$, the ASGD lower bound has already saturated at the accelerated scale.

We defer the full proofs to Appendix B.

4. Discussion and Conclusions

In this work, we have established compute efficiency lower bounds for heavy ball momentum and accelerated SGD, as a function of the problem instance and the batch size. Specializing the problem instances to power law spectra, we have directly compared the 2 algorithms showing that there is a performance to serial runtime tradeoff: one can train with ASGD at a smaller batch size for a longer serial runtime, achieving better final loss. From a theory point of view, our lower bound for ASGD improves over that of [12], most notably due to the bound applying for any power law spectrum and not a specifically constructed problem. We believe that the techniques used in this work to derive the lower bounds can also be applied to deriving upper bounds, and we leave this derivation to future work.

References

[1] Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum further constrains sharpness at the edge of stochastic stability. *arXiv preprint arXiv:2604.14108*, 2026.

[2] Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model. *arXiv preprint arXiv:2602.04774*, 2026.

[3] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[4] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

[5] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[6] Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

[7] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems*, 37:9974–10007, 2024.

[8] Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales sgd. *arXiv preprint arXiv:2505.16098*, 2025.

[9] Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and why momentum accelerates sgd: An empirical study. *arXiv preprint arXiv:2306.09000*, 2023.

[10] Kanan Gupta, Jonathan W Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. *Advances in Neural Information Processing Systems*, 37:20694–20744, 2024.

[11] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

[12] 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.

[13] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023.

[14] Kiwon Lee, Andrew Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of mini-batch momentum: batch size saturation and convergence in high dimensions. *Advances in Neural Information Processing Systems*, 35:36944–36957, 2022.

[15] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

[16] James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. *arXiv preprint arXiv:1804.00325*, 2018.

[17] Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. *arXiv preprint arXiv:1810.06801*, 2018.

[18] Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful. *arXiv preprint arXiv:2507.07101*, 2025.

[19] Alexandru Meterez, Depen Morwani, Costin-Andrei Oncescu, Jingfeng Wu, Cengiz Pehlevan, and Sham Kakade. A simplified analysis of sgd for linear regression with weight averaging. *arXiv preprint arXiv:2506.15535*, 2025.

[20] Depen Morwani, Nikhil Vyas, Hanlin Zhang, and Sham Kakade. Connections between schedule-free optimizers, ademamix, and accelerated sgd variants. *arXiv preprint arXiv:2502.02431*, 2025.

[21] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Dokl. Akad. Nauk. SSSR*, volume 269, page 543, 1983.

[22] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[23] Matteo Pagliardini, Pierre Ablin, and David Grangier. The ademamix optimizer: Better, faster, older. *arXiv preprint arXiv:2409.03137*, 2024.

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

[25] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

[26] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*, 2023.

[27] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International conference on machine learning*, pages 24280–24314. PMLR, 2022.

[28] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems*, 35:33041–33053, 2022.

[29] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32, 2019.

[30] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on learning theory*, pages 4633–4635. PMLR, 2021.

Appendix A. Heavy Ball Analysis

We will recall a basic identity used repeatedly in the stochastic covariance analysis, namely the Gaussian fourth-moment formula.

Proposition 5 *For any deterministic matrix Σ and $x \sim \mathcal{N}(0, H)$ we have the following equation:*

$$\mathbb{E}[xx^\top \Sigma xx^\top] = 2H\Sigma H + \operatorname{Tr}(H\Sigma)H \preceq 3\operatorname{Tr}(H\Sigma)H. \tag{2}$$

For a minibatch empirical covariance $\bar{X} = \frac{1}{B}\sum_{i=1}^{B} x_i x_i^\top$, Proposition 5 gives

$$\mathbb{E}[\bar{X}\Sigma\bar{X}] = \left(1 + \frac{1}{B}\right) H\Sigma H + \frac{1}{B}\operatorname{Tr}(H\Sigma)H \preceq \left(1 + \frac{2}{B}\right)\operatorname{Tr}(H\Sigma)H. \tag{3}$$

A.1. Proof of Proposition 5

Proof The proof is a simple application of Isserlis's theorem. Elementwise, we have that:

$$\mathbb{E}[xx^\top \Sigma xx^\top]_{ij} = \sum_{kl} \Sigma_{kl}\mathbb{E}[x_i x_k x_l x_j] = \sum_{kl} \Sigma_{kl}(H_{ik}H_{lj} + H_{il}H_{kj} + H_{ij}H_{kl})$$

Assembling the result in matrix form gives the first part.

For $v \in \mathbb{R}^d$, define $u := H^{1/2}v$. Then

$$v^\top H\Sigma H v = v^\top H^{1/2}(H^{1/2}\Sigma H^{1/2})H^{1/2}v = u^\top H^{1/2}\Sigma H^{1/2}u.$$

Also,

$$\operatorname{Tr}(H\Sigma)\, v^\top H v = \operatorname{Tr}\Big(H^{1/2}\Sigma H^{1/2}\Big)\, u^\top u.$$

Since $H \succeq 0$ and $\Sigma \succeq 0$, we have $H^{1/2}\Sigma H^{1/2} \succeq 0$. Hence there exists an orthogonal matrix P and a diagonal matrix $D \succeq 0$ such that

$$H^{1/2}\Sigma H^{1/2} = PDP^\top.$$

Since $D \succeq 0$, we have $D \preceq \operatorname{Tr}(D)\, I$. Therefore,

$$u^\top H^{1/2}\Sigma H^{1/2}u = u^\top PDP^\top u \leq u^\top P(\operatorname{Tr}(D)I)P^\top u = \operatorname{Tr}(D)\, u^\top u.$$

Using $\operatorname{Tr}(D) = \operatorname{Tr}\big(H^{1/2}\Sigma H^{1/2}\big) = \operatorname{Tr}(H\Sigma)$, we obtain

$$v^\top H\Sigma H v \leq \operatorname{Tr}(H\Sigma)\, v^\top H v.$$

Since this holds for all $v \in \mathbb{R}^d$, it follows that $H\Sigma H \preceq \operatorname{Tr}(H\Sigma)H$.

■

A.2. Heavy Ball Operator Setup

We can rewrite the HB update from equation equation 1 in the following form:

$$w_{t+1} = w_t - \eta(1-\beta)\,\hat{g}_t + \beta\,(w_t - w_{t-1}), \qquad \eta > 0,\ \beta \in [0,1).$$

We can rewrite the update as a linear recursion, where the vector form is the augmented state:

$$\begin{bmatrix} w_{t+1} - w^* \\ w_t - w^* \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \eta(1-\beta)\bar{X}_t & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} w_t - w^* \\ w_{t-1} - w^* \end{bmatrix} \tag{4}$$

Denote by $z_t = \begin{bmatrix} w_t - w^* \\ w_{t-1} - w^* \end{bmatrix}$ with covariance $\Sigma_t = \mathbb{E}[z_t z_t^\top]$, and denote the transition operator from equation equation 4 as $\hat{A}_t$. We can decompose $\hat{A}_t = M_t + \tilde{A}$, where $\tilde{A}$ will be the deterministic component and M_t the random component as:

$$M_t = \begin{bmatrix} -\eta(1-\beta)(\bar{X}_t - H) & 0 \\ 0 & 0 \end{bmatrix} \qquad \tilde{A} = \begin{bmatrix} (1+\beta)I - \eta(1-\beta)H & -\beta I \\ I & 0 \end{bmatrix}$$

Computing the covariance of the augmented state vector gives:

$$\Sigma_{t+1} = \tilde{A}\Sigma_t\tilde{A}^\top + \mathbb{E}[M_t z_t z_t^\top M_t^\top]$$

For the second expectation, we need to compute a 4th moment Gaussian term coming from $\mathbb{E}[\bar{X}_t z_t z_t^\top \bar{X}_t]$, only on the 11 block (since that is where $\bar{X}_t$ is). We can compute this term by applying Proposition 5 and upper bound $H\Sigma H \preceq \mathrm{Tr}(H\Sigma)H$. With an abuse of notation, we will write the recursion of Σ_{t+1} with equality after applying this bound, since we only lose a small constant factor, thus obtaining:

$$\Sigma_{t+1} = \tilde{A}\Sigma_t\tilde{A}^\top + \begin{bmatrix} \frac{2\eta^2(1-\beta)^2}{B}\,\mathrm{Tr}\big(H\Sigma_t^{11}\big)H & 0 \\ 0 & 0 \end{bmatrix}$$

Henceforth, we will express all matrices in the eigenbasis of H. Thus, after rotation H reduces to Λ with $S_t = Q^\top\Sigma_t Q$ and $A = Q^\top\tilde{A}Q$. Thus, the recursion becomes:

$$S_{t+1} = A\,S_t\,A^\top \;+\; \eta^2(1-\beta)^2 \begin{bmatrix} \frac{1}{B}\,\Lambda\,\mathrm{Tr}(\Lambda S_t^{11}) & 0 \\ 0 & 0 \end{bmatrix} \tag{5}$$

Note that A has a block diagonal structure as $A = \mathrm{blkdiag}(A_1, \ldots, A_d)$, where each per–coordinate 2×2 block is:

$$A_i = \begin{bmatrix} a_i & -\beta \\ 1 & 0 \end{bmatrix}, \qquad a_i := (1+\beta) - \eta(1-\beta)\lambda_i. \tag{6}$$

Let $\mathcal{T}_{H,\eta,\beta,B}$ denote the corresponding linear covariance update operator on symmetric $2d \times 2d$ matrices as defined in equation 5, such that:

$$S_{t+1} = \mathcal{T}_{H,\eta,\beta,B}(S_t)$$

Let $s(H,\eta,\beta,B)$ denote the spectral radius of $\mathcal{T}_{H,\eta,\beta,B}$

$$s(H,\eta,\beta,B) := \rho\big(\mathcal{T}_{H,\eta,\beta,B}\big) := \max\{\,|s| : s \text{ eigenvalue of } \mathcal{T}_{H,\eta,\beta,B}\,\}.$$

As we see later on (in Lemma 9), this spectral radius is attained by a real, nonnegative eigenvalue with a PSD eigenmatrix. We define the *optimal* spectral radius attainable at batch size B as

$$s^*(H,B) := \inf_{\eta>0,\ \beta\in[0,1)} s(H,\eta,\beta,B),$$

A.3. Deriving the Secular Equation

Recall that:

$$S_{t+1} = A\, S_t\, A^\top \;+\; \eta^2(1-\beta)^2 \begin{bmatrix} \frac{1}{B}\,\Lambda\,\mathrm{Tr}(\Lambda S_t^{11}) & 0 \\ 0 & 0 \end{bmatrix}$$

Let $\gamma = \mathrm{Tr}(\Lambda S_t^{11})$. Pushing a vec through this equation coupled with the fact that $\mathrm{vec}(AS_tA^\top) = (A \otimes A)\,\mathrm{vec}(S_t)$ gives us:

$$s\,\mathrm{vec}(S_t) = (A \otimes A)\,\mathrm{vec}(S_t) + \frac{2\eta^2(1-\beta)^2\gamma}{B} \begin{bmatrix} \mathrm{vec}(\Lambda) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\implies (sI - A \otimes A)\,\mathrm{vec}(S_t) = \frac{2\eta^2(1-\beta)^2\gamma}{B} \begin{bmatrix} \mathrm{vec}(\Lambda) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

We can break the above equation up into d, 2×2 equations each using a block of A:

$$(sI - A_i \otimes A_i)\,\mathrm{vec}(S_{t,i}) = \frac{2\eta^2(1-\beta)^2\gamma}{B} \begin{bmatrix} \lambda_i \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\implies \mathrm{vec}(S_{t,i}) = \frac{2\eta^2(1-\beta)^2\gamma}{B} (sI - A_i \otimes A_i)^{-1} \begin{bmatrix} \lambda_i \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

To simplify γ, we multiply the above Equation by $\begin{bmatrix} \lambda_i \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and sum over $i \in \{1, \ldots, d\}$ to get:

$$1 = \frac{2\eta^2(1-\beta)^2}{B} \sum_{i=1}^{d} \begin{bmatrix} \lambda_i & 0 & 0 & 0 \end{bmatrix} (sI - A_i \otimes A_i)^{-1} \begin{bmatrix} \lambda_i \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

Inverting the Kronecker. Suppose M_i is an eigenbasis for A_i, and D_i are the eigenvalues of A_i. Then, $M_i \otimes M_i$ is an eigenbasis for $A_i \otimes A_i$ and $D_i \otimes D_i$ are the eigenvalues of A_i. Thus, we have:

$$(sI - A_i \otimes A_i)^{-1} = M_i \otimes M_i (sI - D_i \otimes D_i)^{-1} M_i^\top \otimes M_i^\top$$

with

$$D_i = \begin{bmatrix} r_{i,+} & 0 \\ 0 & r_{i,-} \end{bmatrix}$$

and

$$M_i = \begin{bmatrix} r_{i,+} & r_{i,-} \\ 1 & 1 \end{bmatrix} \tag{8}$$

Equations (7) and (8) on further simlification give us:

$$1 = \frac{2\eta^2(1-\beta)^2}{B} \sum_i \frac{\lambda_i^2 s(s+\beta)}{(s - r_{i,+}^2)(s - r_{i,-}^2)(s-\beta)}$$

Let

$$\phi_i(s) = \frac{s(s+\beta)}{(s-\beta)\,(s - r_{i,+}^2)\,(s - r_{i,-}^2)}.$$

The final secular Equation can be writen as

$$1 \;=\; \frac{c}{B} \sum_{i=1}^{d} \frac{\lambda_i^2\, \phi_i(s)}{1 - c\lambda_i^2 \phi_i(s)}\,, \qquad c = \eta^2(1-\beta)^2. \tag{9}$$

Validity of vectorizing Equation 16 The vectorized equation is acting on the $4d$ dimensional space defined by the d 2×2 blocks, while the symmetric matrices occupy a $3d$ space here. So the secular equation has extra roots. But we will argue that the maximum eigenvalue of the secular equation is still associated with a per-diagonal block PSD matrix.

First, within the space of per-diagonal symmetric matrices, per-diagonal PSD matrix has the maximum eigenvalue. This follows from KT theorem, as per-diagonal PSD matrix form a total cone. Then, let's consider the solutions of secular equation outside the symmetric space. Let M be an eigen vector of $\mathcal{T}_{H,\eta,\beta,B}$ which is neither symmetric nor anti-symmetric. By linearity of $\mathcal{T}_{H,\eta,\beta,B}$,

$$\mathcal{T}_{H,\eta,\beta,B}(M) = \lambda M \implies \mathcal{T}_{H,\eta,\beta,B}(M^T) = \lambda M^\top$$

Now, let's decompose M into its symmetric and anti-symmetric parts:

$$M_s := \frac{M + M^\top}{2}, \qquad M_a := \frac{M - M^\top}{2}.$$

Then, by linearity,

$$\mathcal{T}_{H,\eta,\beta,B}(M_s) = \frac{\mathcal{T}_{H,\eta,\beta,B}(M) + \mathcal{T}_{H,\eta,\beta,B}(M^\top)}{2} = \frac{\lambda M + \lambda M^\top}{2} = \lambda X_s,$$

and similarly

$$\mathcal{T}_{H,\eta,\beta,B}(M_a) = \frac{\mathcal{T}_{H,\eta,\beta,B}(M) - \mathcal{T}_{H,\eta,\beta,B}(M^\top)}{2} = \frac{\lambda M - \lambda M^\top}{2} = \lambda M_a.$$

Therefore, any eigenvalue carried by an eigenmatrix M is also carried by a symmetric eigenmatrix and an anti-symmetric eigenmatrix. Also, if M is anti-symmetric, $\mathrm{Tr}(\Lambda \mathrm{S}^{11}) = 0 \implies \mathcal{T}_{H,\eta,\beta,B}(M) = \mathcal{T}_\infty(M)$. Thus, the eigenvalue of an anti-symmetric matrix for the stochastic operator coincide with the eigenvalue for the deterministic operator. In addition, note the following: (i) A is block diagonal, (ii) give the stochastic opertor $\mathcal{T}_{H,\eta,\beta,B}$, gauging at the block diagonal eigenmatrices of the determinsitc operator $\mathcal{T}_\infty$ suffices. For a block diagonal eigenmatrix, the eigenvalues of $\mathcal{T}_\infty$ are that of $A_i \otimes A_i$, with the eigenmatrix corresponding to the largest eigenvalue being a symmetric eigenmatrix, with rank-1 block diagonal entries of the form $u_i u_i^T$, where u_i represents the eigenvector of A_i corresponding to its largest eigenvalue. Thus, on the determinisitc operator $\mathcal{T}_\infty$, the eigenvalue of a symmetric eigenmatrix is always greater than or equal to that of an anti-symmetric matrix. By Conjecture 9, the spectral radius for $\mathcal{T}_{H,\eta,\beta,B}$ is obtained by a real eigenvalue and from Lemma 8, the spectral radius of $\mathcal{T}_{H,\eta,\beta,B}$ is lower bounded by the spectral radius for $\mathcal{T}_\infty$, thus, the dominating eigenvalue of the secular equation is given by a symmetric PSD matrix.

A.4. The Zero-Noise Heavy Ball Analysis

In the deterministic setting, the Heavy-Ball iteration in Equation 5 reduces to

$$S_{t+1} = A S_t A^\top \tag{10}$$

The following technical lemma governs the maximal learning rate of the deterministic Heavy–Ball algorithm.

Lemma 6 (Stability band of deterministic HB) *Assume $\beta, \eta > 0$. If the Heavy–Ball iteration in Equation 10 is stable (i.e. all eigenvalues of every A_i lie in the open unit disk), if and only if*

$$\beta < 1, \qquad 0 < \eta < \frac{2(1+\beta)}{(1-\beta)\lambda_{\max}}.$$

Proof Fix $\beta \in [0, 1)$ and first consider a generic 2×2 matrix

$$A = \begin{bmatrix} a & -\beta \\ 1 & 0 \end{bmatrix}, \qquad a \in \mathbb{R}.$$

Let $r_\pm$ be the eigenvalues of A, i.e. the roots of

$$z^2 - az + \beta \;=\; 0.$$

We recall the standard discrete-time stability criterion for a degree–2 polynomial. Writing this polynomial in the form

$$z^2 + pz + q = 0$$

with $p = -a$ and $q = \beta$, the Jury/Schur stability test states that the roots lie strictly inside the unit disk, $|z| < 1$, if and only if

$$|q| < 1, \qquad 1 + p + q > 0, \qquad 1 - p + q > 0, \qquad 1 - q > 0.$$

(These are the degree–2 Jury conditions.)

In our case, substituting $p = -a$ and $q = \beta$ into the above gives

$$\begin{aligned} |q| < 1 &\iff |\beta| < 1, \\ 1 + p + q > 0 &\iff 1 - a + \beta > 0 \iff a < 1 + \beta, \\ 1 - p + q > 0 &\iff 1 + a + \beta > 0 \iff a > -(1 + \beta), \\ 1 - q > 0 &\iff 1 - \beta > 0. \end{aligned}$$

Thus, for the matrix A, the eigenvalues satisfy $|r_\pm| < 1$ if and only if $|\beta| < 1$ and

$$-(1 + \beta) \ < \ a \ < \ 1 + \beta.$$

We now apply this to the deterministic HB iteration on the quadratic. In particular, we apply these conditions to each state matrix A_i, where $a_i = (1+\beta) - \eta(1-\beta)\lambda_i$. Stability for all i requires

$$-(1 + \beta) < a_i < 1 + \beta \quad \text{for all} \quad \lambda_i \in [0, \lambda_{\max}].$$

The upper bound $a_i < 1 + \beta$ is satisfied, since $\eta > 0$ by assumption. The lower bound must hold in particular at the largest curvature $\lambda_{\max}$, where a_i is smallest:

$$(1 + \beta) - \eta(1 - \beta)\lambda_{\max} \ > \ -(1 + \beta).$$

Rearranging,

$$(1+\beta)-\eta(1-\beta)\lambda_{\max} > -(1+\beta) \quad \iff \quad 2(1+\beta) > \eta(1-\beta)\lambda_{\max} \quad \iff \quad \eta < \frac{2(1+\beta)}{(1-\beta)\lambda_{\max}}.$$

Finally, stability also requires $|\beta| < 1$ and $1 - \beta > 0$, which in our nonnegative-β setting is exactly $0 \le \beta < 1$. Combining these conditions gives the claimed deterministic HB stability band. ■

A.5. Compute Efficiency (CE) lower bounds

Let $s(H, \eta, \beta, B)$ denote the spectral radius of $\mathcal{T}_{H,\eta,\beta,B}$:

$$s(H, \eta, \beta, B) := \rho\big(\mathcal{T}_{H,\eta,\beta,B}\big) := \max\{\, |s| : s \text{ eigenvalue of } \mathcal{T}_{H,\eta,\beta,B} \,\}.$$

As we see later on (in Lemma 9), this spectral radius is attained by a real, nonnegative eigenvalue with a PSD eigenmatrix.

We define the *optimal* spectral radius attainable at batch size B as

$$s^*(H, B) := \inf_{\eta>0,\ \beta\in[0,1)} s(H, \eta, \beta, B),$$

and the corresponding *spectral gap* as $1 - s^*(H, B)$.

Theorem 7 (HB–SGD compute efficiency lower bound) *For any covariance matrix H and mini-batch size $B \ge 1$, $\beta, \eta > 0$, the optimal spectral gap satisfies*

$$s^*(H, B) \ \ge \ 1 - 8 \ \min\left\{ \frac{B\,\lambda_{\min}}{\operatorname{tr}(H)} \,,\ \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \right\}.$$

A.6. Helper Lemmas

We will refer to any eigenvalue s solving the secular equation equation 9 as an *observable eigenvalue*, i.e., an eigenvalue whose eigenmode has nonzero coupling to the scalar observable S_w. Recall the secular equation for the observable eigenvalue s of $\mathcal{T}_{H,\eta,\beta,B}$:

$$1 \;=\; F(s) \;:=\; \frac{c}{B}\sum_{i=1}^{d}\frac{\lambda_i^2\,\phi_i(s)}{1-c\,\lambda_i^2\phi_i(s)}\,, \qquad c=\eta^2(1-\beta)^2, \tag{11}$$

with

$$\phi_i(s)=\frac{s(s+\beta)}{(s-\beta)\,(s-r_{i,+}^2)\,(s-r_{i,-}^2)}\,, \quad r_{i,\pm}=\frac{a_i\pm\sqrt{a_i^2-4\beta}}{2}, \quad a_i=(1+\beta)-\eta(1-\beta)\lambda_i. \tag{12}$$

The poles of ϕ_i are at $s=\beta$ and $s=r_{i,\pm}^2$; for a deterministically stable HB choice, all these poles lie in $(0,1)$. We now formalize the intuition that adding stochastic gradient noise can only *slow* convergence. In particular, it cannot yield a larger spectral gap than the deterministic Heavy–Ball dynamics.

In the zero-noise case (full batch, $B=\infty$), the Heavy–Ball iteration on the $[w_t, w_{t-1}]$ state induces the deterministic covariance recursion

$$S_{t+1} \;=\; AS_tA^\top,$$

where $A=\mathrm{blkdiag}(A_1,\dots,A_d)$ is the block-diagonal state matrix in the H-basis (with A_i as defined in Equation 6). Let

$$\mathcal{T}^{(\infty)}(S) := ASA^\top \tag{13}$$

denote this deterministic (zero-noise) covariance operator.

Lemma 8 (Stochastic spectral radius dominates deterministic) *For every (η,β,B) we have*

$$\rho\big(\mathcal{T}_{H,\eta,\beta,B}\big) \;\ge\; \rho\big(\mathcal{T}^{(\infty)}\big) \;=\; \rho(A)^2.$$

Equivalently, the stochastic spectral gap is *no larger* than the deterministic gap:

$$1-s(H,\eta,\beta,B) \;\le\; 1-\rho(A)^2.$$

Before proving this, we record a Perron–Frobenius–type fact for the covariance operators, which follows from the Krein–Rutman theorem (Perron–Frobenius for positive operators on cones).

Corollary 9 (Perron–Frobenius for covariance maps) *Let $\mathcal{S}^{2d}$ denote the space of real symmetric $2d\times 2d$ matrices and let $\mathcal{K} := \{\,S\in\mathcal{S}^{2d} : S\succeq 0\,\}$ be the cone of positive semidefinite (PSD) matrices. Consider either of the covariance operators $\mathcal{T}\in\{\,\mathcal{T}_{H,\eta,\beta,B},\ \mathcal{T}^{(\infty)}\,\}$. Then:*

- *$\mathcal{T}$ is a* positive *operator on the cone $\mathcal{K}$, i.e. if $S\succeq 0$, then $\mathcal{T}(S)\succeq 0$.*
- *The spectral radius*

  $$\rho(\mathcal{T}) := \max\{\,|s| : s \text{ eigenvalue of } \mathcal{T}\,\}$$

  is attained by a real, nonnegative eigenvalue. That is, there exists $S_\star\in\mathcal{K}\setminus\{0\}$ and a real $\lambda_\star\ge 0$ such that

  $$\mathcal{T}(S_\star)=\lambda_\star S_\star, \qquad \lambda_\star=\rho(\mathcal{T}).$$

In particular, each covariance operator admits a principal eigenpair $(\lambda_\star, S_\star)$ *with* $\lambda_\star = \rho(\mathcal{T})$ *and* $S_\star \succeq 0$.

The corollary shows that the leading asymptotic mode of the covariance dynamics can always be represented by a PSD eigenmatrix.

We now prove Lemma 8.

Proof [Proof of Lemma 8] For brevity, write $\mathcal{T}^{(B)} := \mathcal{T}_{H,\eta,\beta,B}$. From the exact recursion and the definition of $\mathcal{T}^{(\infty)}$ in equation 13, we can write

$$\mathcal{T}^{(B)}(S) = \mathcal{T}^{(\infty)}(S) + \mathcal{N}_B(S),$$

where the linear "noise map" $\mathcal{N}_B$ is

$$\mathcal{N}_B(S) := \eta^2(1-\beta)^2 \begin{bmatrix} \frac{1}{B}\,\Lambda\,\mathrm{Tr}(\Lambda\Sigma_{11}) & 0 \\ 0 & 0 \end{bmatrix}, \qquad S = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix}.$$

If $S \succeq 0$ then $\Sigma_{11} \succeq 0$, and both $\Lambda\,\Sigma_{11}\,\Lambda$ and $\Lambda\,\mathrm{tr}(\Lambda\Sigma_{11})$ are PSD. Hence $\mathcal{N}_B(S) \succeq 0$, and therefore

$$\mathcal{T}^{(\infty)}(S) \preceq \mathcal{T}^{(B)}(S) \qquad \text{for all } S \succeq 0.$$

Apply Corollary 9 to the deterministic operator $\mathcal{T}^{(\infty)}$ acting on $\mathcal{K}$. There exists an eigenpair $(\lambda_\infty, S_\infty)$ with

$$S_\infty \succeq 0, \quad S_\infty \neq 0, \qquad \mathcal{T}^{(\infty)}(S_\infty) = \lambda_\infty S_\infty, \qquad \lambda_\infty = \rho\big(\mathcal{T}^{(\infty)}\big).$$

Since $S_\infty \succeq 0$ and $\mathcal{N}_B$ is PSD-valued on $\mathcal{K}$, the domination relation implies

$$\mathcal{T}^{(B)}(S_\infty) = \mathcal{T}^{(\infty)}(S_\infty) + \mathcal{N}_B(S_\infty) \succeq \lambda_\infty S_\infty.$$

By induction,

$$\big(\mathcal{T}^{(B)}\big)^n(S_\infty) \succeq \lambda_\infty^n S_\infty \qquad \text{for all } n \geq 1.$$

Fix a matrix norm $\|\cdot\|$ on $\mathcal{S}^{2d}$ that is monotone on the PSD cone: $0 \preceq Y \preceq X$ implies $\|Y\| \leq \|X\|$. The spectral norm $\|\cdot\|_2$ has this property, since $\|X\|_2 = \lambda_{\max}(X)$ for $X \succeq 0$ and $Y \preceq X$ implies $\lambda_{\max}(Y) \leq \lambda_{\max}(X)$.

The induced operator norm of $(\mathcal{T}^{(B)})^n$ is

$$\big\|\big(\mathcal{T}^{(B)}\big)^n\big\| := \sup_{S \neq 0} \frac{\big\|\big(\mathcal{T}^{(B)}\big)^n(S)\big\|}{\|S\|}.$$

Evaluating this supremum at $S = S_\infty$ and using monotonicity on the cone,

$$\big\|\big(\mathcal{T}^{(B)}\big)^n\big\| \;\geq\; \frac{\big\|\big(\mathcal{T}^{(B)}\big)^n(S_\infty)\big\|}{\|S_\infty\|} \;\geq\; \frac{\big\|\lambda_\infty^n S_\infty\big\|}{\|S_\infty\|} = \lambda_\infty^n.$$

By Gelfand's formula,

$$\rho\big(\mathcal{T}^{(B)}\big) = \lim_{n\to\infty} \big\|\big(\mathcal{T}^{(B)}\big)^n\big\|^{1/n} \;\geq\; \lim_{n\to\infty} \lambda_\infty = \lambda_\infty = \rho\big(\mathcal{T}^{(\infty)}\big).$$

Finally, $\mathcal{T}^{(\infty)}$ acts as $A \otimes A$ on the vectorized state, so its eigenvalues are products $\lambda_i\lambda_j$ of eigenvalues of A. Therefore

$$\rho\big(\mathcal{T}^{(\infty)}\big) = \max_{i,j} |\lambda_i\lambda_j| = \big(\max_i |\lambda_i|\big)^2 = \rho(A)^2.$$

Combining the two inequalities yields the claimed bound. ■

A.7. Proof

We start with a step size bound for HB–SGD.

Lemma 10 (Stability step-size cap) *Assume HB–SGD with parameters (η, β, B) is stable, i.e., all eigenvalues of $\mathcal{T}_{H,\eta,\beta,B}$ satisfy $|s| \le 1$. Then the step size must satisfy*

$$\eta \;\le\; \min\left\{ \frac{2B}{\mathrm{tr}(H)}\,,\; \frac{2(1+\beta)}{(1-\beta)\lambda_{\max}} \right\}.$$

Proof First, let us show that stability forces

$$F(1) \;\le\; 1.$$

Any real s with $F(s) = 1$ corresponds to an eigenvalue s of $\mathcal{T}_{H,\eta,\beta,B}$. In the stable regime, all eigenvalues of $\mathcal{T}_{H,\eta,\beta,B}$ satisfy $|s| \le 1$, so in particular there can be no real eigenvalue $s > 1$. $\phi_i(s) = O(1/s)$ implies $F(s) \to 0$ as $s \to \infty$. Therefore, if $F(1) > 1$, by the Intermediate Value Theorem, there exists $s > 1$ with $F(s) = 1$, which would correspond to an eigenvalue $s > 1$ and hence contradict stability. This proves the claim.

At $s = 1$, we can evaluate $\phi_i(1)$ explicitly. Using

$$(1 - r_{i,+}^2)(1 - r_{i,-}^2) = (1+\beta)^2 - a_i^2 = 2(1+\beta)\,\eta(1-\beta)\lambda_i - \eta^2(1-\beta)^2\lambda_i^2,$$

we obtain

$$\phi_i(1) = \frac{1+\beta}{(1-\beta)\,(1 - r_{i,+}^2)(1 - r_{i,-}^2)} = \frac{1+\beta}{(1-\beta)} \cdot \frac{1}{\eta(1-\beta)\lambda_i} \cdot \frac{1}{2(1+\beta) - \eta(1-\beta)\lambda_i}.$$

In the stable regime, each denominator $1 - c\lambda_i^2\phi_i(1)$ is positive (otherwise F would blow up at or before $s = 1$ and there would be an eigenvalue with $s \ge 1$). Using $x/(1-x) \ge x$ for $x \in [0, 1)$, we have

$$1 \;\ge\; F(1) = \frac{c}{B}\sum_{i=1}^{d} \frac{\lambda_i^2\phi_i(1)}{1 - c\lambda_i^2\phi_i(1)} \;\ge\; \frac{c}{B}\sum_{i=1}^{d} \lambda_i^2\phi_i(1).$$

Substituting $c = \eta^2(1-\beta)^2$ and the expression for $\phi_i(1)$ gives

$$1 \;\ge\; \frac{\eta(1+\beta)}{B}\sum_{i=1}^{d} \frac{\lambda_i}{2(1+\beta) - \eta(1-\beta)\lambda_i}. \tag{14}$$

We first derive the bound $\eta \le 2B/\mathrm{tr}(H)$. Dropping the negative term in each denominator, $2(1+\beta) - \eta(1-\beta)\lambda_i \le 2(1+\beta)$, we obtain

$$1 \;\ge\; \frac{\eta(1+\beta)}{B}\sum_{i=1}^{d} \frac{\lambda_i}{2(1+\beta) - \eta(1-\beta)\lambda_i} \;\ge\; \frac{\eta(1+\beta)}{B}\sum_{i=1}^{d} \frac{\lambda_i}{2(1+\beta)} = \frac{\eta\,\mathrm{tr}(H)}{2B},$$

which completes the proof of this case.

We now show the second bound of $\eta \le 2(1+\beta)/((1-\beta)\lambda_{\max})$ must hold. Lemma 8 implies that if the deterministic HB iteration is unstable, i.e. $\rho(A) > 1$, then $\rho(\mathcal{T}_{H,\eta,\beta,B}) > 1$. Now if $\eta > 2(1+\beta)/((1-\beta)\lambda_{\max})$, then Lemma 6 implies $\rho(A) > 1$, contradicting our assumed stability. This proves our second claim. ∎

Lemma 11 (Gap in terms of η and β) *For any stable HB–SGD parameters (η, β, B), the spectral gap satisfies*

$$1 - s(H, \eta, \beta, B) \;\le\; \min\big\{4\eta\lambda_{\min}, 1-\beta\big\}.$$

Proof By Lemma 8, the stochastic covariance operator $\mathcal{T}_{H,\eta,\beta,B}$ has spectral radius at least that of the deterministic (full-batch) HB operator:

$$s(H, \eta, \beta, B) \;\ge\; \rho\big(\mathcal{T}^{(\infty)}\big) = \rho(A)^2,$$

where $A = \mathrm{blkdiag}(A_1, \ldots, A_d)$ is the deterministic HB state matrix in the $[w_t, w_{t-1}]$ state. Hence

$$1 - s(H, \eta, \beta, B) \;\le\; 1 - \rho(A)^2.$$

We first bound the gap by $1-\beta$. For each coordinate i, the per–coordinate block A_i has eigenvalues $r_{i,\pm}$ satisfying $r_{i,+}r_{i,-} = \beta$. Thus

$$\max\{|r_{i,+}|, |r_{i,-}|\}^2 \;\ge\; |r_{i,+}r_{i,-}| = \beta.$$

Taking a maximum over i yields $\rho(A)^2 \ge \beta$, so

$$1 - \rho(A)^2 \;\le\; 1 - \beta.$$

Therefore

$$1 - s(H, \eta, \beta, B) \;\le\; 1 - \beta.$$

We now show $1 - s(H, \eta, \beta, B) \le 4\eta\lambda_{\min}$. Let $A_{\min}$ be the 2×2 HB block corresponding to the smallest eigenvalue $\lambda_{\min}$, with eigenvalues $r_{\min,\pm}$ solving

$$z^2 - a_{\min} z + \beta = 0, \qquad a_{\min} = (1+\beta) - \eta(1-\beta)\lambda_{\min}.$$

As $A_{\min}$ is a block of A, we have

$$\rho(A) \;\ge\; \max\{|r_{\min,+}|, |r_{\min,-}|\}.$$

Hence

$$1 - s(H, \eta, \beta, B) \;\le\; 1 - \rho(A)^2 \;\le\; 1 - \max\{|r_{\min,+}|, |r_{\min,-}|\}^2.$$

We distinguish the real and complex cases for $r_{\min,\pm}$.

For case of real roots (the overdamped case), Assume $r_{\min,\pm} \in \mathbb{R}$, which corresponds to $a_{\min}^2 > 4\beta$. We have $1 - \rho(A)^2 \le (1 - r_{\min,+}^2) \le 2(1 - r_{\min,+})$. The roots satisfy

$$(r_{\min,+} - 1)(r_{\min,-} - 1) = 1 - (r_{\min,+} + r_{\min,-}) + r_{\min,+}r_{\min,-} = 1 - a_{\min} + \beta = \eta(1-\beta)\lambda_{\min}.$$

Since the roots are real and satisfy $r_+ r_- = \beta$ with $r_- \le r_+$, we must have $r_- \le \sqrt{\beta}$. Using that $\beta \le 1$ (stability),

$$1 - \rho(A)^2 \le 2(1 - r_{\min,+}) = 2\frac{\eta(1-\beta)\lambda_{\min}}{1 - r_-} \le 2\frac{\eta(1-\beta)\lambda_{\min}}{1 - \sqrt{\beta}} = 2\eta(1+\sqrt{\beta})\lambda_{\min} \le 4\eta\lambda_{\min},$$

which completes the proof of this case.

For case of complex roots (underdamped case), assume $r_{\min,\pm}$ are complex conjugates, where $a_{\min}^2 < 4\beta$. Then both have modulus $\sqrt{\beta}$, so

$$1 - \rho(A)^2 \;\le\; 1 - |r_{\min,+}|^2 = 1 - \beta.$$

For the underdamped case (complex roots), due to that $a_{\min} = (1+\beta) - \eta(1-\beta)\lambda_{\min}$, the condition $a_{\min}^2 < 4\beta$ is equivalent to:

$$\eta(1-\beta)\lambda_{\min} > (1-\sqrt{\beta})^2 \quad\Longrightarrow\quad \eta(1+\sqrt{\beta})\lambda_{\min} > 1 - \sqrt{\beta}.$$

Therefore,

$$1 - \beta \;=\; (1-\sqrt{\beta})(1+\sqrt{\beta}) \le (1+\sqrt{\beta})^2\eta\lambda_{\min} \;\le 4\eta\lambda_{\min},$$

which completes the proof of the complex case. ■

Now we are equipped to complete the proof of Theorem 7.

Proof [Proof of Theorem 7] Fix H and any stable HB–SGD parameters (η, β, B).

Combining Lemma 11 (the step size cap) with Lemma 10 (the bound on $1 - s(H, \eta, \beta, B)$), we obtain three simultaneous upper bounds on the same quantity $1 - s(H, \eta, \beta, B)$:

$$1 - s(H,\eta,\beta,B) \;\le\; \min\{4\eta\lambda_{\min},\; 1-\beta\} \;\le\; \min\left\{\frac{8B\lambda_{\min}}{\operatorname{tr}(H)},\; \frac{8(1+\beta)\lambda_{\min}}{(1-\beta)\lambda_{\max}},\; 1-\beta\right\}.$$

Using the definition of $s^*(H, B)$, it remains to bound:

$$1 - s^*(H,B) \;\le\; \sup_{\beta\in[0,1)} \min\left\{\frac{8B\lambda_{\min}}{\operatorname{tr}(H)},\; \frac{8(1+\beta)\lambda_{\min}}{(1-\beta)\lambda_{\max}},\; 1-\beta\right\}.$$

The first term inside the minimum does not depend on β), so the proof consists in bounding:

$$\sup_{\beta\in[0,1)} \min\left\{\frac{8(1+\beta)\lambda_{\min}}{(1-\beta)\lambda_{\max}},\; 1-\beta\right\}.$$

The first function in the min is increasing in $\beta \in [0, 1)$, while the second is decreasing. Consequently, the sup is achived at the at the crossing point $\beta^\star$, where

$$\frac{8(1+\beta^\star)\lambda_{\min}}{(1-\beta^\star)\lambda_{\max}} = 1 - \beta^\star \quad\Longleftrightarrow\quad (1-\beta^\star)^2 = 8(1+\beta^\star)\,\frac{\lambda_{\min}}{\lambda_{\max}}.$$

Hence

$$\sup_{\beta\in[0,1)} \min\left\{\frac{8(1+\beta)\lambda_{\min}}{(1-\beta)\lambda_{\max}},\; 1-\beta\right\} = 1-\beta^\star = \sqrt{8(1+\beta^\star)\,\frac{\lambda_{\min}}{\lambda_{\max}}} = 4\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}},$$

using $1 + \beta^\star \le 2$. This completes the proof. ■

Corollary 12 (HB on power-law spectra) *Assume the eigenvalues of H satisfy $\lambda_i \eqsim i^{-a}$ for some $a > 1$. Then, the optimal spectral radius of HB satisfies:*

$$s^*(H,B) \;\gtrsim\; 1 - \min\left\{Bd^{-a},\; d^{-a/2}\right\},$$

where $\gtrsim$ absorbs universal constants. In particular, the transition to the accelerated regime occurs at batch size $B_{\text{HB}}^{\text{crit}} \eqsim d^{a/2}$

Proof Under the power-law spectrum

$$\lambda_i \eqsim i^{-a}, \qquad \lambda_{\max} \eqsim 1, \qquad \lambda_{\min} \eqsim d^{-a}, \qquad a > 1,$$

Theorem 7 gives

$$s^*(H, B) \;\geq\; 1 - 8 \min\left\{ \frac{B\lambda_{\min}}{\mathrm{tr}(H)}, \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \right\}.$$

Now,

$$\frac{B\lambda_{\min}}{\mathrm{tr}(H)} \;\eqsim\; \frac{B\, d^{-a}}{\sum_{i=1}^{d} i^{-a}}.$$

Since $a > 1$, we have

$$\sum_{i=1}^{d} i^{-a} \eqsim 1 \implies \frac{B\lambda_{\min}}{\mathrm{tr}(H)} \eqsim B\, d^{-a}.$$

Also,

$$\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \;\eqsim\; \sqrt{\frac{d^{-a}}{1}} = d^{-a/2}.$$

Therefore

$$\min\left\{ \frac{B\lambda_{\min}}{\mathrm{tr}(H)}, \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \right\} \eqsim \min\left\{ B\, d^{-a},\, d^{-a/2} \right\},$$

$$\implies s^*(H, B) \;\gtrsim\; 1 - \min\left\{ B\, d^{-a},\, d^{-a/2} \right\}.$$

■

Appendix B. Accelerated SGD Analysis

We now turn our attention to Accelerated SGD (ASGD) algorithm and we establish rates for it in a similar technical way. We begin by establishing the rates for the deterministic operator $\mathcal{T}_\infty$, followed by deterministic and stochastic conditions for the learning rate.

B.1. ASGD Operator Setup

The analysis for ASGD follows a very similar pattern as HB in Section 3.1. Note that we assume $0 < \zeta \leq 1$ and $0 < \beta < 1$. The ASGD update rule from Equation 1 can be written as a linear recursion in the following augmented state:

$$\begin{bmatrix} w_{t+1} - w^* \\ w_{t+1} - w^* + \eta\mu_t \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \eta(\zeta + 1 - \beta)\bar{X}_t & -\beta I \\ I - \eta\zeta\bar{X}_t & 0 \end{bmatrix} \begin{bmatrix} w_t - w^* \\ w_t - w^* + \eta\mu_{t-1} \end{bmatrix} \tag{15}$$

We again decompose the transition matrix into a deterministic and a stochastic component:

$$M_t = \begin{bmatrix} -\eta(\zeta + 1 - \beta)(\bar{X}_t - H) & 0 \\ -\eta\zeta(\bar{X}_t - H) & 0 \end{bmatrix} \qquad \tilde{A} = \begin{bmatrix} (1+\beta)I - \eta(\zeta + 1 - \beta)\, H & -\beta I \\ I - \eta\zeta H & 0 \end{bmatrix}$$

Note that all the randomness is in M_t. Denoting by $z_t = \begin{bmatrix} w_t - w^* \\ w_t - w^* + \eta\mu_{t-1} \end{bmatrix}$ and computing its covariance we get:

$$\Sigma_{t+1} = A\Sigma_t A^T + \frac{\eta^2}{B}\begin{bmatrix} (\zeta + (1-\beta))^2 \operatorname{Tr}\big(H\Sigma_t^{11}\big)H & \zeta(\zeta+1-\beta)\operatorname{Tr}\big(H\Sigma_t^{11}\big)H \\ \zeta(\zeta+1-\beta)\operatorname{Tr}\big(H\Sigma_t^{11}\big)H & \zeta^2 \operatorname{Tr}\big(H\Sigma_t^{11}\big)H \end{bmatrix}$$

Similarly, we will express all matrices in the eigenbasis of H. Reusing the same notation as in Section 3.1 after rotation in the eigenbasis of H we get Λ and $S_t = Q^\top \Sigma_t Q$ and $A = Q^\top \tilde{A} Q$. After computing the 4th moment term using Proposition 5, the recursion becomes:

$$S_{t+1} = AS_tA^\top + \frac{\eta^2}{B}\begin{bmatrix} (\zeta + (1-\beta))^2 \operatorname{Tr}\big(\Lambda S_t^{11}\big)\Lambda & \zeta(\zeta+1-\beta)\operatorname{Tr}\big(\Lambda S_t^{11}\big)\Lambda \\ \zeta(\zeta+1-\beta)\operatorname{Tr}\big(\Lambda S_t^{11}\big)\Lambda & \zeta^2 \operatorname{Tr}\big(\Lambda S_t^{11}\big)\Lambda \end{bmatrix} \tag{16}$$

Let $\mathcal{T}_{H,\eta,\beta,\zeta,B}$ denote the corresponding linear covariance update operator for this update rule.
Denoting $\mathbf{p} = [(\zeta + (1-\beta))^2, \zeta(\zeta+1-\beta), \zeta(\zeta+1-\beta), \zeta^2]$.
Let $\gamma = \operatorname{Tr}(HS_t^{11})$ and $A = \operatorname{blkdiag}(A_1, \ldots, A_d)$, where each per–coordinate 2×2 block is

$$A_i = \begin{bmatrix} (1+\beta) - \eta(\zeta+1-\beta)\lambda_i & -\beta \\ 1 - \eta\zeta\lambda_i & 0 \end{bmatrix}$$

Pushing a vec through Equation 16 coupled with the fact that $\operatorname{vec}(AS_tA^\top) = (A \otimes A)\operatorname{vec}(S_t)$ gives us:

$$s\operatorname{vec}(S_t) = (A \otimes A)\operatorname{vec}(S_t) + \frac{\eta^2\gamma}{B}(\mathbf{p} \otimes \operatorname{vec}(\Lambda))$$

$$\implies (sI - A \otimes A)\operatorname{vec}(S_t) = \frac{\eta^2\gamma}{B}(\mathbf{p} \otimes \operatorname{vec}(\Lambda))$$

We can break the above equation up into d, 2×2 equations each using a block of A:

$$(sI - A_i \otimes A_i)\operatorname{vec}(S_{t,i}) = \frac{\eta^2\gamma\lambda_i}{B}\mathbf{p}$$

$$\implies \operatorname{vec}(S_{t,i}) = \frac{\eta^2\gamma\lambda_i}{B}(sI - A_i \otimes A_i)^{-1}\mathbf{p} \tag{17}$$

Let,

$$q := \zeta+1-\beta, \qquad a := (1+\beta)-\eta(\zeta+1-\beta)\lambda_i, \qquad b := -\beta, \qquad c := 1-\eta\zeta\lambda_i, \quad M(s)_i := (sI-A_i\otimes A_i)^{-1}.$$

Let $\{e_1, e_2, \ldots, e_d\}$ represent the standard basis vector. To simplify γ, we multiply Equation 17 by $\lambda_i e_1$ and sum over $i \in \{1, \ldots, d\}$ to get:

$$1 = F(s) := \frac{\eta^2}{B}\sum_i \lambda_i^2 e_1^\top M(s)_i\,\mathbf{p} \tag{18}$$

with

$$e_1^\top M(s)_i \, \mathbf{p} = \frac{(s-bc)\left(sq^2+b^2\zeta^2\right)+2ab\,s\,\zeta\,q}{(s+bc)\left((s-bc)^2-a^2s\right)}$$

$$= \frac{(s+\beta)\big(s(\zeta+1-\beta)^2+\beta^2\zeta^2\big)-2\beta(1+\beta)s\zeta(\zeta+1-\beta)+\beta\eta\zeta\lambda_i\big(s(\zeta+1-\beta)^2-\beta^2\zeta^2\big)}{(s-\beta+\beta\eta\zeta\lambda_i)\Big[(s-1)(s-\beta^2)+2\eta\lambda_i\big(s(1+\zeta-\beta^2)-\beta^2\zeta\big)+\eta^2\lambda_i^2\big(\beta^2\zeta^2-s(\zeta+1-\beta)^2\big)\Big]}$$

As argued for the Heavy-Ball case, one can show that the maximum eigenvalue of $\mathcal{T}_{H,\eta,\beta,\zeta,B}$ will occur at a symmetric PSD matrix, thus, justifying the vectorization of Equation 16.

Bounding $\mathcal{T}_\infty$. By the same argument as in the Heavy–Ball case, the spectral radius of the stochastic covariance operator dominates that of the deterministic operator:

$$\rho(\mathcal{T}_{H,\eta,\beta,\zeta,B}) \geq \rho(\mathcal{T}_\infty), \qquad \mathcal{T}_\infty(M) = AMA^\top.$$

Since $A = \mathrm{blkdiag}(A_1, \ldots, A_d)$, restricting to the space of symmetric matrices gives

$$\rho(\mathcal{T}_\infty) = \rho(A \otimes A) = \rho(A)^2 = \max_i \rho(A_i)^2.$$

Thus it suffices to understand the eigenvalues of each block

$$A_i = \begin{bmatrix} (1+\beta)-\eta(\zeta+1-\beta)\lambda_i & -\beta \\ 1-\eta\zeta\lambda_i & 0 \end{bmatrix}.$$

We now record two lower bounds on $\rho(A)^2$.

First bound (product bound / complex-root regime). Let $r_{i,+}, r_{i,-}$ be the two roots of the characteristic polynomial of A_i. Their product is

$$r_{i,+}r_{i,-} = \beta(1-\eta\zeta\lambda_i).$$

Hence

$$\rho(A_i)^2 = \max\{|r_{i,+}|, |r_{i,-}|\}^2 \geq |r_{i,+}r_{i,-}| = \beta(1-\eta\zeta\lambda_i).$$

Taking i corresponding to $\lambda_{\min}$, we obtain

$$\rho(A)^2 \geq \beta(1-\eta\zeta\lambda_{\min}),$$

and therefore

$$1-\rho(A)^2 \leq (1-\beta)+\eta\beta\zeta\lambda_{\min}. \tag{19}$$

Second bound (real-root regime). Let $A_{\min}$ denote the block corresponding to $\lambda_{\min}$, and let $r_+ \geq r_-$ be its two real roots. Then

$$\rho(A)^2 \geq \rho(A_{\min})^2 \geq r_+^2,$$

so

$$1-\rho(A)^2 \leq 1-r_+^2 \leq 2(1-r_+).$$

From the characteristic polynomial at $\lambda_{\min}$,

$$(1 - r_+)(1 - r_-) = \eta\lambda_{\min}(1 - \beta)(1 + \zeta).$$

Also, since $r_- + r_+ = a_{\min} \le 1 + \beta$, we have

$$2r_- \le 1 + \beta \qquad \Longrightarrow \qquad 1 - r_- \ge \frac{1 - \beta}{2}.$$

Therefore

$$1 - r_+ = \frac{\eta\lambda_{\min}(1 - \beta)(1 + \zeta)}{1 - r_-} \le 2\eta\lambda_{\min}(1 + \zeta),$$

and hence

$$1 - \rho(A)^2 \le 4\eta\lambda_{\min}(1 + \zeta). \tag{20}$$

Combining equation 19 and equation 20, we obtain

$$1 - \rho(A)^2 \lesssim \min\Big\{\eta\lambda_{\min}(1 + \zeta),\ (1 - \beta) + \eta\beta\zeta\lambda_{\min}\Big\}. \tag{21}$$

B.2. Deterministic learning-rate stability

Lemma 13 (Deterministic ASGD stability band) *Assume $\eta > 0$. The deterministic ASGD iteration*

$$S_{t+1} = AS_tA^\top$$

is stable, i.e. all eigenvalues of every A_i lie in the open unit disk, if and only if

$$0 < \eta < \frac{2(1 + \beta)}{\big(1 - \beta + \zeta(1 + \beta)\big)\lambda_{\max}}.$$

Proof For each coordinate i, write

$$A_i = \begin{bmatrix} a_i & -\beta \\ c_i & 0 \end{bmatrix}, \qquad a_i = (1 + \beta) - \eta(\zeta + 1 - \beta)\lambda_i, \qquad c_i = 1 - \eta\zeta\lambda_i.$$

Its characteristic polynomial is

$$r^2 - a_ir + \beta c_i = 0.$$

Applying the degree–2 Jury criterion to

$$r^2 + pr + q = 0 \qquad \text{with} \qquad p = -a_i, \quad q = \beta c_i,$$

the roots lie in the open unit disk if and only if

$$|q| < 1, \qquad 1 + p + q > 0, \qquad 1 - p + q > 0.$$

The condition $1 + p + q > 0$ becomes

$$1 - a_i + \beta c_i > 0,$$

which simplifies to

$$\eta\lambda_i(1 - \beta)(1 + \zeta) > 0.$$

This holds automatically for $\eta > 0$, $\lambda_i > 0$, $0 < \beta < 1$, and $\zeta > 0$.

The condition $1 - p + q > 0$ becomes

$$1 + a_i + \beta c_i > 0,$$

that is,

$$2(1+\beta) - \eta\lambda_i\Big((\zeta + 1 - \beta) + \beta\zeta\Big) > 0.$$

Since

$$(\zeta + 1 - \beta) + \beta\zeta = 1 - \beta + \zeta(1+\beta),$$

this is equivalent to

$$\eta < \frac{2(1+\beta)}{\big(1 - \beta + \zeta(1+\beta)\big)\lambda_i}.$$

Imposing this for every i yields the claimed condition at $\lambda_{\max}$. ■

B.3. A stochastic step-size cap

Let

$$s(H, \eta, \beta, \zeta, B) := \rho\big(\mathcal{T}_{H,\eta,\beta,\zeta,B}\big), \qquad \alpha(H, \eta, \beta, \zeta, B) := 1 - s(H, \eta, \beta, \zeta, B).$$

Recall from the secular equation equation 18 that any real solution of $F(s) = 1$ is an eigenvalue s of $\mathcal{T}_{H,\eta,\beta,\zeta,B}$. Since $F(s) \to 0$ as $s \to \infty$, stability implies

$$F(1) \le 1.$$

Evaluating the secular equation at $s = 1$ gives

$$F(1) = \frac{\eta}{B}\sum_{i=1}^{d} \lambda_i \frac{(1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta + 1 - \beta + \beta\zeta)}{\big((1-\beta) + \beta\eta\zeta\lambda_i\big)\big(2(1+\beta) - \eta\lambda_i(1 - \beta + \zeta(1+\beta))\big)}. \tag{22}$$

Define

$$S := \{i : \beta\eta\zeta\lambda_i \le 1 - \beta\}, \qquad S^c := \{i : \beta\eta\zeta\lambda_i > 1 - \beta\}. \tag{23}$$

Let

$$x_i := 1 - \beta(1 - \eta\zeta\lambda_i) = (1-\beta) + \beta\eta\zeta\lambda_i, \qquad N_i := (1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta + 1 - \beta + \beta\zeta).$$

Following the definition of x_i and N_i, Equation 22 can be written as,

$$F(1) = \frac{\eta}{B}\sum_{i} \lambda_i \frac{N_i}{x_i\,\big(2(1+\beta) - \eta\lambda_i\big(1 - \beta + \zeta(1+\beta)\big)\big)},$$

Since,

$$2(1+\beta) - \eta\lambda_i\big(1 - \beta + \zeta(1+\beta)\big) \le 2(1+\beta),$$

we have,

$$\frac{1}{x_i\,\big(2(1+\beta) - \eta\lambda_i\big(1 - \beta + \zeta(1+\beta)\big)\big)} \ge \frac{1}{2(1+\beta)x_i}.$$

$$\implies F(1) \geq \frac{\eta}{2B(1+\beta)} \sum_i \lambda_i \frac{N_i}{x_i}.$$

$$= 1 \frac{\eta}{2B(1+\beta)} \sum \lambda_i \frac{(1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta+1-\beta+\beta\zeta)}{\big(1-\beta(1-\eta\zeta\lambda_i)\big)}$$

Splitting in two cases:

Small eigenvalues: $\beta\eta\zeta\lambda_i \leq 1-\beta$

$$x_i = (1-\beta) + \beta\eta\zeta\lambda_i \leq 2(1-\beta) \implies \frac{1}{x_i} \geq \frac{1}{2(1-\beta)}.$$

Hence

$$F(1) \geq \frac{\eta}{4B(1+\beta)} \sum_{i \in S} \lambda_i \frac{N_i}{1-\beta}.$$

Now

$$N_i = (1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta+1-\beta+\beta\zeta) \geq (1-\beta^2)(1+\zeta),$$
$$\implies \frac{N_i}{1-\beta} \geq \frac{(1-\beta^2)(1+\zeta)}{1-\beta} = (1+\beta)(1+\zeta).$$

Substituting back we get,

$$F(1) \geq \frac{\eta(1+\zeta)}{4B} \sum_{i \in S} \lambda_i. \tag{24}$$

Higher eigenvalues: $\beta\eta\zeta\lambda_i \geq 1-\beta$

$$x_i = (1-\beta) + \beta\eta\zeta\lambda_i \leq 2\beta\eta\zeta\lambda_i \implies \frac{1}{x_i} \geq \frac{1}{2\beta\eta\zeta\lambda_i}.$$

Therefore

$$F(1) \geq \frac{\eta}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i \frac{N_i}{\beta\eta\zeta\lambda_i}.$$

$$\implies F(1) \geq \frac{1}{4B(1+\beta)\beta\zeta} \sum_{i \in S^c} \left[(1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta+1-\beta+\beta\zeta)\right].$$

$$= \frac{(1-\beta^2)(1+\zeta)}{4B(1+\beta)\beta\zeta} |S^c| + \frac{\eta(\zeta+1-\beta+\beta\zeta)}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i.$$

$$\implies F(1) \geq \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta} |S^c| + \frac{\eta(\zeta+1-\beta+\beta\zeta)}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i. \tag{25}$$

Summing up Equation 24, 25 we get:

$$F(1) \geq \frac{\eta(1+\zeta)}{4B} \sum_{i \in S} \lambda_i + \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta} |S^c| + \frac{\eta(\zeta+1-\beta+\beta\zeta)}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i.$$

Since,

$$\frac{\zeta+1-\beta+\beta\zeta}{1+\beta} = \zeta + \frac{1-\beta}{1+\beta},$$

we have,

$$F(1) \geq \frac{\eta(1+\zeta)}{4B}\sum_{i\in S}\lambda_i + \frac{\eta\zeta}{4B}\sum_{i\in S^c}\lambda_i + \frac{\eta(1-\beta)}{4B(1+\beta)}\sum_{i\in S^c}\lambda_i + \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta}|S^c|.$$

Since

$$\frac{\eta(1+\zeta)}{4B}\sum_{i\in S}\lambda_i = \frac{\eta}{4B}\sum_{i\in S}\lambda_i + \frac{\eta\zeta}{4B}\sum_{i\in S}\lambda_i,$$

we get

$$F(1) \geq \frac{\eta}{4B}\sum_{i\in S}\lambda_i + \frac{\eta\zeta}{4B}\operatorname{Tr}(H) + \frac{\eta(1-\beta)}{4B(1+\beta)}\sum_{i\in S^c}\lambda_i + \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta}|S^c|. \tag{26}$$

We now assume the power-law spectrum

$$\lambda_i \eqsim i^{-a}, \qquad \lambda_{\max} \eqsim 1, \qquad \lambda_{\min} \eqsim d^{-a}, \qquad a > 1.$$

Let $k := |S^c|$. Since the spectrum is monotone, S^c consists of the top k eigenvalues up to constants, so

$$\sum_{i\in S}\lambda_i \eqsim \sum_{i=k+1}^{d} i^{-a}, \qquad \operatorname{Tr}(H) \eqsim 1.$$

For $a > 1$ we have:

$$\sum_{i=k+1}^{d} i^{-a} \eqsim k^{1-a}.$$

Since all terms in equation 26 are nonnegative, we may drop the last two terms and obtain

$$F(1) \geq \frac{\eta}{4B}\sum_{i\in S}\lambda_i + \frac{\eta\zeta}{4B}\operatorname{Tr}(H).$$

Using the power-law estimates above, this gives

$$F(1) \gtrsim \frac{\eta}{B}k^{1-a} + \frac{\eta\zeta}{B}.$$

Since stability implies $F(1) \leq 1$, we obtain

$$1 \gtrsim \frac{\eta}{B}k^{1-a} + \frac{\eta\zeta}{B},$$

hence the stochastic step-size cap

$$\eta \lesssim \frac{B}{k^{1-a}+\zeta}. \tag{27}$$

B.4. Proof of the ASGD rate bound

We define the optimal ASGD spectral radius and gap at batch size B by

$$s^*(H,B) := \inf_{\eta>0,\ \beta\in[0,1),\ \zeta>0} s(H,\eta,\beta,\zeta,B), \qquad \alpha^*(H,B) := 1 - s^*(H,B).$$

For every stable choice of (η,β,ζ), the deterministic dominance argument gives

$$\alpha(H,\eta,\beta,\zeta,B) \le 1-\rho(A)^2.$$

Combining this with equation 21, we obtain

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \min\Big\{\eta\lambda_{\min}(1+\zeta),\ (1-\beta)+\eta\beta\zeta\lambda_{\min}\Big\}. \tag{28}$$

We split into two cases.

Case 1: $\zeta \ge 1$**.** In this regime, using equation 28 and equation 27,

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \eta\lambda_{\min}(1+\zeta) \lesssim \frac{1+\zeta}{\zeta}\lambda_{\min} \eqsim d^{-a}.$$

Thus the branch $\zeta \ge 1$ is at best of order d^{-a}, so it cannot yield acceleration.

Case 2: $0 < \zeta < 1$**.** In this regime, $1+\zeta \eqsim 1$, so the first term in equation 28 is

$$\eta\lambda_{\min}(1+\zeta) \eqsim \eta d^{-a}.$$

For the second term, using equation 23,

$$(1-\beta)+\eta\beta\zeta\lambda_{\min} \le (1-\beta)+\eta\zeta d^{-a} \lesssim \eta\zeta k^{-a}+\eta\zeta d^{-a} \lesssim \eta\zeta k^{-a},$$

since $k \le d$ implies $k^{-a} \ge d^{-a}$. Therefore

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \eta\,\min\{d^{-a},\,\zeta k^{-a}\}. \tag{29}$$

Now combine equation 29 with the two step-size caps equation 27 and Lemma 13. For every stable choice of (η,β,ζ) with $0<\zeta<1$,

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \min\left\{\frac{B\,d^{-a}}{k^{1-a}+\zeta},\ \frac{d^{-a}}{\zeta},\ \frac{B\,\zeta\,k^{-a}}{k^{1-a}+\zeta},\ k^{-a}\right\}. \tag{30}$$

At this point, η and β have been eliminated: the right-hand side is an upper bound valid for every stable (η,β,ζ), and it depends only on the induced threshold k and on ζ. Thus

$$\alpha^*(H,B) \lesssim \sup_{1\le k\le d,\ 0<\zeta<1} \min\left\{\frac{B\,d^{-a}}{k^{1-a}+\zeta},\ \frac{d^{-a}}{\zeta},\ \frac{B\,\zeta\,k^{-a}}{k^{1-a}+\zeta},\ k^{-a}\right\}. \tag{31}$$

We now optimize over ζ for fixed k. Set

$$\zeta_k := \left(\frac{k}{d}\right)^a.$$

This is the crossover point where

$$d^{-a} = \zeta k^{-a}.$$

For fixed k, define

$$A_k(\zeta) := \frac{B\, d^{-a}}{k^{1-a} + \zeta}, \qquad B_k(\zeta) := \frac{d^{-a}}{\zeta}, \qquad C_k(\zeta) := \frac{B\, \zeta\, k^{-a}}{k^{1-a} + \zeta}, \qquad D_k := k^{-a}.$$

Then A_k and B_k are decreasing in ζ, C_k is increasing in ζ, and D_k is constant.

If $0 < \zeta \le \zeta_k$, then $\zeta k^{-a} \le d^{-a}$, so

$$A_k(\zeta) \ge C_k(\zeta), \qquad B_k(\zeta) \ge D_k.$$

Hence in this region

$$\min\{A_k(\zeta), B_k(\zeta), C_k(\zeta), D_k\} = \min\{C_k(\zeta), D_k\},$$

which is nondecreasing in ζ.

If $\zeta_k \le \zeta < 1$, then $d^{-a} \le \zeta k^{-a}$, so

$$A_k(\zeta) \le C_k(\zeta), \qquad B_k(\zeta) \le D_k.$$

Hence in this region

$$\min\{A_k(\zeta), B_k(\zeta), C_k(\zeta), D_k\} = \min\{A_k(\zeta), B_k(\zeta)\},$$

which is nonincreasing in ζ.

Therefore, for each fixed k, the right-hand side of equation 31 is maximized at the crossover point $\zeta = \zeta_k$. Substituting $\zeta_k = (k/d)^a$, we obtain

$$\alpha^*(H, B) \lesssim \sup_{1 \le k \le d} \min \left\{ \frac{B\, d^{-a}}{k^{1-a} + \left(\frac{k}{d}\right)^a},\ k^{-a} \right\}. \tag{32}$$

We now optimize over k.

Regime I: $B \lesssim 1$**.** Let

$$f(k) := k^{1-a} + \left(\frac{k}{d}\right)^a.$$

Its minimum is obtained by balancing the two terms:

$$k^{1-a} \eqsim \left(\frac{k}{d}\right)^a \qquad \Longrightarrow \qquad k \eqsim d^{\frac{a}{2a-1}}.$$

At this value,

$$f(k) \eqsim d^{-\frac{a(a-1)}{2a-1}},$$

and therefore

$$\frac{B\, d^{-a}}{f(k)} \eqsim B\, d^{-\frac{a^2}{2a-1}}.$$

Also,

$$k^{-a} \eqsim d^{-\frac{a^2}{2a-1}}.$$

Hence, when $B \lesssim 1$, the first term is the smaller one, and we obtain

$$\alpha^*(H, B) \lesssim B\, d^{-\frac{a^2}{2a-1}}.$$

Regime II: $1 \lesssim B \lesssim d^{1/2}$**.** In this regime, the optimum is obtained by balancing the two terms in equation 32:

$$\frac{B\, d^{-a}}{k^{1-a} + \left(\frac{k}{d}\right)^a} \eqsim k^{-a}.$$

Equivalently,

$$B \eqsim d^a k^{-a} \left(k^{1-a} + \left(\frac{k}{d}\right)^a \right) = d^a k^{1-2a} + 1.$$

For $B \gtrsim 1$, this gives

$$k \eqsim \left(\frac{d^a}{B}\right)^{\frac{1}{2a-1}}.$$

Substituting back,

$$\alpha^*(H, B) \lesssim k^{-a} \eqsim B^{\frac{a}{2a-1}}\, d^{-\frac{a^2}{2a-1}}.$$

Regime III: deterministic ceiling. Finally, stochastic noise can only slow convergence, so

$$\alpha^*(H, B) \leq \alpha^*_{\text{det}}(H),$$

where $\alpha^*_{\text{det}}(H)$ is the optimal full-batch ASGD gap. The standard deterministic ASGD optimization gives

$$\alpha^*_{\text{det}}(H) \lesssim d^{-a/2}.$$

Therefore

$$\alpha^*(H, B) \lesssim d^{-a/2} \qquad \text{for all } B.$$

The crossover expression

$$B^{\frac{a}{2a-1}}\, d^{-\frac{a^2}{2a-1}}$$

matches the deterministic ceiling $d^{-a/2}$ exactly when $B \eqsim d^{1/2}$. Combining the three regimes, we obtain

$$\alpha^*(H, B) \lesssim \begin{cases} B\, d^{-\frac{a^2}{2a-1}}, & B \lesssim 1, \\ B^{\frac{a}{2a-1}}\, d^{-\frac{a^2}{2a-1}}, & 1 \lesssim B \lesssim d^{1/2}, \\ d^{-a/2}, & B \gtrsim d^{1/2}. \end{cases} \tag{33}$$

Appendix C. Related Work

Acceleration with Noisy Gradients. The suboptimality of GD [3] for quadratic models has been well studied in literature [22]. In their seminal works, Polyak [24] and Nesterov [21] have proposed different variants of momentum, namely heavy ball and Nesterov's accelerated gradient method, which improve the deterministic rate to $\mathcal{O}(\sqrt{\kappa})$. Several works have shown that for SGD with batch size 1, both HB and NAG do not improve its compute efficiency [11, 12]. Extensions to these algorithms have been studied by introducing an extra momentum buffer [10–12], a family of algorithms commonly refered to as accelerated SGD [8, 20], which provably improve the contraction at batch size 1.

Closely related to our work is the work of Lee et al. [14], who analyze heavy-ball momentum for high-dimensional random least squares in a proportional asymptotic regime. They consider mini-batch sizes β satisfying $\beta/n \to \zeta > 0$ as the sample size n and feature dimension d grow with d/n fixed, and show that acceleration appears only once the batch fraction crosses a spectrum-dependent implicit conditioning ratio (ICR), a notion analogous to a critical batch fraction. For upper bounds, Liu and Belkin [15] established upper bounds at arbitrary batch size for an algorithm named MaSS, which is schematically similar to ASGD.

ASGD Variants in Practice. Several momentum schemes based on ASGD have been used in practice, in particular for large language model (LLM) pretraining. Morwani et al. [20] have shown that the Schedule-Free algorithm [7] can be rewritten as an ASGD equivalent. AdEMAMix [23] uses a similar scheme, based on a fast and slow momentum buffer updated with different EMA parameters. Interestingly, Lion [4], an algorithm discovered via genetic algorithms, also interpolates between the current gradient and momentum, before applying `sign` on the update. Several other methods have been proposed in the literature that aggregate over more than 1 momentum buffer [16, 17]. DANA [2, 8] is parametrically related to ASGD: it uses a single momentum buffer together with a direct gradient path in the parameter update, but, notably, chooses the parameters as a function of the model size and training time, rather than keeping them constant.

Nature of the bounds. Our results are finite-dimensional, discrete-time lower bounds. For HB, the bound holds for arbitrary finite spectra; for ASGD, the stated comparison holds under the two-sided power-law spectral assumption. In both cases, the bounds hold uniformly over all stable parameter choices, including aggressive choices near the edge of stability, and only hide universal constants. Although we summarize the power-law consequences using large-d scaling notation, these scalings come from finite-d inequalities rather than asymptotic limits. For heavy ball, Lee et al. [14] provide a lower bound in a proportional limit, showing that there is no acceleration in the small batch size regime.

Divergence thresholds. The techniques used in this work to derive the lower bounds also, implicitly, characterize the edge of stability [5, 6] conditions of the discussed algorithms. Recently, Andreyev et al. [1] have derived stability thresholds for heavy ball and Nesterov momentum, which we also implicitly recover through the proof of Theorem 1.

Appendix D. Experiments

We run synthetic experiments in linear regression on quadratics with power law data with $a = 2.0$, showing that the observed scaling is consistent with the lower bounds. We set the problem size to be $D = 50000$ and train for $N = 500000$, at batch sizes $B \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 1024\}$, averaged over 50 seeds. For both HB and ASGD we sweep over learning rate $\eta \in \{10^{-5}, 3 \cdot 10^{-5}, 10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}, 10^{-1}, 3 \cdot 10^{-1}, 1.0, 2.0, 3.0, 5.0, 10.0\}$, momentum EMA parameter $\beta \in \{0.8, 0.9, 0.95, 0.99, 0.999, 0.9999\}$ and ASGD hyperparameter $\zeta \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$, with $\zeta = 0$ for HB, and we plot each curve at the best set of hyperparameters. We plot the number of steps required to reach a target loss as a function of the batch size.