Compute Efficiency and Serial Runtime Tradeoffs for Stochastic Momentum Methods

Depen Morwani* Harvard University
Alexandru Meterez* Harvard University
Pranav Ajit Nair* Harvard University
Sham Kakade Harvard University

Abstract

Stochastic momentum methods such as heavy ball (HB), Nesterov momentum, and variants of Accelerated SGD (ASGD) [Kidambi et al., 2018] 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

Training deep neural networks is typically done via a first order optimization method, such as gradient descent (GD), or in the case of large scale datasets and models that are memory prohibitive through minibatch Stochastic Gradient Descent (SGD) [Robbins and Monro, 1951], with the latter method taking gradient steps on a random subsample of the population. While most open source large scale models [Grattafiori et al., 2024, Team et al., 2026, Guo et al., 2025, Yang et al., 2025] are trained with preconditioned SGD, typically Adam variants [Kingma and Ba, 2014, Loshchilov and Hutter, 2017, Vyas et al., 2024], or the more recent approximate second order methods [Martens and Grosse, 2015, Gupta et al., 2018, Jordan et al., 2024], they all generally bake in a version of momentum. Assuming a noiseless quadratic model, under exact gradients classical momentum methods such as heavy ball [Polyak, 1964] and Nesterov's accelerated gradient [Nesterov, 1983, 2013] improve over the GD rate of $\mathcal{O}(\kappa)$ to $\mathcal{O}(\sqrt{\kappa})$. However, this acceleration does not extend to the stochastic regime [Kidambi et al., 2018], 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 [Shallue et al., 2019, Fu et al., 2023, Kunstner et al., 2023, Wang et al., 2023, Zhang et al., 2019]. Most recently, Marek et al. [2025] 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

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

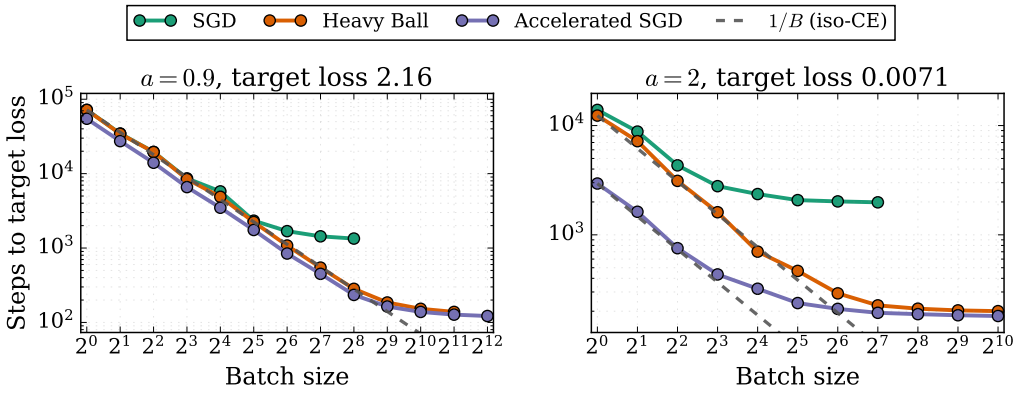

Figure 1: Batch-size tradeoffs for stochastic momentum methods in synthetic linear regression with power-law spectra. Plots show the number of serial steps needed to reach a fixed target loss as the batch size varies, after tuning hyperparameters for each method. **(Left)** For slow decaying power law 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 begins trading off CE for serial runtime.

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. [2018] and the asymptotic characterization of Lee et al. [2022], Ferbach et al. [2025]. We refer the reader to Section 2 where we discuss these works further. Based on the theoretical understanding, new momentum schemes have been proposed combining multiple momentum buffers that provably achieve acceleration at batch size 1 [Jain et al., 2018, Vaswani et al., 2019, Liu and Belkin, 2018, Gupta et al., 2024, Ferbach et al., 2025, Pagliardini et al., 2024], a family of methods we will refer to as Accelerated SGD (ASGD) [Kidambi et al., 2018, Jain et al., 2018, Ferbach et al., 2025, Lee et al., 2022]. While these methods achieve accelerated rates and thus better compute efficiency at batch size 1, this regime is far from practical from a serial runtime point of view, especially in the current pretraining setups where the number of tokens can scale to several trillions.

Thus, in large-scale pretraining, the key question is not only whether an optimizer improves sample complexity, but also whether it can maintain those gains at the large batch sizes needed to reduce serial training time. While keeping the data size fixed, one can linearly increase the batch size until the critical batch size (CBS) [Zhang et al., 2024, Shallue et al., 2019, Merrill et al., 2025, Meterez et al., 2025b] while achieving a linear decrease in the serial cost. 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, with matching upper bounds remaining an open question.

1.1 Contributions

In the deterministic, full-batch regime, the classical story is clear: on a κ-conditioned quadratic, GD has contraction gap of order κ^{-1}, while HB achieves the accelerated gap of order $\kappa^{-1/2}$. Equivalently, HB reduces the number of serial iterations by a factor of order $\sqrt{\kappa}$. Our goal is to understand what remains of this acceleration in the stochastic mini-batch regime, where we care about *both* the serial runtime and the compute efficiency.

A useful way to compare stochastic momentum methods is to separate these two quantities. For a given algorithm, let $s^*(H, B)$ denote the best-achievable spectral radius, or contraction factor, at batch size B and with covariance spectrum H (suppressing dependence on the algorithm). We define the corresponding contraction gap by

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

In the full-batch limit, this gap satisfies $\alpha_{\mathrm{GD}}(\infty) \asymp \kappa^{-1}$ and $\alpha_{\mathrm{HB}}(\infty) \asymp \kappa^{-1/2}$, which is the classical deterministic acceleration of HB over GD.

Algorithm	Batch-size regime	Serial runtime $\alpha(B)^{-1}$	Compute efficiency $\alpha(B)/B$
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 compute-efficiency 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 compute efficiency scales as $\alpha(B)/B$, where $\alpha(B) = 1 - s^*(H, B)$ is the best-achievable spectral gap. SGD and HB have the same compute-efficiency frontier $\lambda_{\min}/\mathrm{Tr}(H)$, but HB preserves this frontier over a larger batch-size window, up to $B \asymp (\mathrm{Tr}(H)/\lambda_{\max})\sqrt{\kappa}$, thereby reducing serial runtime from κ to $\sqrt{\kappa}$ before saturating.

The number of serial steps needed to reach a fixed target error scales as $1/\alpha(B)$, while the total gradient-query cost , or total FLOPs up to constants , scales as $B/\alpha(B)$. Equivalently, the compute efficiency (CE) scales as

$$\frac{\alpha(B)}{B}.$$

Thus, higher CE means fewer total FLOPs to reach a target accuracy . If $\alpha(B)$ grows linearly with B, increasing the batch size reduces serial runtime without hurting CE. If $\alpha(B)$ grows sublinearly with B, larger batches still improve serial runtime, but only by sacrificing CE.

Our first main result characterizes how this deterministic acceleration appears as the batch size is varied, where we compare HB to SGD, and is summarized in Table 1. The result holds for *arbitrary* spectra.

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

More precisely, we prove a lower bound showing that HB cannot improve over the optimal CE scaling of SGD at any batch size. At small batch sizes, HB has the same linear- scaling behavior as SGD. Above the SGD critical batch size, however, HB can continue to convert larger batches into fewer serial steps while preserving the same CE scale. This continues until HB reaches its deterministic accelerated scale. Thus, HB's benefit is not a better best-achievable CE frontier, but rather a larger batch-size window over which SGD-level CE can be used to buy improved serial runtime .

In practice, several newer stochastic momentum methods are also being considered. The majority of these schemes are parametrically equivalent to ASGD, which is known to improve CE over SGD at small batch sizes. Our second main result characterizes how this CE improvement changes with batch size.

ASGD has a higher CE ceiling than HB at small batch sizes, but it decreases as batch size grows.

We show that ASGD's CE gains are concentrated in the small-batch regime. Unlike the case for HB, the serial-runtime picture for ASGD is more dependent on the full spectra. Increasing the batch size reduces ASGD's serial runtime, but it does so by spending some of ASGD's small-batch CE advantage .

More concretely, for rapidly decaying spectra, the resulting picture is as follows. At small batch sizes, ASGD has better CE than HB, consistent with known upper bounds at $B = 1$. As B increases, ASGD moves along a CE–serial tradeoff curve: serial runtime improves, but CE decreases relative to the small-batch optimum. Eventually, ASGD reaches the deterministic accelerated serial-runtime scale. Beyond this point, further increasing the batch size no longer improves serial runtime and only decreases CE at the usual $1/B$ rate .

Thus, ASGD improves the CE–serial frontier most strongly at small batch sizes; for larger batch sizes the advantage is more modest, because part of the small-batch CE gain has already been converted into serial-time improvement . For more slowly decaying spectra, ASGD and HB are comparable .

Table 2 summarizes the corresponding power-law scalings, and Figure 2 plots them in terms of compute (gradient queries $B/\alpha(B)$) and serial steps. Section 5 tests the qualitative picture in

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

Table 2: Serial-runtime and compute-efficiency implications of our lower bounds under power-law spectra $\lambda_i \asymp i^{-a}$ with $a > 1$. Serial runtime scales as $\alpha(B)^{-1}$ and compute efficiency scales as $\alpha(B)/B$, where $\alpha(B) = 1 - s^*(H, B)$ is the best-achievable spectral gap. For these spectra, SGD has only a constant-size linear-scaling window. HB preserves SGD-level CE over the larger window $B \lesssim d^{a/2}$, while ASGD improves small-batch CE but loses CE as B grows through the intermediate regime.

synthetic linear regression (Figure 1). For a slowly decaying spectrum, ASGD and HB nearly overlap, while HB improves serial runtime over SGD at essentially the same CE as SGD. For a rapidly decaying spectrum, ASGD improves small-batch CE over HB/SGD, and then spends this CE advantage to reduce serial runtime as the batch size grows.

In summary, HB preserves SGD-level CE over a larger batch-size range and uses this to improve serial runtime, but it does not improve the CE frontier itself. ASGD improves the small-batch CE frontier, but larger batches buy serial-time improvements by spending some of ASGD's CE advantage.

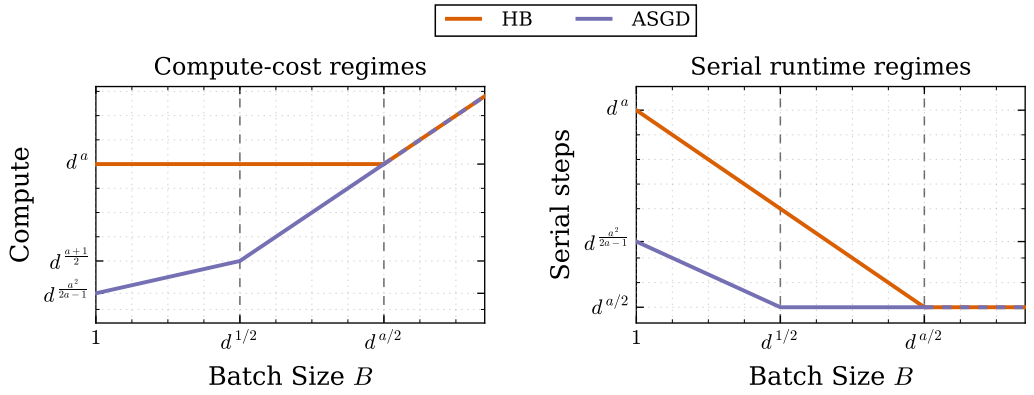

Figure 2: Illustration of the batch-size regimes implied by our bounds for power-law spectra (Table 2). **(Left)** Compute cost $B/\alpha(B)$ required to reach a fixed accuracy. HB stays at the SGD level d^a up to $B \asymp d^{a/2}$. ASGD starts lower, at $d^{a^2/(2a-1)}$, grows as $B^{(a-1)/(2a-1)}$ up to $d^{(a+1)/2}$ at $B \asymp d^{1/2}$, and then grows linearly. **(Right)** Serial steps $1/\alpha(B)$. HB decreases linearly until $B \asymp d^{a/2}$; ASGD decreases as $B^{-a/(2a-1)}$ and reaches the accelerated floor $d^{a/2}$ already at $B \asymp d^{1/2}$. Beyond $d^{a/2}$ the two coincide (dashed).

2 Related Work

Acceleration with Noisy Gradients. The suboptimality of GD [Cauchy et al., 1847] for quadratic models has been well studied in literature [Nesterov, 2013]. In their seminal works, Polyak [1964] and Nesterov [1983] 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 [Jain et al., 2018, Kidambi et al., 2018]. Extensions to these algorithms have been studied by introducing an extra momentum buffer [Kidambi et al., 2018, Jain et al., 2018, Gupta et al., 2024], a family of

algorithms commonly referred to as accelerated SGD [Morwani et al., 2025, Ferbach et al., 2025], which provably improve the contraction at batch size 1.

Closely related to our work is the work of Lee et al. [2022], 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 [2018] 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. [2025] have shown that the Schedule-Free algorithm [Defazio et al., 2024] can be rewritten as an ASGD equivalent. AdEMAMix [Pagliardini et al., 2024] uses a similar scheme, based on a fast and slow momentum buffer updated with different EMA parameters. Interestingly, Lion [Chen et al., 2023], 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 [Lucas et al., 2018, Ma and Yarats, 2018]. DANA [Ferbach et al., 2025, Bordelon and Mori, 2026] 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 and constants that may depend on the fixed power-law exponent a. 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. [2022] 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 [Cohen et al., 2021, 2022] conditions of the discussed algorithms. Recently, Andreyev et al. [2026] have derived stability thresholds for heavy ball and Nesterov momentum, which we also implicitly recover through the proof of Theorem 1.

3 Background and Preliminaries

In this section, we first introduce the setup necessary for our main theoretical results.

Definitions. We denote as $f(x) \lesssim g(x)$ if there exists a constant $c > 0$ such that $f(x) \leq cg(x)$ for any x in the domain. Moreover, we denote as $f(x) \eqsim g(x)$ if $g(x) \lesssim f(x) \lesssim g(x)$. For 2 matrices A and B we say that $A \preceq B$ if and only if $B - A \succeq 0$, where a matrix $B \succeq 0$ denotes that B is positive semidefinite (PSD). We study stochastic optimization for a noiseless linear regression problem with Gaussian covariates i.e. a consistent linear system.

We define our notion of compute to be 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 [Wu et al., 2022a, Zou et al., 2021, Meterez et al., 2025a], the error contracts geometrically at a rate governed by the spectral radius of $\mathcal{T}$.

We train online, without repeating over the data. At each iteration, we sample from the population a minibatch of size B of covariates $\{(x_i, y_i)\}_{i=1}^{B}$ such that:

$$x_i \sim \mathcal{N}(0, H), \qquad y_i = \langle w^*, x_i \rangle,$$

where $w^* \in \mathbb{R}^d$ is the minimizer and $H \in \mathbb{R}^{d\times d}$, $H \succ 0$ is the population covariance matrix of the covariates. We denote the eigendecomposition of H by

$$H = Q\Lambda Q^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \qquad \lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_d = \lambda_{\min} > 0$$

We will use $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ to denote the conditioning number of H. Where stated, we will additionally specialize to a power-law spectrum with exponent a defined as:

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

Moreover, we denote the population risk as:

$$\mathcal{R}(w) := \frac{1}{2}\,\mathbb{E}\big[(\langle w, x\rangle - y)^2\big]$$

where the expectation is taken over the randomness induced by the stochastic gradient updates. We will recall a basic identity used repeatedly in the stochastic covariance analysis, namely the Gaussian fourth-moment formula.

Proposition 1. *For any deterministic matrix $\Sigma \succeq 0$ and $x \sim \mathcal{N}(0, H)$ we have the following equation:*

$$\mathbb{E}[xx^\top \Sigma xx^\top] = 2H\Sigma H + \mathrm{Tr}(H\Sigma)H \preceq 3\,\mathrm{Tr}(H\Sigma)H. \tag{1}$$

For a minibatch of size B of covariates with empirical covariance $\bar{X} = \frac{1}{B}\sum_{i=1}^B x_i x_i^\top$, Proposition 1 generalizes as:

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

Background. We analyze the ASGD algorithm [Jain et al., 2018, Kidambi et al., 2018, Morwani et al., 2025], with the following update rule:

$$\begin{aligned} \mu_t &= \beta\mu_{t-1} + (1-\beta)g_t \\ w_{t+1} &= w_t - \eta(\mu_t + \zeta g_t) \end{aligned} \tag{3}$$

We denote as μ_t the momentum buffer at time t, the current iterate w_t and the current minibatch gradient g_t. The hyperparameters of this algorithm are the learning rate η, the momentum EMA parameter β and the weight placed on the current gradient ζ. Note that by setting $\zeta = 0$, we recover the HB update:

$$\begin{aligned} \mu_t &= \beta\mu_{t-1} + (1-\beta)g_t \\ w_{t+1} &= w_t - \eta\mu_t \end{aligned} \tag{4}$$

Therefore, intuitively, since ASGD has an extra hyperparameter we can optimize over, it is possible that it can improve over the rates achieved by HB.

4 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 [Wu et al., 2022a, Zou et al., 2021, Wu et al., 2022b, Meterez et al., 2025a], 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.

We first state a general HB lower bound valid for arbitrary spectra, and then specialize it to power-law spectra to identify the critical batch-size transition. We next give an analogous bound for ASGD, showing that for power-law spectra the ASGD lower bound has a strictly better scaling than the HB lower bound in the small-batch regime and reaches the accelerated scale at a smaller batch size.

4.1 Heavy Ball

We can rewrite the HB update from equation 4 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:

$$\begin{bmatrix} w_{t+1} - w^* \\ w_t - w^* \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \eta(1-\beta)\bar{X}_t & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} w_t - w^* \\ w_{t-1} - w^* \end{bmatrix} \tag{5}$$

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 5 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 1 and upper bound $H\Sigma H \preceq \mathrm{Tr}(H\Sigma)H$. We have a bound on Σ_{t+1} on both sides now:

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

For the lower-bound analysis, we consider the lower comparison recursion. 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:

$$S_{t+1} = A\,S_t\,A^\top \;+\; \eta^2(1-\beta)^2 \begin{bmatrix} \frac{1}{B}\,\Lambda\,\mathrm{Tr}(\Lambda S_t^{11}) & 0 \\ 0 & 0 \end{bmatrix} \tag{6}$$

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:

$$A_i = \begin{bmatrix} a_i & -\beta \\ 1 & 0 \end{bmatrix}, \qquad a_i := (1+\beta) - \eta(1-\beta)\lambda_i. \tag{7}$$

Let $\mathcal{T}_{H,\eta,\beta,B}$ denote the corresponding linear covariance update operator on symmetric $2d \times 2d$ matrices as defined in equation 6, 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 3), 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

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

This brings us to our first main result.

Theorem 1 (HB Compute Efficiency Lower Bound)**.** *For any data covariance matrix H and mini-batch 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 lower bound reaches the accelerated rate at batch size $B_{\mathrm{HB}}^{\mathrm{crit}} = \frac{\mathrm{Tr}(H)\sqrt{\kappa}}{\lambda_{\max}}$.

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 $B < B_{\text{HB}}^{\text{crit}}$, the algorithm cannot improve over the SGD scaling, whereas for batch size $B \geq B_{\text{HB}}^{\text{crit}}$ the best-achievable spectral gap is bounded above by the asymptotic contraction rate . Therefore, one can conclude that in order to minimize the number of serial steps without wasting compute, the optimal batch size to train at is $B_{\text{HB}}^{\text{crit}}$. In Corollary 1, we specialize the HB lower bound to a power law spectrum, in order to compare the rate with the ASGD bound in Section 4.2.

Corollary 1 (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:*

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

where $\gtrsim$ absorbs universal constants. The lower bound reaches the accelerated scale at batch size: $B_{\text{HB}}^{\text{crit}} \eqsim d^{a/2}$

We defer the full proofs to Appendix A. In the following section, we establish spectral lower bounds for ASGD.

4.2 Accelerated SGD

The analysis for ASGD follows a very similar pattern as HB in Section 4.1. Note that we assume $\zeta > 0$ and $0 < \beta < 1$. The ASGD update rule from Equation 3 can be written as a linear recursion in the following augmented state:

$$\begin{bmatrix} w_{t+1} - w^* \\ w_{t+1} - w^* + \eta\mu_t \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \eta(\zeta + 1 - \beta)\bar{X}_t & -\beta I \\ I - \eta\zeta\bar{X}_t & 0 \end{bmatrix} \begin{bmatrix} w_t - w^* \\ w_t - w^* + \eta\mu_{t-1} \end{bmatrix} \tag{8}$$

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 4.1 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 1, the recursion becomes:

$$S_{t+1} = AS_tA^\top + \frac{\eta^2}{B}\begin{bmatrix} (\zeta + (1-\beta))^2\, \operatorname{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda & \zeta(\zeta + 1 - \beta)\, \operatorname{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda \\ \zeta(\zeta + 1 - \beta)\, \operatorname{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda & \zeta^2\, \operatorname{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda \end{bmatrix} \tag{9}$$

Let $\mathcal{T}_{H,\eta,\beta,\zeta,B}$ denote the corresponding linear covariance update operator for this update rule. Then, we have the following statement.

Theorem 2 (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 gap of ASGD satisfies:*

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

Constants may depend on a and the spectral comparison constants. In particular, the lower bound saturates at the accelerated scale at batch size $B_{\text{ASGD}}^{\text{crit}} \eqsim d^{1/2}$.

Theorem 2 distinguishes 2 scaling regimes as a function of batch size. In the first regime, there are *diminishing returns* indicating that increasing the batch size gives us a sublinear decrease in the serial runtime, since for $a > 1$ the batch exponent $a/(2a-1) < 1$ (note that for large a this term behaves roughly like $\sqrt{B}$). Finally, for large batch size we reach a *saturation* regime, saturating the lower bound at the accelerated scale. Increasing the batch size any further would not yield any improvements in the serial runtime, and will be, instead, wasting 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 2.

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

$$B^{\text{crit}}_{\text{HB}} \eqsim d^{a/2}, \qquad B^{\text{crit}}_{\text{ASGD}} \eqsim d^{1/2}.$$

Hence

$$\frac{B^{\text{crit}}_{\text{HB}}}{B^{\text{crit}}_{\text{ASGD}}} \eqsim d^{(a-1)/2},$$

so 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.

Compute-serial runtime tradeoff. We now summarize the implications of the lower bounds above. Theorem 1 shows that, for arbitrary spectra, HB cannot improve over the SGD CE frontier. However, note that HB can use larger batches to reduce serial runtime, but this does not correspond to an improved compute-efficiency frontier over SGD. Beyond the HB critical batch size, the lower bound saturates at the deterministic accelerated scale.

Under power-law spectra with $a > 1$, Theorem 2 shows that ASGD has a better small-batch spectral-gap scaling than HB and that its lower bound reaches the accelerated scale at the smaller batch size $B^{\text{crit}}_{\text{ASGD}} \asymp d^{1/2}$, as stated in Corollary 2. Thus ASGD offers a sharper compute–serial tradeoff: small batches preserve its compute-efficiency advantage , while increasing the batch size up to $B^{\text{crit}}_{\text{ASGD}}$ converts part of this advantage into reduced serial runtime . We empirically confirm these regimes in synthetic linear-regression experiments in Section 5.

5 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. Figure 1 shows the results.

6 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 theoretical perspective, our lower bound for ASGD

improves over that of [Kidambi et al., 2018], 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.

Acknowledgements

The authors would like to thank Alex Damian and Jingfeng Wu, for helpful discussions. The authors would also like to thank Max Shad and Bala Desinghu for their help with the cluster. AM, DM, PN acknowledge the support of a Kempner Institute Graduate Research Fellowship. AM, SK, DM and PN acknowledge that this work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. SK and DM acknowledge support from the Office of Naval Research under award N0001422-1-2377 and the National Science Foundation Grant under award #IIS 2229881. DM is also supported by a Simons Investigator Fellowship, NSF grant DMS-2134157, DARPA grant W911NF2010021,and DOE grant DE-SC0022199.

References

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.

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.

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

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.

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.

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.

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.

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

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

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.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

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.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

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

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.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

William Merrill, Shane Arora, Dirk Groeneveld, and Hannaneh Hajishirzi. Critical batch size revisited: A simple empirical approach to large-batch language model training. *arXiv preprint arXiv:2505.23971*, 2025.

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*, 2025a.

Alexandru Meterez, Depen Morwani, Jingfeng Wu, Costin-Andrei Oncescu, Cengiz Pehlevan, and Sham Kakade. Seesaw: Accelerating training by balancing learning rate and batch size scheduling. *arXiv preprint arXiv:2510.14717*, 2025b.

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.

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.

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

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

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

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.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

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.

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, 2022a.

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, 2022b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? *arXiv preprint arXiv:2410.21676*, 2024.

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.

A Heavy Ball Analysis

A.1 Proof of Proposition 1

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,

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

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 $D \preceq \mathrm{Tr}(D)\, I$. Therefore,

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

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

$$v^\top H\Sigma H v \le \mathrm{Tr}(H\Sigma)\, v^\top H v.$$

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

□

A.2 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:

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

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

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:

$$1 = \frac{2\eta^2(1-\beta)^2}{B}\sum_{i=1}^{d}[\lambda_i \quad 0 \quad 0 \quad 0](sI - A_i \otimes A_i)^{-1}\begin{bmatrix}\lambda_i\\0\\0\\0\end{bmatrix} \tag{10}$$

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

$$M_i = \begin{bmatrix} r_{i,+} & r_{i,-} \\ 1 & 1\end{bmatrix} \tag{11}$$

Equations (10) and (11) on further simplification 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 written as

$$1 \;=\; \frac{c}{B}\sum_{i=1}^{d}\frac{\lambda_i^2\,\phi_i(s)}{1 - c\lambda_i^2\phi_i(s)}\,, \qquad c = \eta^2(1-\beta)^2. \tag{12}$$

Validity of vectorizing Equation 9 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,

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

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) given the stochastic operator $\mathcal{T}_{H,\eta,\beta,B}$, gauging at the block diagonal eigenmatrices of the deterministic 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 deterministic 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 3, the spectral radius for $\mathcal{T}_{H,\eta,\beta,B}$ is obtained by a real eigenvalue and from Lemma 2, 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.3 The Zero-Noise Heavy Ball Analysis

In the deterministic setting, the Heavy-Ball iteration in Equation 6 reduces to

$$S_{t+1} = A S_t A^\top \tag{13}$$

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

Lemma 1 (Stability band of deterministic HB)**.** *Assume $\beta, \eta > 0$. If the Heavy–Ball iteration in Equation 13 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.4 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 3), 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 3 (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.5 Helper Lemmas

We will refer to any eigenvalue s solving the secular equation equation 12 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 $\mathcal{T}_{H,\eta,\beta,B}$:

$$1 \;=\; F(s) \;:=\; \frac{c}{B} \sum_{i=1}^{d} \frac{\lambda_i^2\,\phi_i(s)}{1 - c\,\lambda_i^2 \phi_i(s)}\,, \qquad c = \eta^2(1 - \beta)^2, \tag{14}$$

with

$$\phi_i(s) = \frac{s(s + \beta)}{(s - \beta)\,(s - r_{i,+}^2)\,(s - r_{i,-}^2)}\,, \quad r_{i,\pm} = \frac{a_i \pm \sqrt{a_i^2 - 4\beta}}{2}, \quad a_i = (1 + \beta) - \eta(1 - \beta)\lambda_i. \tag{15}$$

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 7). Let

$$\mathcal{T}^{(\infty)}(S) := ASA^\top \tag{16}$$

denote this deterministic (zero-noise) covariance operator.

Lemma 2 (Stochastic spectral radius dominates deterministic)**.** *For every (η, β, B) we have*

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

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

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

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 3 (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 $\mathcal{T} \in \big\{\, \mathcal{T}_{H,\eta,\beta,B},\ \mathcal{T}^{(\infty)} \,\big\}$. 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 $\lambda_\star \geq 0$ 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 2.

Proof of Lemma 2. 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 16, 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 3 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.6 Proof

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

Lemma 3 (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 $|s| \leq 1$. Then the step size must satisfy*

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

Proof. First, let us show that stability forces

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

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 $|s| \leq 1$, 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 $s \geq 1$). Using $x/(1-x) \geq x$ for $x \in [0, 1)$, we have

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

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

$$1 \;\geq\; \frac{\eta(1+\beta)}{B} \sum_{i=1}^{d} \frac{\lambda_i}{2(1+\beta) - \eta(1-\beta)\lambda_i}. \tag{17}$$

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

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

which completes the proof of this case.

We now show the second bound of $\eta \leq 2(1+\beta)/((1-\beta)\lambda_{\max})$ must hold. Lemma 2 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 1 implies $\rho(A) > 1$, contradicting our assumed stability. This proves our second claim. □

Lemma 4 (Gap in terms of η and β)**.** *For any stable HB–SGD parameters (η, β, B), the spectral gap satisfies*

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

Proof. By Lemma 2, the stochastic covariance operator $\mathcal{T}_{H,\eta,\beta,B}$ has spectral radius at least that of the deterministic (full-batch) HB operator:

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

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

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

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

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

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

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

Therefore

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

We now show $1 - s(H, \eta, \beta, B) \leq 4\eta\lambda_{\min}$. 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

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

Hence

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

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 $1 - \rho(A)^2 \leq (1 - r_{\min,+}^2) \leq 2(1 - r_{\min,+})$. 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 $r_- \leq r_+$, we must have $r_- \leq \sqrt{\beta}$. Using that $\beta \leq 1$ (stability),

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

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 3.

Proof of Theorem 3. Fix H and any stable HB–SGD parameters (η, β, B).

Combining Lemma 4 (the step size cap) with Lemma 3 (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 achieved 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 4 (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 $B_{\mathrm{HB}}^{\mathrm{crit}} \eqsim d^{a/2}$

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 3 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

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

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

□

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.

Recall that :

$$S_{t+1} = AS_tA^\top + \frac{\eta^2}{B}\begin{bmatrix} (\zeta + (1-\beta))^2\ \mathrm{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda & \zeta(\zeta+1-\beta)\ \mathrm{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda \\ \zeta(\zeta+1-\beta)\ \mathrm{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda & \zeta^2\ \mathrm{Tr}\big(\Lambda S_t^{11}\big)\,\Lambda \end{bmatrix}. \tag{18}$$

Let $\mathcal{T}_{H,\eta,\beta,\zeta,B}$ denote the corresponding linear covariance update operator on symmetric $2d \times 2d$ matrices, defined by

$$S_{t+1} = \mathcal{T}_{H,\eta,\beta,\zeta,B}(S_t) \tag{19}$$

Denoting $\mathbf{p} = [(\zeta + (1-\beta))^2, \zeta(\zeta+1-\beta), \zeta(\zeta+1-\beta), \zeta^2]$.

Let $\gamma = \mathrm{Tr}(HS_t^{11})$ and $A = \mathrm{blkdiag}(A_1, \ldots, A_d)$, 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 9 coupled with the fact that $\mathrm{vec}(AS_tA^\top) = (A \otimes A)\mathrm{vec}(S_t)$ gives us:

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

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

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

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

$$\implies \text{vec}(S_{t,i}) = \frac{\eta^2 \gamma \lambda_i}{B}(sI - A_i \otimes A_i)^{-1}\mathbf{p} \tag{20}$$

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 20 by $\lambda_i e_1$ and sum over $i \in \{1, \ldots, d\}$ to get:

$$1 = F(s) := \frac{\eta^2}{B}\sum_i \lambda_i^2 e_1^\top M(s)_i\,\mathbf{p} \tag{21}$$

with

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

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

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 9 .

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 $A = \text{blkdiag}(A_1, \ldots, A_d)$, 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

$$1 - \rho(A)^2 \leq (1-\beta) + \eta\beta\zeta\lambda_{\min}. \tag{22}$$

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 $r_- + r_+ = a_{\min} \leq 1 + \beta$, we have

$$2r_- \leq 1 + \beta \qquad \Longrightarrow \qquad 1 - r_- \geq \frac{1 - \beta}{2}.$$

Therefore

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

and hence

$$1 - \rho(A)^2 \leq 4\eta\lambda_{\min}(1 + \zeta). \tag{23}$$

Combining equation 22 and equation 23, we obtain

$$1 - \rho(A)^2 \lesssim \min\Big\{\eta\lambda_{\min}(1 + \zeta),\ (1 - \beta) + \eta\beta\zeta\lambda_{\min}\Big\}. \tag{24}$$

B.1 Deterministic learning-rate stability

Lemma 5 (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.2 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 21 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

$$F(1) = \frac{\eta}{B}\sum_{i=1}^{d} \lambda_i \frac{(1-\beta^2)(1+\zeta) + \beta\eta\zeta\lambda_i(\zeta+1-\beta+\beta\zeta)}{\big((1-\beta)+\beta\eta\zeta\lambda_i\big)\big(2(1+\beta)-\eta\lambda_i(1-\beta+\zeta(1+\beta))\big)}. \tag{25}$$

Define

$$S := \{i : \beta\eta\zeta\lambda_i \le 1-\beta\}, \qquad S^c := \{i : \beta\eta\zeta\lambda_i > 1-\beta\}. \tag{26}$$

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 25 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,

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

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

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

Splitting in two cases:

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

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

Hence

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

Now

$$N_i = (1-\beta^2)(1+\zeta)+\beta\eta\zeta\lambda_i(\zeta+1-\beta+\beta\zeta) \ge (1-\beta^2)(1+\zeta),$$

$$\implies \frac{N_i}{1-\beta} \ge \frac{(1-\beta^2)(1+\zeta)}{1-\beta} = (1+\beta)(1+\zeta).$$

Substituting back we get,

$$F(1) \ge \frac{\eta(1+\zeta)}{4B}\sum_{i\in S}\lambda_i. \tag{27}$$

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

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

Therefore

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

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

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

$$\implies F(1) \geq \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta}|S^c| + \frac{\eta(\zeta+1-\beta+\beta\zeta)}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i. \tag{28}$$

Summing up Equation 27 , 28 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

$$F(1) \geq \frac{\eta}{4B} \sum_{i \in S} \lambda_i + \frac{\eta\zeta}{4B} \operatorname{Tr}(H) + \frac{\eta(1-\beta)}{4B(1+\beta)} \sum_{i \in S^c} \lambda_i + \frac{(1-\beta)(1+\zeta)}{4B\beta\zeta}|S^c|. \tag{29}$$

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 29 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

$$\eta \lesssim \frac{B}{k^{1-a} + \zeta}. \tag{30}$$

B.3 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 24, we obtain

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \min\Big\{\eta\lambda_{\min}(1+\zeta),\ (1-\beta)+\eta\beta\zeta\lambda_{\min}\Big\}. \tag{31}$$

We split into two cases.

Case 1: $\zeta \ge 1$**.** In this regime, using equation 31 and equation 30 ,

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

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 31 is

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

For the second term, using equation 26,

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

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

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \eta\,\min\{d^{-a},\,\zeta k^{-a}\}. \tag{32}$$

Now combine equation 32 with the two step-size caps equation 30 and Lemma 5. For every stable choice of (η,β,ζ) with $0<\zeta<1$,

$$\alpha(H,\eta,\beta,\zeta,B) \lesssim \min\left\{\frac{B\,d^{-a}}{k^{1-a}+\zeta},\ \frac{d^{-a}}{\zeta},\ \frac{B\,\zeta\,k^{-a}}{k^{1-a}+\zeta},\ k^{-a}\right\}. \tag{33}$$

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

$$\alpha^*(H,B) \lesssim \sup_{1\le k\le d,\ 0<\zeta<1} \min\left\{\frac{B\,d^{-a}}{k^{1-a}+\zeta},\ \frac{d^{-a}}{\zeta},\ \frac{B\,\zeta\,k^{-a}}{k^{1-a}+\zeta},\ k^{-a}\right\}. \tag{34}$$

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 34 is maximized at the crossover point $\zeta = \zeta_k$. Substituting $\zeta_k = (k/d)^a$, we obtain

$$\alpha^*(H, B) \lesssim \sup_{1 \le k \le d} \min\left\{ \frac{B\, d^{-a}}{k^{1-a} + \left(\frac{k}{d}\right)^a},\ k^{-a} \right\}. \tag{35}$$

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 35:

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

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

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

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

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

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

$$\alpha^*(H, B) \;\lesssim\; \begin{cases} B\, d^{-\frac{a^2}{2a-1}}, & B \lesssim 1, \\ B^{\frac{a}{2a-1}}\, d^{-\frac{a^2}{2a-1}}, & 1 \lesssim B \lesssim d^{1/2}, \\ d^{-a/2}, & B \gtrsim d^{1/2}. \end{cases} \tag{36}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract reflects the content.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We do not have large empirical studies, and the ASGD proof is not spectrum independent.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proofs are in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: None.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are averaged over 50 seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: Experiments are small scale synthetic plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: None.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: None.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: None.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.