MOMENTUM IN LARGE-BATCH TRAINING: POLYAK ENLARGES THE CRITICAL BATCH SIZE, NESTEROV IMPROVES DATA EFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study when and how momentum improves large-batch training in the one-pass regime, using power-law kernel regression as a tractable setting. We first characterize risk stability through the *critical learning rate*, defined as the largest learning rate for stable training, and obtain $\eta_{\mathrm{sgd}}^{\mathrm{crit}} \eqsim 1$, $\eta_{\mathrm{polyak}}^{\mathrm{crit}} \eqsim \min\{1, B(1-\beta)\}$, and $\eta_{\mathrm{nesterov}}^{\mathrm{crit}} \eqsim \min\{1, B^q(1-\beta)\}$, where B is the batch size, β is the momentum factor, and $q > 1$ is the capacity exponent. Within this admissible region, we derive scaling laws for the full risk dynamics, capturing the progression from an early transient, through power-law decay, to a noise floor. We then minimize the final-step risk over the admissible learning rates and momentum factors under a fixed data budget, yielding a three-regime batch-size phase diagram that reveals how the role of momentum changes with batch size. Notably, Polyak enlarges the critical batch size, the largest batch size preserving the best small-batch data-scaling exponent, thereby enabling greater parallelism without sacrificing data efficiency. In contrast, Nesterov achieves better data efficiency in the large-batch regime because its look-ahead mechanism suppresses noise accumulation. Numerical experiments validate the predicted stability boundaries, risk dynamics, and batch-size phase diagram.

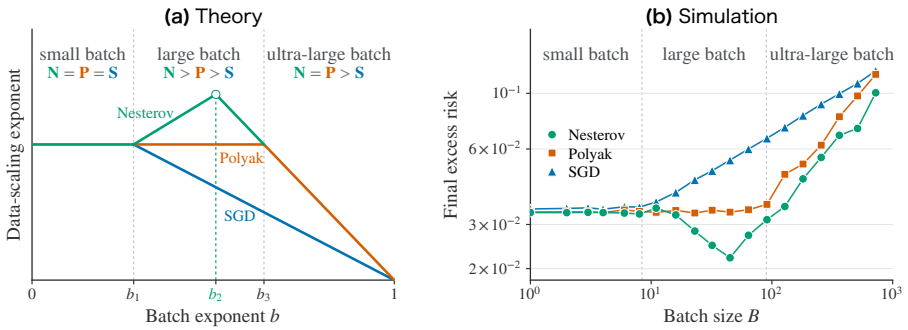


Figure 1: **Batch-size dependence of momentum benefits:** **Polyak** enlarges the critical batch size, while **Nesterov** improves data efficiency. **(a)** Theoretical batch-size phase diagram from Theorem 6.3, showing the optimal data-scaling exponent as a function of the batch exponent b, defined by the scaling $B = D^{b+o(1)}$. **(b)** Best-tuned final excess risk versus batch size B at fixed data budget $D = 2^{14}$ and $(s, q) = (0.3, 4)$. The simulation recovers the same qualitative three-regime structure predicted by the theory. Experimental details are provided in Section 7.

1 INTRODUCTION

Momentum has become a central mechanism in optimizer design for modern large-scale training, underlying popular optimizers such as Adam (Kingma & Ba, 2015) and Muon (Jordan et al., 2024; Liu et al., 2025). Classical theory views momentum primarily as an acceleration mechanism, quantified through reduced iteration complexity (Nesterov, 1983; Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2021; Gitman et al., 2019; Velikanov et al., 2023; Wang et al., 2024; Liu & Belkin, 2018; Gupta et al., 2024). Yet iteration complexity alone does not fully capture the role of

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

momentum in modern large-scale training, where the optimization regime itself changes as training resources, such as data and hardware memory, scale up.

Modern large-scale training, particularly language-model pretraining, often operates in a *one-pass training* regime (Brown et al., 2020). Under a fixed data budget D, the batch size B determines the number of sequential updates, $N = D/B$, directly trading greater hardware parallelism for fewer optimization steps. Recent language-model studies further show that the relative advantages of different optimizers can change substantially with batch size, with some exhibiting markedly stronger data efficiency in the large-batch regime (Marek et al., 2026; Shah et al., 2025; Semenov et al., 2025). This motivates a systematic characterization of momentum across batch sizes:

Question: *Can momentum enable larger batches without sacrificing data efficiency, or even improve data efficiency itself?*

We address this question for both Polyak (Polyak, 1964) and Nesterov (Nesterov, 1983) methods, using power-law kernel (PLK) regression as a tractable learning problem. Its infinite-dimensional power-law spectrum spans a broad range of learning timescales, making the effects of momentum across spectral directions analytically accessible. The problem is characterized by two exponents: the capacity exponent $q > 1$, which controls the decay of the feature spectrum, and the source exponent $s > 0$, which controls how the target-function energy is distributed across the spectrum. Related power-law settings have been widely used to understand scaling laws in large-scale training across a variety of settings (Lin et al., 2024; Bahri et al., 2024; Bordelon et al., 2025; Li et al., 2025; Wang et al., 2026; Li et al., 2026). Building on the functional scaling law (FSL) framework of Li et al. (2025), we extend the analysis to momentum methods.

Our analysis of one-pass training follows the hierarchy

$$\mathcal{A}_B \longrightarrow \mathcal{E}(k,\eta,\beta,B) \longrightarrow \min_{(\eta,\beta)\in\mathcal{A}_B} \mathcal{E}(N,\eta,\beta,B)\,, \quad N = \frac{D}{B}.$$

Here, $\mathcal{A}_B$ denotes the admissible set of learning rates η and momentum β at batch size B, and $\mathcal{E}(k,\eta,\beta,B)$ denotes the expected excess risk. The three stages correspond, respectively, to **1)** identifying the admissible hyperparameter region through stability analysis, **2)** deriving scaling laws that characterize the risk dynamics, and **3)** optimizing the final-step risk subject to both the admissibility constraint $(\eta,\beta) \in \mathcal{A}_B$ and the one-pass constraint $N = D/B$.

Our contributions. Our main contributions are summarized as follows.

- **Risk stability.** We characterize risk stability through the *critical learning rate* η^{crit}, the largest learning rate for which the stochastic dynamics remain stable (Section 4). In the large-batch, high-momentum regime, we obtain the sharp scalings

$$\eta^{\mathrm{crit}}_{\mathrm{sgd}} \eqsim 1, \quad \eta^{\mathrm{crit}}_{\mathrm{polyak}} \eqsim \min\{1, B(1-\beta)\}, \quad \eta^{\mathrm{crit}}_{\mathrm{nesterov}} \eqsim \min\{1, B^q(1-\beta)\}. \tag{1}$$

  Thus, stronger momentum requires a larger batch to maintain stability. This compensation scales as B for Polyak but as B^q for Nesterov. Since $q > 1$, Nesterov permits substantially higher momentum in large-batch training.

- **Scaling laws for risk dynamics.** Within the stable region, we extend the functional scaling law framework (Li et al., 2025) to Polyak and Nesterov and obtain accurate descriptions of their risk dynamics (Section 5). In the high-momentum regime, both methods first exhibit an exponential transient on the timescale $O((1-\beta)^{-1})$, after which they enter the scaling-law regime. Writing $\eta_\beta := \eta/(1-\beta)$, Polyak follows the same signal–noise tradeoff as SGD, whose risk scales as $\mathcal{E}_{\mathrm{sgd}}(k) \eqsim (k\eta)^{-s} + \sigma^2\eta/B$, but with the enlarged effective learning rate η_β:

$$\mathcal{E}_{\mathrm{polyak}}(k) \eqsim (k\eta_\beta)^{-s} + \frac{\sigma^2\eta_\beta}{B}.$$

  Nesterov retains the same accelerated signal term while further suppressing noise accumulation:

$$\mathcal{E}_{\mathrm{nesterov}}(k) \eqsim (k\eta_\beta)^{-s} + \frac{\sigma^2}{B}\min\{\eta_\beta, \eta_\beta^{1/q}\}.$$

- **Data scaling and batch-size phase diagram.** Let $B = D^{b+o(1)}$. Optimizing the remaining hyperparameters yields three characteristic batch exponents $b_1 < b_2 < b_3$ and a three-regime phase diagram (Section 6). For $b \leqslant b_1$, SGD, Polyak, and Nesterov achieve the same data-scaling exponent. For $b_1 < b < b_3$, Nesterov outperforms Polyak, which in turn outperforms SGD, with Nesterov attaining its best data-scaling rate at $b = b_2$. For $b \geqslant b_3$, Polyak and Nesterov again share the same exponent and both outperform SGD. Hence, b_1 and b_3 mark the critical batch sizes of SGD and Polyak, respectively. Polyak retains the best small-batch scaling over a much wider range of batch sizes, while Nesterov further improves the data-scaling rate in the large-batch regime.

Figure 1 summarizes how the relative advantage of the three methods changes with batch size, and Section 7 empirically validates these theoretical predictions.

Remark 1.1 (Why momentum helps large-batch training). The batch-size dependence of momentum can be understood through the *per-sample effective learning rate* (ELR),

$$\eta_{\text{sample}} := \frac{\eta_{\text{eff}}}{B}, \qquad \eta_{\text{eff}} = \begin{cases} \eta, & \text{SGD}, \\ \eta_\beta := \eta/(1-\beta), & \text{momentum}. \end{cases}$$

Ignoring the early transient, SGD and Polyak obey the same data-level signal–noise tradeoff,

$$\mathcal{E}_D(\eta) \eqsim (D\eta_{\text{sample}})^{-s} + \sigma^2 \eta_{\text{sample}}. \tag{2}$$

The difference is the admissible range of the per-sample ELR. Stability condition (1) gives $\eta_{\text{sample}} \lesssim 1/B$ for SGD, whereas Polyak allows $\eta_{\text{sample}} \lesssim 1$. Thus, as the batch size grows, the admissible per-sample ELR of SGD shrinks as $1/B$, while Polyak can keep it at order one by increasing η_β proportionally with B. Polyak therefore extends the SGD signal–noise tradeoff to substantially larger batch sizes without changing its optimum, explaining both the enlarged critical batch size and the unchanged best data-scaling exponent.

In contrast, Nesterov changes the picture more fundamentally. Its look-ahead mechanism suppresses noise accumulation, improving the signal–noise tradeoff itself and thereby achieving better data efficiency in the large-batch regime.

2 Related work

Theory of momentum. Classical analyses explain momentum through accelerated convergence, Lyapunov functions, and continuous-time analysis (Nesterov, 1983; Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2021; Shi et al., 2022; Kovachki & Stuart, 2021). Under stochastic training, subsequent work has studied convergence, stability, and asymptotic behavior (Vidyasagar, 2025; Liu et al., 2020; Yuan et al., 2016; Gitman et al., 2019; Velikanov et al., 2023; Zhang et al., 2024). However, existing theory does not characterize how the role of momentum changes with batch size under a fixed data budget. We instead treat batch size as a scaling variable and compare Polyak and Nesterov across the resulting training regimes.

Scaling-law theory. Scaling laws characterize how learning performance scales with resources such as data, model, and compute (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Grattafiori et al., 2024; DeepSeek-AI et al., 2026). A growing theoretical literature has derived such laws in linear, kernel, and related tractable models (Sharma & Kaplan, 2022; Hutter, 2021; Maloney et al., 2022; Wei et al., 2022; Jain et al., 2024; Michaud et al., 2023; Nam et al., 2024; Atanasov et al., 2026; Bahri et al., 2024; Dohmatob et al., 2024; Bordelon et al., 2024; Lin et al., 2024; Paquette et al., 2024; Bordelon et al., 2025; Yan et al., 2026; Lin et al., 2025; Sous & Winer, 2026; Li et al., 2025; Zhang et al., 2025; Wang et al., 2026; Li et al., 2026; Ding et al., 2025). Functional scaling laws further characterize risk dynamics through signal learning and noise accumulation (Li et al., 2025). We extend this framework to momentum methods. Closely related, Ferbach et al. (2026) study momentum scaling laws in noiseless power-law random features at fixed batch size, showing that Polyak preserves the SGD exponents while adapted Nesterov-type methods can improve them. We instead study the noisy setting with batch size scaling under a fixed data budget, deriving a phase diagram for canonical Polyak and Nesterov momentum.

Stability of gradient-based methods. Stability characterizes the admissible hyperparameter region and has been studied extensively for gradient-based methods, with connections to implicit bias, generalization, and optimization geometry (Zhu et al., 2019; Wu et al., 2018; Nar & Sastry, 2018;

Ma & Ying, 2021; Wu et al., 2022; Wu & Su, 2023; Chemnitz & Engel, 2025; Mulayoff et al., 2021; Nacson et al., 2023; Qiao et al., 2024; Kaur et al., 2023; Wang & Wu, 2023). For stochastic momentum methods, the joint dependence of stability on batch size and momentum remains less understood. We derive sharp risk stability boundaries for Polyak and Nesterov.

Concurrent work. Morwani et al. (2026), which appeared while this manuscript was in preparation, studies stochastic momentum in finite-dimensional linear regression and derives lower bounds on contraction rates, suggesting that Polyak may remain data-efficient over a wider range of batch sizes; matching upper bounds remain open. In contrast, we derive sharp scaling laws for the full risk dynamics and, after optimizing the learning rate and momentum factor, obtain the data-scaling exponent across batch sizes. This yields a complete phase diagram, including a large-batch regime where Nesterov outperforms Polyak.

3 PRELIMINARIES

Notation. We use $\eqsim$ to denote equivalence up to a constant factor, and $\lesssim$ (resp. $\gtrsim$) indicates inequality up to a constant factor. For two non-negative functions f, g, we write $f(x) \lesssim g(x)$ (resp. $f(x) \gtrsim g(x)$) if there exists a constant $C > 0$ such that $f(x) \leqslant Cg(x)$ (resp. $g(x) \leqslant Cf(x)$) for all x under consideration. We write $f(x) \eqsim g(x)$ if both $f(x) \lesssim g(x)$ and $f(x) \gtrsim g(x)$ hold. In this paper, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2$ denote the standard inner product and the associated 2-norm, respectively.

Power-law kernel (PLK) regression. Consider regression on a Hilbert space $\mathcal{H}$ with the input $\mathbf{x} \sim \mathcal{D}$ and the observed label $y = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta}^* \rangle + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{x}$. Assume that the feature function $\boldsymbol{\phi}(\mathbf{x})$ is Gaussian: $\boldsymbol{\phi}(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{H})$, where $\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\boldsymbol{\phi}(\mathbf{x}) \otimes \boldsymbol{\phi}(\mathbf{x})]$ denotes the feature covariance operator. Let $\{\lambda_j\}_{j\geqslant 1}$ denote the eigenvalues of $\mathbf{H}$, arranged in decreasing order. We learn the target function using the linear student model $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta} \rangle$ by minimizing the population risk: $\mathcal{R}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\mathbf{x},y}\left[(f(\mathbf{x}; \boldsymbol{\theta}) - y)^2\right]$. A direct calculation gives $\mathcal{R}(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{\theta}) + \frac{\sigma^2}{2}$, where $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2}\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \rangle$ is the excess risk. Since SGD, Polyak, and Nesterov momentum are orthogonally equivariant, we make the following diagonal assumption without loss of generality; see Appendix A.1.

Assumption 3.1 (Diagonal covariance)**.** $\mathbf{H} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots > 0$.

Assumption 3.2 (Constant-order label noise)**.** Assume $\sigma \eqsim 1$.

Assumption 3.3 (Source and capacity conditions)**.** There exist $q > 1$ and $s > 0$ such that

$$\lambda_j \eqsim j^{-q}, \quad \lambda_j |\theta_j^*|^2 \eqsim j^{-(1+sq)}.$$

Here q is the **capacity exponent**, which controls the decay rate of the feature spectrum, and s is the **source exponent**, which describes how the target energy decays across spectral directions. We also refer to s as the **relative difficulty** of the task: smaller s means that more target energy lies in low-curvature, small-eigenvalue directions, making the problem harder to learn. We remark that similar assumptions have been widely used in the analysis of kernel methods (Caponnetto & De Vito, 2005; 2007; Spigler et al., 2020; Bordelon et al., 2020; Maloney et al., 2022).

Optimization methods. At each step, we draw a mini-batch $\mathcal{B}_k = \{(\mathbf{x}_{k,i}, y_{k,i})\}_{i=1}^B$ and compute the stochastic gradient $\mathbf{g}(\boldsymbol{\theta}; \mathcal{B}_k) = \nabla_{\boldsymbol{\theta}}(\frac{1}{2B}\sum_{(\mathbf{x},y)\in\mathcal{B}_k}(f(\mathbf{x}; \boldsymbol{\theta}) - y)^2)$. SGD and Polyak update as follows:

$$\text{SGD:} \qquad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}(\boldsymbol{\theta}_k; \mathcal{B}_k), \tag{3}$$

$$\text{Polyak:} \qquad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}(\boldsymbol{\theta}_k; \mathcal{B}_k) + \beta(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}), \tag{4}$$

where η is the learning rate and $\beta \in [0, 1)$ is the momentum factor. Nesterov momentum differs from Polyak momentum by evaluating the gradient at a look-ahead point:

$$\text{Nesterov:} \qquad \boldsymbol{\theta}_k^{\mathrm{la}} = \boldsymbol{\theta}_k + \beta(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}) \qquad \text{(look ahead)}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \mathbf{g}(\boldsymbol{\theta}_k^{\mathrm{la}}; \mathcal{B}_k) + \beta(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}), \tag{5}$$

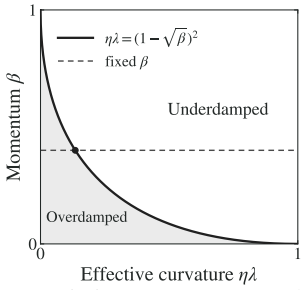
where $\boldsymbol{\theta}_k^{\mathrm{la}}$ is the look-ahead point obtained by extrapolating the current iterate along the previous update direction. This look-ahead mechanism distinguishes Nesterov from Polyak.

For clarity, we decompose $\mathbf{g}(\boldsymbol{\theta}; \mathcal{B}_k) = \nabla\mathcal{R}(\boldsymbol{\theta}) + \boldsymbol{\xi}_k(\boldsymbol{\theta})$, where $\nabla\mathcal{R}(\boldsymbol{\theta})$ denotes the population gradient and $\boldsymbol{\xi}_k(\boldsymbol{\theta})$ represents the stochastic gradient noise satisfying $\mathbb{E}[\boldsymbol{\xi}_k(\boldsymbol{\theta})] = \mathbf{0}$, and $\mathbb{E}[\boldsymbol{\xi}_k(\boldsymbol{\theta}) \otimes \boldsymbol{\xi}_k(\boldsymbol{\theta})] = \frac{1}{B}\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the gradient-noise covariance of batch size 1.

Regime-dependent damping in momentum dynamics. Consider full-batch Polyak for minimizing $f(x) = \lambda x^2/2$, where $\lambda > 0$ denotes the curvature. Since $\nabla f(x) = \lambda x$, the Polyak update is:

$$x_{k+1} = x_k - \eta\lambda x_k + \beta(x_k - x_{k-1}). \tag{6}$$

The corresponding characteristic polynomial is $r^2 - (1+\beta-\eta\lambda)r + \beta = 0$. The relevant transition from real to complex roots occurs at $\eta\lambda = (1-\sqrt{\beta})^2$. Hence, different curvature directions can exhibit two qualitatively different damping behaviors.



- **Overdamped regime**: $\lambda\eta < (1-\sqrt{\beta})^2$. In this regime, the characteristic roots $r_\pm$ are real with the dominant root satisfying $r_+ \approx 1 - \frac{\eta\lambda}{1-\beta}$ when $1-\beta = o(1)$. For sufficiently large k, it holds that $x_k = c_+ r_+^k + c_- r_-^k \approx c_+ r_+^k \approx c_+ \exp(-\frac{\eta\lambda}{1-\beta}k)$. Hence, the resulting dynamics behave like gradient descent with the *momentum-rescaled learning rate*

$$\eta_\beta = \eta/(1-\beta). \tag{7}$$

- **Underdamped regime**: $\lambda\eta > (1-\sqrt{\beta})^2$. In this regime, the characteristic roots form a complex-conjugate pair with modulus $\sqrt{\beta}$. The iterate therefore has the form $x_k = \beta^{k/2}\big[c_1\cos(k\omega) + c_2\sin(k\omega)\big]$, where c_1, c_2, and ω depend on the initialization and curvature. Thus, the iterate exhibits oscillatory convergence to zero, with the amplitude decaying as $\beta^{k/2}$.

The key difference between the two regimes lies in how curvature affects the dynamics. In the overdamped regime, the decay rate $\eta\lambda/(1-\beta)$ is proportional to the curvature, so flatter directions converge more slowly. In the underdamped regime, the oscillation frequency depends on the curvature, but the amplitude envelope $\beta^{k/2}$ does not.

Nesterov look-ahead induces curvature-adaptive damping. On the same quadratic objective, Nesterov updates using the gradient at the look-ahead point $x_k^{\text{la}} = x_k + \beta(x_k - x_{k-1})$:

$$\begin{aligned} x_{k+1} &= x_k - \eta\lambda[x_k + \beta(x_k - x_{k-1})] + \beta(x_k - x_{k-1}) \\ &= x_k - \eta\lambda x_k + \beta(1-\eta\lambda)(x_k - x_{k-1}), \end{aligned} \tag{8}$$

which is equivalent to Polyak with the momentum factor $\beta(1-\eta\lambda)$. For an underdamped direction, it accelerates convergence: $\beta^{k/2} \to [\beta(1-\eta\lambda)]^{k/2}$, which damps sharper directions more rapidly. By contrast, in sufficiently flat directions where $\eta\lambda \ll 1$, we have $\beta(1-\eta\lambda) \approx \beta$, and Nesterov retains approximately the same dynamics as Polyak.

4 Risk Stability and Critical Learning Rates

We first characterize the maximal learning rates that ensure stable training and study their dependence on the momentum factor and batch size.

Definition 4.1 (Risk stability)**.** Consider a stochastic optimization algorithm for minimizing a population risk $\mathcal{R}(\cdot)$, producing model parameters $\{\boldsymbol{\theta}_k\}_{k\geqslant 0}$. It is said to be *risk stable* if, for every admissible initialization with finite population risk, $\sup_{k\geqslant 0}\mathbb{E}[\mathcal{R}(\boldsymbol{\theta}_k)] < \infty$.

Definition 4.2 (Critical learning rate)**.** Consider a stochastic optimization algorithm $\mathcal{A}$ parameterized by a learning rate η and additional hyperparameters $\mathbf{h}$. For fixed $\mathbf{h}$, let

$$\eta_{\mathcal{A}}^{\text{crit}}(\mathbf{h}) := \sup\{\eta > 0 : \mathcal{A} \text{ is risk stable}\}.$$

Critical learning rate is the largest learning rate compatible with risk stability. When the dependence on $\mathbf{h}$ is clear from context, we simply write $\eta_{\mathcal{A}}^{\text{crit}}$. In our setting, $\mathbf{h} = B$ for SGD and $\mathbf{h} = (B, \beta)$ for Polyak and Nesterov. The following theorem characterizes how the critical learning rate scales with batch size and momentum; the proof is deferred to Appendix B.

Theorem 4.3 (Critical learning-rate scaling)**.** For sufficiently large B and sufficiently small $1-\beta$,

$$\eta_{\text{sgd}}^{\text{crit}} \eqsim 1, \tag{9}$$

$$\eta_{\text{polyak}}^{\text{crit}} \eqsim \min\{1,\, B(1-\beta)\}, \tag{10}$$

$$\eta_{\text{nesterov}}^{\text{crit}} \eqsim \min\{1,\, B^q(1-\beta)\}, \tag{11}$$

where the implicit constants are independent of B and β.

Joint dependence on momentum and batch size. Theorem 4.3 reveals a direct coupling among the learning rate, momentum factor, and batch size in determining training stability. Increasing β toward one reduces the critical learning rate, whereas increasing the batch size enlarges it. In the high-momentum regime, $\eta^{\mathrm{crit}}_{\mathrm{polyak}} \eqsim B(1-\beta), \eta^{\mathrm{crit}}_{\mathrm{nesterov}} \eqsim B^q(1-\beta)$, before reaching their order-one ceilings. Thus, the batch-size effect on stability is substantially stronger for Nesterov, with a B^q dependence rather than the B dependence of Polyak. This difference originates from the curvature adaptivity induced by Nesterov's look-ahead mechanism.

Multiplicative-noise feedback governs stability through the renewal kernel. For all three methods, Appendix B shows that the risk dynamics admit, up to constant factors, a renewal equation $\ell_k \eqsim h_k + \sum_{j=0}^{k-1} K_{k-1-j}\big(\ell_j + \sigma^2\big)$, where ℓ_k is the risk moment, h_k is the free response, and the renewal kernel K describes how stochastic perturbations propagate through the optimizer dynamics. Additive and multiplicative noise play different roles: additive label noise acts only as an external forcing, whereas multiplicative noise feeds the current risk back through the renewal kernel, producing the feedback term $K * \ell$. Risk stability is therefore determined by the strength of this feedback, characterized by $\|K\|_1 < 1$; see Appendix B.1. Since K is independent of the label-noise scale σ, the stability condition in Theorem 4.3 is also independent of σ.

5 Scaling laws for risk dynamics across damping regimes

In this section, we establish scaling laws for the full risk dynamics of Polyak and Nesterov. For reference, the **SGD scaling law** for PLK regression is

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim (\eta k)^{-s} + \frac{\sigma^2}{B}\eta, \qquad k\eta \gtrsim 1. \tag{12}$$

This follows from Li et al. (2025, Theorem 5.2); see Appendix C.4 for details. The first term, $(\eta k)^{-s}$, describes signal learning, with the decay rate determined by the source exponent s: larger s corresponds to an easier target and faster learning. The second term, $\sigma^2\eta/B$, captures stochastic noise accumulation and determines the asymptotic noise floor.

For momentum methods, increasing the momentum parameter β toward 1 leads to three regimes: fully overdamped dynamics, mixed damping, and eventual risk instability. In the mixed regime, sharp modes become underdamped while flatter modes remain overdamped.

Theorem 5.1 (Scaling law for Polyak)**.** Assume $\eta \lesssim 1$ and $k \gtrsim \max(\eta_\beta^{-1}, (1-\beta)^{-1})$. Then

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \begin{cases} (\eta_\beta k)^{-s} + \dfrac{\sigma^2}{B}\eta_\beta, & 1-\beta \gtrsim \sqrt{\eta}, \\ \mathbb{T}_P(k) + (\eta_\beta k)^{-s} + \dfrac{\sigma^2}{B}\eta_\beta, & \dfrac{\eta}{B} \lesssim 1-\beta \lesssim \sqrt{\eta}, \end{cases} \tag{13}$$

where $\mathbb{T}_P$ denotes the underdamped transient and satisfies $0 \leqslant \mathbb{T}_P(k) \lesssim \beta^k$. If $1-\beta \lesssim \eta/B$, the dynamics are not risk stable.

The proof is deferred to Appendix C.2. In the fully overdamped regime, Polyak follows the SGD scaling law with effective learning rate η_β. In the mixed-damping regime, the transient $\mathbb{T}_P(k)$ arises from underdamped sharp modes, whereas the power-law term $(\eta_\beta k)^{-s}$ accounts for overdamped flat modes. Balancing the two terms gives the crossover timescale

$$\tau_\beta \eqsim \frac{1}{1-\beta}\log\frac{\eta}{(1-\beta)^2}, \tag{14}$$

after which the underdamped transient becomes negligible and the usual scaling-law dynamics dominate. Thus, **increasing β accelerates learning in the flat modes through a larger η_β, but also prolongs the relaxation of the sharp modes**. Importantly, Polyak preserves the SGD signal–noise tradeoff while allowing $\eta_\beta \lesssim B$, according to Theorem 4.3, and hence enlarges the critical batch size without improving the optimal data-scaling exponent.

For Nesterov, the same damping structure persists, but its look-ahead mechanism modifies both the underdamped transient and noise accumulation in the mixed-damping regime.

Theorem 5.2 (Scaling law for Nesterov)**.** Assume $\eta \lesssim 1$ and $k \gtrsim \max(\eta_\beta^{-1}, (1-\beta)^{-1})$. Then

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \begin{cases} (\eta_\beta k)^{-s} + \dfrac{\sigma^2}{B}\eta_\beta, & 1-\beta \gtrsim \sqrt{\eta}, \\ \mathbb{T}_N(k) + (\eta_\beta k)^{-s} + \dfrac{\sigma^2}{B}\min\{\eta_\beta, \eta_\beta^{1/q}\}, & \dfrac{\eta}{B^q} \lesssim 1-\beta \lesssim \sqrt{\eta}, \end{cases} \tag{15}$$

where $\mathbb{T}_N$ denotes the underdamped transient and satisfies $0 \leqslant \mathbb{T}_N(k) \lesssim \min\{1, (\eta k)^{-s}\}\beta^k$. If $1-\beta \lesssim \eta/B^q$, the dynamics are not risk stable.

The proof is deferred to Appendix C.3. The distinction between Nesterov and Polyak emerges in the mixed-damping regime and stems from the curvature-adaptive damping induced by Nesterov's look-ahead mechanism; see Eq. (8). This accelerates the decay of the underdamped transient from $\mathbb{T}_P(k) \lesssim \beta^k$ to $\mathbb{T}_N(k) \lesssim \min\{1, (\eta k)^{-s}\}\beta^k$, while also suppressing stochastic noise accumulation. In particular, when $\eta_\beta \gg 1$, the noise floors scale as

$$\frac{\sigma^2}{B}\eta_\beta \qquad \text{for Polyak,} \qquad \frac{\sigma^2}{B}\eta_\beta^{1/q} \qquad \text{for Nesterov.}$$

For asymptotic data scaling, this reduced noise accumulation is the key advantage of Nesterov.

6 Optimal data scaling across batch-size regimes

We now optimize the final-step risk using the scaling laws derived above, subject to the one-pass constraint $N = D/B$ and the stability constraints. We focus on the large-batch setting, allowing B to grow with the data budget D rather than remain fixed.

Definition 6.1 (Batch-size exponent)**.** For a sequence of training problems with $D \to \infty$, the batch size has exponent $b \in [0, 1)$ if $B = D^{b+o(1)}$, or equivalently, $b = \lim_{D\to\infty} \frac{\log B}{\log D}$.

Definition 6.2 (Batch-size-dependent data-scaling exponent)**.** Fix $b \in [0, 1)$, so that $B = D^{b+o(1)}$ and $N = D/B$. For an optimization method with admissible hyperparameter region $\mathcal{A}_B$, define its optimal data-scaling exponent $r(b)$ by

$$\inf_{(\eta,\beta)\in\mathcal{A}_B} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] = D^{-r(b)+o(1)}.$$

Here, β is omitted for SGD. We denote the exponents for SGD, Polyak, and Nesterov by $r_{\text{sgd}}(b)$, $r_{\text{polyak}}(b)$, and $r_{\text{nesterov}}(b)$, respectively. A larger $r(b)$ indicates better data efficiency.

Define the three transition exponents

$$b_1 = \frac{s}{s+1}, \qquad b_2 = \frac{2s+1/q}{2s+1+1/q}, \qquad b_3 = \frac{2s+1}{2s+2}. \tag{16}$$

Since $q > 1$ and $s > 0$, it holds $0 < b_1 < b_2 < b_3 < 1$. The following theorem gives the optimal data-scaling exponents across batch-size regimes; the proof is deferred to Appendix D.

Theorem 6.3 (Batch-size-dependent data scaling)**.** *For $B = D^{b+o(1)}$ with $b \in [0, 1)$,*

$$r_{\text{sgd}}(b) = \begin{cases} \frac{s}{s+1}, & 0 \leqslant b < b_1, \\ s(1-b), & b_1 \leqslant b < 1, \end{cases} \qquad r_{\text{polyak}}(b) = \begin{cases} \frac{s}{s+1}, & 0 \leqslant b < b_3, \\ 2s(1-b), & b_3 \leqslant b < 1, \end{cases}$$

and

$$r_{\text{nesterov}}(b) = \begin{cases} \frac{s}{s+1}, & 0 \leqslant b < b_1, \\ \frac{s(1+(q-1)b)}{sq+1}, & b_1 \leqslant b < b_2, \\ 2s(1-b), & b_2 \leqslant b < 1. \end{cases}$$

This theorem shows that the achievable data efficiency depends sharply on batch-size scaling. Thus, Polyak enlarges the batch-size range over which the best small-batch data scaling is preserved, whereas Nesterov further improves the achievable data-scaling exponent in the large-batch regime.

- **Small batch** ($b \leqslant b_1$). All three methods achieve the same optimal data-scaling exponent:
$$r_{\text{nesterov}}(b) = r_{\text{polyak}}(b) = r_{\text{sgd}}(b).$$
- **Large batch** ($b_1 < b < b_3$). The three methods separate:
$$r_{\text{nesterov}}(b) > r_{\text{polyak}}(b) > r_{\text{sgd}}(b),$$
with Nesterov attaining its best data-scaling exponent at $b = b_2$.
- **Ultra-large batch** ($b \geqslant b_3$). Polyak and Nesterov share the same exponent and outperform SGD:
$$r_{\text{nesterov}}(b) = r_{\text{polyak}}(b) > r_{\text{sgd}}(b).$$

Mechanism. Remark 1.1 shows that SGD and Polyak share the same signal–noise tradeoff. Maintaining a fixed per-sample ELR requires the linear-scaling rule for ELR $\eta_{\text{eff}} \propto B$. For SGD, this eventually hits the stability ceiling $\eta \lesssim 1$. Polyak can continue the linear scaling by increasing momentum, so that $\eta_{\text{eff}} = \eta/(1-\beta)$ grows with B even after η saturates. This extends the optimal scaling to much larger batch sizes. As β increases, the underdamped transient becomes progressively longer and decays more slowly. The Polyak critical batch size is reached when this transient spans the entire training horizon, leaving insufficient time to enter the scaling-law regime and resulting in suboptimal scaling.

7 Numerical Experiments

7.1 PLK regression

Risk stability and risk-dynamics scaling laws. We validate the stability predictions of Theorem 4.3 with $(s, q) = (0.3, 1.5)$. Figure 2 (left) fixes $\beta = 1 - 2^{-6}$ and varies B, confirming $\eta^{\text{crit}}_{\text{polyak}} \eqsim B$ and $\eta^{\text{crit}}_{\text{nesterov}} \eqsim B^q$ before saturation, while $\eta^{\text{crit}}_{\text{sgd}}$ remains order one. Figure 2 (middle) fixes $B = 4$ and varies β, confirming the predicted linear dependence of both momentum stability boundaries on $1 - \beta$. Appendix E.1.1 provides further experimental details and additional results.

Figure 2 (right) shows that the momentum scaling laws accurately capture the risk dynamics of both Polyak and Nesterov for $(s, q) = (0.3, 4)$. In particular, it confirms three key predictions: an underdamped transient with transition timescale dominated by $(1-\beta)^{-1}$, a common power-law signal decay, and distinct asymptotic noise floors. Once the transient becomes negligible, Polyak and Nesterov follow the same power law, while Nesterov reaches a substantially lower noise floor. Appendix E.1.2 provides further experimental details and additional results.

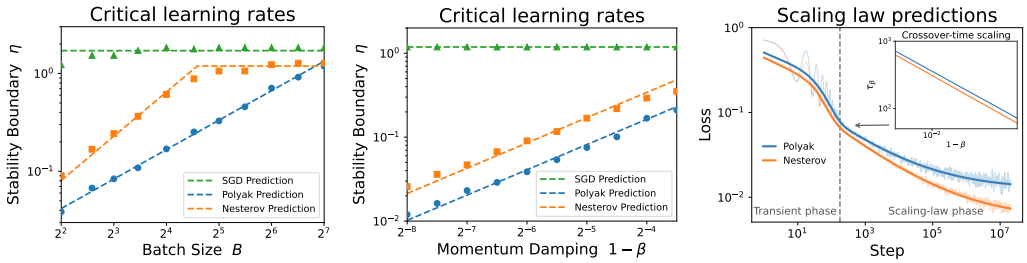


Figure 2: **Validation of stability and risk-dynamics scaling laws. Left:** Critical learning rate versus batch size at fixed momentum; markers denote numerical estimates and dashed lines denote theoretical predictions. **Middle:** Critical learning rate versus $1 - \beta$ at fixed batch size, confirming the predicted linear scaling before the order-one ceiling. **Right:** Polyak and Nesterov risk dynamics; thin fluctuating lines show empirical risk curves, while thick lines show theoretical predictions. The inset confirms the dominant $(1-\beta)^{-1}$ scaling of the crossover timescale τ_β (see Eq. 14).

Batch-size phase diagram. We first test the three-regime structure predicted by Theorem 6.3 at a finite data budget. We fix $D = 2^{14}$ and, for each batch size, tune (η, β) to minimize the final excess risk. Figure 1 (right) exhibits a progression across the three regimes. At small batch sizes, SGD, Polyak, and Nesterov achieve nearly identical risk. As B increases, SGD deteriorates first, while Polyak maintains the small-batch performance over a wider range, reflecting its enlarged critical

batch size. Nesterov goes further: in the large-batch regime, its risk continues to decrease beyond the best level attained by SGD and Polyak, before eventually increasing at ultra-large batch sizes. Appendix E.1.3 provides the experimental setup and additional results.

We next test the predicted data-scaling exponents at representative points in each regime. We continue with $(s, q) = (0.3, 4)$ and choose $B \eqsim 1$, $B \eqsim D^{0.4}$, and $B \eqsim D^{0.8}$, representing small batch, large batch, and ultra-large batch, respectively. For each data budget D, we set (η, β) according to the theoretically prescribed scalings. Figure 3 shows that the empirical slopes closely match Theorem 6.3: all three methods share the same exponent for small batch; for large batch, Nesterov outperforms Polyak, which outperforms SGD; and for ultra-large batch, Polyak and Nesterov again share the same exponent and both outperform SGD. These results quantitatively validate the batch-dependent data-scaling exponents underlying the phase diagram. Experimental details and additional settings are provided in Appendix E.1.4.

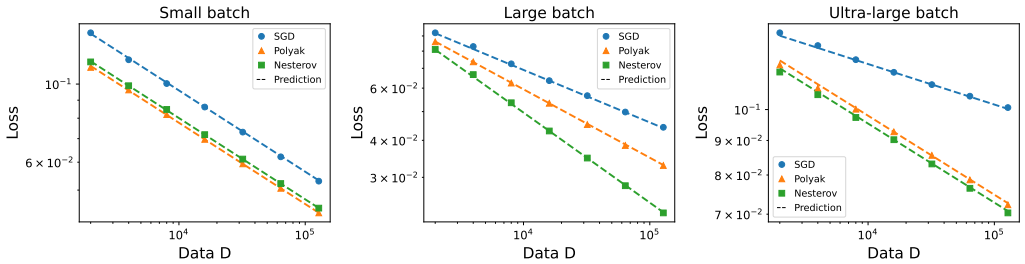
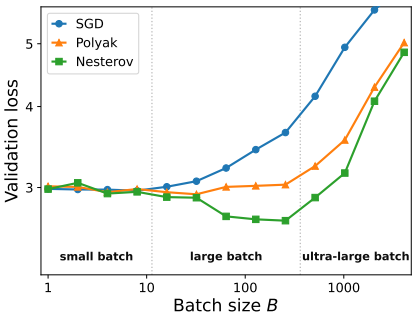


Figure 3: **Data-scaling exponents across the three batch-size regimes.** We choose representative batch-size scalings $B \eqsim 1$ **(left)**, $B \eqsim D^{0.4}$ **(middle)**, and $B \eqsim D^{0.8}$ **(right)**, and vary the data budget D. Markers show empirical final excess risks, while dashed lines indicate the scaling exponents predicted by Theorem 6.3. For these representative choices, the predicted ordering is recovered: Nesterov = Polyak = SGD at small batch, Nesterov > Polyak > SGD at large batch, and Nesterov = Polyak > SGD at ultra-large batch.

7.2 IMAGENET-1K

We examine the batch-size dependence of the three optimizers when training ResNet-18 on ImageNet-1K. For each batch size, we tune the learning rate and, for Polyak and Nesterov, the momentum factor over grids, and report the best achievable risk. The results, shown in the right figure, display the same qualitative batch-size dependence predicted by the PLK analysis. Polyak achieves a similar best risk as SGD, but maintains this performance over a substantially larger batch-size range, allowing fewer sequential updates without sacrificing data efficiency. Nesterov exhibits a further advantage in the large-batch regime, where it attains lower risk than both SGD and Polyak. These results suggest that the phase structure identified in PLK regression extends qualitatively to a standard deep-learning benchmark. Additional experimental details are provided in Appendix E.2.1.



8 CONCLUSION

We developed a unified analysis of Polyak and Nesterov momentum in large-batch, one-pass training, covering risk stability, risk dynamics, and optimal data scaling. The main conclusion is that the benefit of momentum depends critically on batch size. At small batches, neither method improves over SGD. At larger batches, Polyak preserves the optimal data-scaling rate over a much wider batch-size range, while Nesterov further improves data efficiency. More broadly, our results show that learning rate, batch size, and momentum should be considered jointly rather than tuned in isolation. These hyperparameters together determine the admissible stability region, which in turn constrains the effective optimization dynamics and achievable data efficiency. Extending this joint perspective beyond the present setting may provide a useful framework for understanding and designing large-scale training recipes.

AI USE STATEMENT

We used AI tools for drafting and language editing throughout the manuscript. In Appendices B–D, they were also used to accelerate algebraic calculations and asymptotic estimates. The proof strategies and mathematical arguments were developed by the authors, who verified all AI-assisted calculations and estimates. We take responsibility for the accuracy, originality, and integrity of the final manuscript, including all claims, derivations, and conclusions.

REPRODUCIBILITY STATEMENT

We believe that the experimental results in this work are reproducible. Appendix E provides the experimental setups, training and evaluation procedures, hyperparameter settings, and results aggregated across independent runs. The synthetic data are generated from the specified PLK model, and CIFAR-10 and ImageNet-1K are publicly available datasets.

REFERENCES

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026. (cited on page 3)

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), 2024. doi: 10.1073/pnas.2311878121. (cited on pages 2 and 3)

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020. (cited on page 4)

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382, 2024. (cited on page 3)

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. In *International Conference on Learning Representations*, 2025. (cited on pages 2 and 3)

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. (cited on page 2)

Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm. 2005. (cited on page 4)

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007. (cited on page 4)

Dennis Chemnitz and Maximilian Engel. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning. *Journal of Machine Learning Research*, 26(134):1–46, 2025. (cited on page 4)

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. (cited on page 3)

Shihong Ding, Haihan Zhang, Hanzhen Zhao, and Cong Fang. Scaling law for stochastic gradient descent in quadratically parameterized linear regression. *arXiv preprint arXiv:2502.09106*, 2025. (cited on page 3)

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning*, 2024. (cited on page 3)

William Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 1991. (cited on page 23)

Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales SGD. *Advances in Neural Information Processing Systems*, 38:112780–112977, 2026. (cited on page 3)

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9633–9643, 2019. (cited on pages 1 and 3)

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. (cited on page 3)

Kanan Gupta, Jonathan W. Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20694–20744, 2024. (cited on page 1)

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. (cited on page 3)

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. (cited on page 3)

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021. (cited on page 3)

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. (cited on page 3)

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/. (cited on page 1)

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. (cited on page 3)

Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum Hessian eigenvalue and generalization. In *Proceedings on "I Can't Believe It's Not Better! – Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pp. 51–65, 2023. (cited on page 4)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. (cited on page 1)

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. (cited on page 3)

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *Advances in Neural Information Processing Systems*, 38:101211–101269, 2025. (cited on pages 2, 3, 6, 31, 32, and 47)

Binghui Li, Zilin Wang, Fengling Chen, Shiyang Zhao, Ruiheng Zheng, and Lei Wu. Optimal learning-rate schedules under functional scaling laws: Power decay and warmup-stable-decay. *arXiv preprint arXiv:2602.06797*, 2026. (cited on pages 2 and 3)

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. (cited on pages 2 and 3)

Licong Lin, Jingfeng Wu, and Peter L. Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems*, volume 38, 2025. (cited on page 3)

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018. (cited on page 1)

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. (cited on page 1)

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems*, volume 33, 2020. (cited on page 3)

Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. (cited on page 4)

Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022. (cited on pages 3 and 4)

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla SGD works, and why gradient accumulation is wasteful. *Advances in Neural Information Processing Systems*, 38: 148837–148862, 2026. (cited on page 2)

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, 2023. (cited on page 3)

Depen Morwani, Alexandru Meterez, Pranav Nair, and Sham Kakade. Compute efficiency and serial runtime tradeoffs for stochastic momentum methods. *arXiv preprint arXiv:2606.19179*, 2026. (cited on page 4)

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, volume 34, 2021. (cited on page 4)

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *International Conference on Learning Representations*, 2023. (cited on page 4)

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, and Ard Louis. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. In *Advances in Neural Information Processing Systems*, volume 37, 2024. (cited on page 3)

Kamil Nar and S. Shankar Sastry. Step size matters in deep learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3436–3444, 2018. (cited on page 3)

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl Akad Nauk SSSR*, volume 269, pp. 543, 1983. (cited on pages 1, 2, and 3)

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024. (cited on page 3)

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR computational mathematics and mathematical physics*, 4(5):1–17, 1964. (cited on page 2)

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. (cited on page 4)

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025. (cited on page 2)

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025. (cited on page 2)

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022. (cited on page 3)

Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1–2): 79–148, 2022. doi: 10.1007/s10107-021-01681-8. (cited on page 3)

John Sous and Michael Winer. Asymmetric scaling laws from sparse features. *arXiv preprint arXiv:2605.23591*, 2026. (cited on page 3)

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020. (cited on page 4)

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016. (cited on pages 1 and 3)

Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch SGD via generating functions: Conditions of convergence, phase transitions, benefit from negative momenta. In *International Conference on Learning Representations*, 2023. (cited on pages 1 and 3)

Mathukumalli Vidyasagar. Convergence of momentum-based optimization algorithms with time-varying parameters. *arXiv preprint arXiv:2506.11904*, 2025. (cited on page 3)

Jinbo Wang, Binghui Li, Zhanpeng Zhou, Mingze Wang, Jiaqi Zhang, Xunliang Cai, and Lei Wu. Fast catch-up, late switching: Optimal batch size scheduling via functional scaling laws. In *International Conference on Learning Representations*, volume 2026, pp. 7310–7340, 2026. (cited on pages 2 and 3)

Mingze Wang and Lei Wu. A theoretical analysis of noise geometry in stochastic gradient descent. *arXiv preprint arXiv:2310.00692*, 2023. (cited on pages 4 and 18)

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *International Conference on Learning Representations*, 2024. (cited on page 1)

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23549–23588, 2022. (cited on page 3)

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113 (47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113. (cited on pages 1 and 3)

Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021. (cited on pages 1 and 3)

Lei Wu and Weijie J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37656–37684, 2023. (cited on page 4)

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8279–8288, 2018. (cited on page 3)

Lei Wu, Mingze Wang, and Weijie J. Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4680–4693, 2022. (cited on pages 4 and 18)

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. In *International Conference on Learning Representations*, 2026. (cited on page 3)

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. (cited on page 3)

Haihan Zhang, Yuanshi Liu, Qianwen Chen, and Cong Fang. The optimality of (accelerated) SGD for high-dimensional quadratic optimization. *arXiv preprint arXiv:2409.09745*, 2024. (cited on page 3)

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham M. Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations*, 2025. (cited on page 3)

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pp. 7654–7663, 2019. (cited on pages 3 and 18)

Appendix

Table of Contents

A Technical preliminaries **16**
- A.1 Kernel representation and rotational equivariance . . . 16
- A.2 Noise-curvature alignment . . . 17
- A.3 Second-order recursions . . . 19
- A.4 Power-law spectral sums . . . 21
- A.5 A convolution inequality . . . 22

B Risk stability analysis **23**
- B.1 Renewal formulation of risk stability . . . 23
- B.2 Risk stability of SGD . . . 23
- B.3 Risk stability of Polyak . . . 25
- B.4 Risk stability of Nesterov . . . 28

C Functional scaling laws for risk dynamics **31**
- C.1 Functional scaling law for second-order iterations . . . 31
- C.2 Scaling law for Polyak (Proof of Theorem 5.1) . . . 32
- C.3 Scaling law for Nesterov (Proof of Theorem 5.2) . . . 39
- C.4 Scaling law for SGD . . . 47

D Optimal data scaling across batch-size regimes **47**
- D.1 Data scaling for SGD . . . 48
- D.2 Data scaling for Polyak . . . 48
- D.3 Data scaling for Nesterov . . . 49

E Experimental details and additional results **51**
- E.1 Power-law kernel regression . . . 52
- E.2 Neural-network experiments . . . 54

A TECHNICAL PRELIMINARIES

A.1 KERNEL REPRESENTATION AND ROTATIONAL EQUIVARIANCE

We explain the kernel interpretation of the PLK model and justify the spectral representation used throughout the analysis.

Kernel representation. Given a positive semidefinite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, kernel regression learns over its associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$. Given an input distribution $\mathcal{D}$, under mild conditions, Mercer's theorem ensures that the kernel admits the following eigendecomposition

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j e_j(\mathbf{x}) e_j(\mathbf{x}'),$$

where $\{e_j\}_{j \geqslant 1}$ are orthonormal in $L^2(\mathcal{D})$. The standard capacity condition assumes $\lambda_j \eqsim j^{-q}$. Target regularity is commonly described through the interpolation spaces

$$\mathcal{H}_K^r = \left\{ \sum_{j=1}^{\infty} a_j e_j : \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j^r} < \infty \right\},$$

with a source condition of the form $f^* \in \mathcal{H}_K^r$.

Our formulation is the corresponding feature-map representation. Under Assumption 3.1, define $e_j := \phi_j / \lambda_j^{1/2}$. Then $\{e_j\}_{j \geqslant 1}$ is orthonormal in $L^2(\mathcal{D})$, and

$$K_\phi(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \phi_j(\mathbf{x}) \phi_j(\mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j e_j(\mathbf{x}) e_j(\mathbf{x}').$$

Thus, Assumption 3.3, $\lambda_j \eqsim j^{-q}$, is precisely the standard capacity condition.

For the target function,

$$f^* = \sum_{j=1}^{\infty} \theta_j^* \phi_j = \sum_{j=1}^{\infty} a_j^* e_j, \qquad a_j^* := \lambda_j^{1/2} \theta_j^*.$$

Our task-difficulty assumption gives $|a_j^*|^2 \eqsim j^{-1} \lambda_j^s$. Hence, for any $\delta \in (0, s)$,

$$\sum_{j=1}^{\infty} \frac{|a_j^*|^2}{\lambda_j^{s-\delta}} \eqsim \sum_{j=1}^{\infty} j^{-1} \lambda_j^{\delta} \eqsim \sum_{j=1}^{\infty} j^{-1-q\delta} < \infty.$$

Therefore, $f^* \in \mathcal{H}_{K_\phi}^{s-\delta}$ for every $\delta \in (0, s)$. At the endpoint, $\sum_j |a_j^*|^2 / \lambda_j^s \eqsim \sum_j j^{-1} = \infty$, so $f^* \notin \mathcal{H}_{K_\phi}^s$. Thus, q is the capacity exponent, while s marks the boundary of the target source regularity.

When $0 < s \leqslant 1$, the relation $|\theta_j^*|^2 \eqsim j^{-1-q(s-1)}$ shows that the target sequence is not square-summable and hence does not represent an element of $\mathcal{H}$. All subsequent expressions involving $\boldsymbol{\theta}^*$ are understood coordinatewise in the eigenbasis of $\mathbf{H}$, with prediction series interpreted in $L^2(\mathcal{D})$, which is well defined since $\sum_j \lambda_j |\theta_j^*|^2 < \infty$.

Rotational equivariance. We now verify that SGD, Polyak, and Nesterov are equivariant under rotations. Directions in $\ker \mathbf{H}$ have zero feature variance and do not affect predictions, so we restrict attention to its orthogonal complement. Since $\mathbf{H} = \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}) \otimes \boldsymbol{\phi}(\mathbf{x})]$ is positive, self-adjoint, and trace class, it admits a spectral decomposition

$$\mathbf{H} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^*, \qquad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots),$$

for some unitary operator $\boldsymbol{U}$. Define the rotated variables

$$\boldsymbol{\theta}' = \boldsymbol{U}^* \boldsymbol{\theta}, \qquad \boldsymbol{\phi}'(\mathbf{x}) = \boldsymbol{U}^* \boldsymbol{\phi}(\mathbf{x}).$$

Unitary invariance of the inner product gives

$$f'(\mathbf{x};\boldsymbol{\theta}') = \langle \boldsymbol{\phi}'(\mathbf{x}), \boldsymbol{\theta}' \rangle = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta} \rangle = f(\mathbf{x};\boldsymbol{\theta}),$$

and hence, for every mini-batch $\mathcal{B}_k$,

$$\mathbf{g}'(\boldsymbol{\theta}';\mathcal{B}_k) = \boldsymbol{U}^* \mathbf{g}(\boldsymbol{\theta};\mathcal{B}_k).$$

It follows immediately that the SGD and Polyak recursions are preserved under $\boldsymbol{\theta}'_k = \boldsymbol{U}^* \boldsymbol{\theta}_k$. For Nesterov, the look-ahead point transforms as

$$\boldsymbol{\theta}^{\mathrm{la}\prime}_k = \boldsymbol{U}^* \boldsymbol{\theta}^{\mathrm{la}}_k = \boldsymbol{\theta}'_k + \beta(\boldsymbol{\theta}'_k - \boldsymbol{\theta}'_{k-1}),$$

so its recursion is preserved as well. Finally,

$$\mathbf{H}' = \mathbb{E}[\boldsymbol{\phi}'(\mathbf{x}) \otimes \boldsymbol{\phi}'(\mathbf{x})] = \boldsymbol{U}^* \mathbf{H} \boldsymbol{U} = \boldsymbol{\Lambda}.$$

Thus, SGD, Polyak, and Nesterov are equivariant under unitary changes of basis, while their predictions and risks are unchanged. We may therefore work in the eigenbasis of $\mathbf{H}$, where the covariance operator is diagonal, without loss of generality.

The target is the same function f^* in both representations. If $\boldsymbol{\theta}^* \in \mathcal{H}$, its coefficients transform as $\boldsymbol{\theta}^{*\prime} = \boldsymbol{U}^* \boldsymbol{\theta}^*$. Otherwise, use the finite spectral truncations $f^*_N = \sum_{j=1}^N \theta^*_j \phi_j$, which converge to f^* in $L^2(\mathcal{D})$. The prediction and risk identities then pass to the limit, without applying $\boldsymbol{U}^*$ to a vector outside $\mathcal{H}$.

A.2 Noise-curvature alignment

We derive the covariance identities and noise–curvature alignment property needed for the risk stability and scaling law analyses.

Gaussian fourth moments. The following lemma expresses Gaussian fourth moments in terms of the covariance operator.

Lemma A.1 (Gaussian fourth-moment identities)**.** *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$ in a separable Hilbert space $\mathcal{H}$, where $\mathbf{H}$ is a positive trace-class covariance operator. For any $\mathbf{u}, \mathbf{v} \in \mathcal{H}$,*

$$\mathbb{E}\big[\langle \mathbf{x}, \mathbf{u} \rangle^2 \langle \mathbf{x}, \mathbf{v} \rangle^2\big] = \langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle \langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle + 2\langle \mathbf{u}, \mathbf{H}\mathbf{v} \rangle^2. \tag{17}$$

Equivalently, for any self-adjoint trace-class operator $\mathbf{A}$,

$$\mathbb{E}[(\mathbf{x} \otimes \mathbf{x})\mathbf{A}(\mathbf{x} \otimes \mathbf{x})] = 2\mathbf{H}\mathbf{A}\mathbf{H} + \mathbf{H}\mathrm{Tr}(\mathbf{H}\mathbf{A}). \tag{18}$$

Proof. The directional identity follows directly from Isserlis' formula. For any $\mathbf{u}, \mathbf{v} \in \mathcal{H}$, the random variables $\langle \mathbf{x}, \mathbf{u} \rangle$ and $\langle \mathbf{x}, \mathbf{v} \rangle$ are jointly Gaussian, and hence

$$\mathbb{E}\big[\langle \mathbf{x}, \mathbf{u} \rangle^2 \langle \mathbf{x}, \mathbf{v} \rangle^2\big] = \langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle \langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle + 2\langle \mathbf{u}, \mathbf{H}\mathbf{v} \rangle^2.$$

We next prove the operator identity. Let $\mathbf{A} = \sum_j a_j \mathbf{e}_j \otimes \mathbf{e}_j$ be the spectral decomposition of the self-adjoint trace-class operator $\mathbf{A}$. For arbitrary $\mathbf{u}, \mathbf{v} \in \mathcal{H}$, Isserlis' formula gives

$$\begin{aligned}
&\langle \mathbf{u}, \mathbb{E}[(\mathbf{x} \otimes \mathbf{x})\mathbf{A}(\mathbf{x} \otimes \mathbf{x})]\, \mathbf{v} \rangle \\
&\quad = \sum_j a_j \mathbb{E}\big[\langle \mathbf{u}, \mathbf{x} \rangle \langle \mathbf{e}_j, \mathbf{x} \rangle^2 \langle \mathbf{x}, \mathbf{v} \rangle\big] \\
&\quad = 2\sum_j a_j \langle \mathbf{u}, \mathbf{H}\mathbf{e}_j \rangle \langle \mathbf{e}_j, \mathbf{H}\mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{H}\mathbf{v} \rangle \sum_j a_j \langle \mathbf{e}_j, \mathbf{H}\mathbf{e}_j \rangle \\
&\quad = 2\langle \mathbf{u}, \mathbf{H}\mathbf{A}\mathbf{H}\mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{H}\mathbf{v} \rangle \mathrm{Tr}(\mathbf{H}\mathbf{A}).
\end{aligned}$$

Since this holds for all $\mathbf{u}, \mathbf{v} \in \mathcal{H}$,

$$\mathbb{E}[(\mathbf{x} \otimes \mathbf{x})\mathbf{A}(\mathbf{x} \otimes \mathbf{x})] = 2\mathbf{H}\mathbf{A}\mathbf{H} + \mathbf{H}\mathrm{Tr}(\mathbf{H}\mathbf{A}).$$

□

Corollary A.2 (Empirical covariance fluctuation)**.** *Let*

$$\widehat{\mathbf{H}} := \frac{1}{B}\sum_{b=1}^{B} \mathbf{x}_b \otimes \mathbf{x}_b, \qquad \mathbf{\Sigma} := \widehat{\mathbf{H}} - \mathbf{H}, \qquad \mathbf{x}_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{H}).$$

Then, for any self-adjoint trace-class operator $\mathbf{A}$*,*

$$\mathbb{E}[\mathbf{\Sigma A \Sigma}] = \frac{1}{B}\left(\mathbf{HAH} + \mathbf{H}\mathrm{Tr}(\mathbf{HA})\right). \tag{19}$$

Proof. Write $\mathbf{\Sigma} = B^{-1}\sum_{b=1}^{B} \mathbf{Z}_b$, where $\mathbf{Z}_b := \mathbf{x}_b \otimes \mathbf{x}_b - \mathbf{H}$ are independent and centered. Hence

$$\mathbb{E}[\mathbf{\Sigma A \Sigma}] = \frac{1}{B}\mathbb{E}[\mathbf{ZAZ}] = \frac{1}{B}\left(\mathbb{E}[(\mathbf{x}\otimes\mathbf{x})\mathbf{A}(\mathbf{x}\otimes\mathbf{x})] - \mathbf{HAH}\right),$$

and the result follows from Lemma A.1. □

Lemma A.3 (Noise–curvature alignment)**.** *It holds for any* $\boldsymbol{\nu}$ *that*

$$\mathbb{E}\Big[\left|\langle \boldsymbol{\xi}_k(\boldsymbol{\theta}_k), \boldsymbol{\nu}\rangle\right|^2 \Big| \boldsymbol{\theta}_k\Big] \eqsim \frac{1}{B}\big(\mathcal{E}(\boldsymbol{\theta}_k) + \sigma^2\big)\,\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle\,.$$

Here, $\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle$ is the curvature along $\boldsymbol{\nu}$. The two terms in Lemma A.3 have distinct origins and roles. The term proportional to $\mathcal{E}(\boldsymbol{\theta}_k)$ arises from mini-batch sampling, depends on the current iterate, and vanishes at the optimum, acting as *multiplicative noise*. The term proportional to σ^2 arises from label noise and persists at the optimum, acting as *additive noise*. Both are curvature-aligned but affect the dynamics differently: multiplicative noise feeds back into the dynamics and determines stability, whereas additive noise determines the asymptotic noise floor. This noise–curvature alignment has been observed in neural networks (Zhu et al., 2019; Wu et al., 2022) and established rigorously in toy models (Wang & Wu, 2023).

Proof. We now apply these identities to the mini-batch gradient noise and prove Lemma A.3. For an evaluation point $\boldsymbol{\theta}$ independent of the current mini-batch of size B, write $\boldsymbol{\xi}_k(\boldsymbol{\theta}) = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{\zeta}_{k,i}(\boldsymbol{\theta})$, where $\boldsymbol{\zeta}_{k,i}(\boldsymbol{\theta})$ are conditionally independent, centered single-sample gradient noises. Hence, for any unit vector $\boldsymbol{\nu} \in \mathcal{H}$,

$$\mathbb{E}\big[\,|\langle \boldsymbol{\xi}_k(\boldsymbol{\theta}), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}\big] = \frac{1}{B}\mathbb{E}\big[\,|\langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}\big]\,.$$

Let $\boldsymbol{\delta} := \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and $\boldsymbol{\phi} := \boldsymbol{\phi}(\mathbf{x})$. Since $y = \langle \boldsymbol{\phi}, \boldsymbol{\theta}^*\rangle + \epsilon$, the single-sample gradient is $\mathbf{g}_{\mathbf{x},y}(\boldsymbol{\theta}) = \big(\langle \boldsymbol{\phi}, \boldsymbol{\delta}\rangle - \epsilon\big)\boldsymbol{\phi}$, while $\nabla\mathcal{R}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\delta}$. Therefore,

$$\boldsymbol{\zeta}(\boldsymbol{\theta}) = \langle \boldsymbol{\phi}, \boldsymbol{\delta}\rangle\boldsymbol{\phi} - \mathbf{H}\boldsymbol{\delta} - \epsilon\boldsymbol{\phi}.$$

Using the independence of ϵ and $\boldsymbol{\phi}$, $\mathbb{E}[\epsilon] = 0$, and $\mathbb{E}[\boldsymbol{\phi}\otimes\boldsymbol{\phi}] = \mathbf{H}$, we obtain

$$\mathbb{E}\big[\,|\langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}\big] = \mathbb{E}\big[\langle \boldsymbol{\phi}, \boldsymbol{\delta}\rangle^2\langle \boldsymbol{\phi}, \boldsymbol{\nu}\rangle^2\big] - \langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\nu}\rangle^2 + \sigma^2\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle.$$

If $\boldsymbol{\delta} \notin \mathcal{H}$, apply Lemma A.1 first to its finite spectral truncations. Since $\sum_j \lambda_j\delta_j^2 = 2\mathcal{E}(\boldsymbol{\theta}) < \infty$, the corresponding Gaussian prediction errors converge in L^4, so their fourth-moment identities pass to the limit. Thus the lemma gives

$$\mathbb{E}\big[\langle \boldsymbol{\phi}, \boldsymbol{\delta}\rangle^2\langle \boldsymbol{\phi}, \boldsymbol{\nu}\rangle^2\big] = \langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\delta}\rangle\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle + 2\langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\nu}\rangle^2.$$

Thus,

$$\mathbb{E}\big[\,|\langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}\big] = \big(\sigma^2 + \langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\delta}\rangle\big)\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle + \langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\nu}\rangle^2. \tag{20}$$

The last term is nonnegative and, by Cauchy–Schwarz in the $\mathbf{H}$-inner product,

$$\langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\nu}\rangle^2 \leqslant \langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\delta}\rangle\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle.$$

Since $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2}\langle \boldsymbol{\delta}, \mathbf{H}\boldsymbol{\delta}\rangle$, Eq. (20) implies

$$\mathbb{E}\big[\,|\langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}\big] \eqsim \big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2\big)\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle.$$

Combining this with the $1/B$ variance reduction from mini-batching and setting $\boldsymbol{\theta} = \boldsymbol{\theta}_k$ yields

$$\mathbb{E}\big[\,|\langle \boldsymbol{\xi}_k(\boldsymbol{\theta}_k), \boldsymbol{\nu}\rangle|^2 \big|\, \boldsymbol{\theta}_k\big] \eqsim \frac{1}{B}\big(\mathcal{E}(\boldsymbol{\theta}_k) + \sigma^2\big)\langle \boldsymbol{\nu}, \mathbf{H}\boldsymbol{\nu}\rangle.$$

□

A.3 SECOND-ORDER RECURSIONS

We collect solution formulas and bounds for second-order recursions. These tools separate the evolution of the initial error from the response to gradient noise in the scaling law analysis.

Consider the homogeneous recursion

$$x_{k+1} + ax_k + bx_{k-1} = 0, \qquad k \geqslant 0, \tag{21}$$

with prescribed initial values x_{-1}, x_0. Its characteristic polynomial is

$$r^2 + ar + b = 0. \tag{22}$$

Lemma A.4 (Second-order iteration)**.** *Let r_+ and r_- be the roots of (22). If $r_+ \neq r_-$, then the unique solution of (21) with initial values x_{-1}, x_0 is*

$$x_k = c_+ r_+^k + c_- r_-^k, \qquad k \geqslant 0, \tag{23}$$

where

$$c_+ = \frac{r_+(x_0 - r_- x_{-1})}{r_+ - r_-}, \qquad c_- = \frac{r_-(r_+ x_{-1} - x_0)}{r_+ - r_-}. \tag{24}$$

If $r_+ = r_- =: r$, then

$$x_k = (c_1 + c_2 k) r^k, \tag{25}$$

where $c_1 = x_0$ and $c_2 = x_0 - r x_{-1}$ when $r \neq 0$. If the repeated root is $r = 0$, then $a = b = 0$ and $x_k = 0$ for $k \geqslant 1$, with x_{-1}, x_0 prescribed.

Proof. For distinct roots, substituting (24) into (23) gives the prescribed value at $k = 0$ and satisfies the recursion at $k = 0$ with the prescribed x_{-1}. Each root then generates a solution for $k \geqslant 1$. The repeated-root formula follows by taking the limit as the two roots approach the same value. When $r = 0$, the recursion directly gives $x_k = 0$ for $k \geqslant 1$. In each case, uniqueness follows by induction. □

For the normalized initial values $x_{-1} = x_0 = 1$, the same formula reads

$$x_k = r_+^k + r_-(1 - r_+) D_k(r_+, r_-), \qquad D_k(r_+, r_-) := \sum_{\ell=0}^{k-1} r_+^{k-1-\ell} r_-^{\ell}, \tag{26}$$

where $D_0 = 0$, with the convention $0^0 = 1$. For distinct roots, $D_k = (r_+^k - r_-^k)/(r_+ - r_-)$, while $D_k(r, r) = k r^{k-1}$ for $k \geqslant 1$. In particular, $|D_k| \leqslant k R^{k-1}$ for $k \geqslant 1$ if $|r_\pm| \leqslant R$ and $R > 0$. This bound remains valid when the two roots coincide, so it controls modes near the damping threshold without dividing by a small root separation.

Lemma A.5 (Forced second-order iteration)**.** *Consider the forced recursion*

$$\tilde{x}_{k+1} + a\tilde{x}_k + b\tilde{x}_{k-1} + z_k = 0, \qquad k \geqslant 0, \tag{27}$$

with arbitrary initial values $\tilde{x}_{-1}, \tilde{x}_0$. Let x_k^{det} be the solution of the homogeneous recursion

$$x_{k+1}^{\mathrm{det}} + a x_k^{\mathrm{det}} + b x_{k-1}^{\mathrm{det}} = 0, \qquad x_{-1}^{\mathrm{det}} = \tilde{x}_{-1}, \qquad x_0^{\mathrm{det}} = \tilde{x}_0. \tag{28}$$

Let G_k be the Green function associated with (21), defined by

$$G_{k+1} + aG_k + bG_{k-1} = 0, \qquad G_{-1} = 0, \qquad G_0 = 1. \tag{29}$$

Then the unique solution of (27) is

$$\tilde{x}_k = x_k^{\mathrm{det}} - \sum_{m=0}^{k-1} G_{k-1-m} z_m, \qquad k \geqslant 0. \tag{30}$$

If $r_+ \neq r_-$, then

$$G_k = \frac{r_+^{k+1} - r_-^{k+1}}{r_+ - r_-}, \qquad k \geqslant 0, \tag{31}$$

and at a repeated root $r_+ = r_- = r$ this is interpreted by continuity as $G_k = (k+1) r^k$.

Proof. Set $w_k = -\sum_{m=0}^{k-1} G_{k-1-m} z_m$, with $w_{-1} = w_0 = 0$. Using the recursion for G, all terms cancel in $w_{k+1} + a w_k + b w_{k-1}$ except $-G_0 z_k = -z_k$. Thus $x^{\mathrm{det}} + w$ has the prescribed initial values and satisfies (27). Uniqueness follows by induction. The initial values $G_{-1} = 0$, $G_0 = 1$ and Lemma A.4 give (31). Taking the limit as the two roots approach the same value gives $G_k = (k+1)r^k$, with $G_0 = 1$ also when $r = 0$. □

The next lemma gives the total energy of the Green function and a bound on its tail. We use these results to estimate the forgetting kernel sums in the label noise proofs and to control multiplicative noise.

Lemma A.6 (Energy and tail bounds for the Green function)**.** *Consider the recursion*

$$G_{k+1} + aG_k + bG_{k-1} = 0, \qquad G_{-1} = 0, \quad G_0 = 1.$$

Let $r_\pm$ be the roots of $r^2 + ar + b = 0$, and assume $|r_\pm| < 1$. Then the following statements hold.

1. *The total energy satisfies*

$$\sum_{k=0}^{\infty} G_k^2 = \frac{1+b}{(1-b)\big((1+b)^2 - a^2\big)}. \tag{32}$$

2. *For any $z > 1$ satisfying $\sqrt{z}\,|r_\pm| < 1$, the weighted energy satisfies*

$$\sum_{\tau=0}^{\infty} z^\tau G_\tau^2 = \frac{1+bz}{(1-bz)\big((1+bz)^2 - a^2 z\big)}, \tag{33}$$

3. *For the same z, the tail satisfies*

$$\sum_{\tau=k}^{\infty} G_\tau^2 \leqslant z^{-k} \sum_{\tau=0}^{\infty} z^\tau G_\tau^2. \tag{34}$$

Proof. Let $Q = \sum_{k\geqslant 0} G_k^2$ and $R = \sum_{k\geqslant 0} G_k G_{k-1}$. The root condition ensures convergence. Multiplying the recursion by G_k and summing gives $(1+b)R = -aQ$. Squaring it and summing gives $Q - 1 = (a^2 + b^2)Q + 2abR$. Eliminating R yields

$$1 = Q\,\frac{1-b}{1+b}\big((1+b)^2 - a^2\big),$$

which is (32). For the weighted identity, set $\widehat{G}_k := z^{k/2} G_k$. This sequence satisfies the recursion with coefficients $a\sqrt{z}$ and bz, with $\widehat{G}_{-1} = 0$ and $\widehat{G}_0 = 1$. Its characteristic roots are $\sqrt{z}\,r_\pm$, whose moduli are less than one by assumption. Applying (32) to this sequence gives (33). Finally, $z^\tau \geqslant z^k$ for $\tau \geqslant k$, which proves (34). □

We next bound the overdamped coordinate decay for Polyak and Nesterov. These estimates give the signal learning bounds in Appendix C.2 and C.3.

Lemma A.7 (Overdamped coordinate decay for Polyak)**.** *Let $0 \leqslant \beta < 1$ and $x > 0$. Consider the recursion*

$$h_{k+1} = (1 + \beta - x)h_k - \beta h_{k-1}, \qquad h_{-1} = h_0 = 1.$$

If $x \leqslant c_0(1-\beta)^2$ for a sufficiently small absolute constant $c_0 > 0$, then

$$c e^{-Cxk/(1-\beta)} \leqslant h_k \leqslant C e^{-cxk/(1-\beta)}, \qquad k \geqslant 0, \tag{35}$$

for absolute constants $c, C > 0$.

Proof. The characteristic polynomial is $p(r) = r^2 - (1 + \beta - x)r + \beta$. To estimate the distance of its roots from 1, set $r = 1 - u$. Then

$$p(1-u) = u^2 - \gamma u + x, \qquad \gamma = 1 - \beta + x.$$

Thus $1-\beta \leqslant \gamma \leqslant 1-\beta+2x$. Since $\gamma \eqsim 1-\beta$ and $\gamma^2-4x \geqslant (1-4c_0)(1-\beta)^2 > 0$, the roots $u_+ \leqslant u_-$ are real and positive. Set $r_+ = 1-u_+$ and $r_- = 1-u_-$. Then

$$u_+ = \frac{2x}{\gamma + \sqrt{\gamma^2-4x}} \eqsim \frac{x}{1-\beta}, \qquad r_+ - r_- = \sqrt{\gamma^2-4x} \eqsim 1-\beta.$$

Since $u_+ \lesssim c_0$, we have $r_+ > 0$. The identity $r_+ r_- = \beta \geqslant 0$ gives $0 \leqslant r_- \leqslant r_+ < 1$. By (26),

$$h_k = r_+^k + d(r_+^k - r_-^k), \qquad 0 \leqslant d := \frac{r_-(1-r_+)}{r_+ - r_-} \lesssim \frac{x}{(1-\beta)^2} \leqslant c_0.$$

Thus $r_+^k \leqslant h_k \leqslant (1+Cc_0)r_+^k$. Since $-\log r_+ \eqsim 1-r_+ \eqsim x/(1-\beta)$, this proves the stated bounds. □

Lemma A.8 (Overdamped coordinate decay for Nesterov)**.** *Let $0 \leqslant \beta < 1$ and $x > 0$. Consider the recursion*

$$h_{k+1} = (1+\beta)(1-x)h_k - \beta(1-x)h_{k-1}, \qquad h_{-1} = h_0 = 1.$$

If $x \leqslant c_0(1-\beta)^2$ for a sufficiently small absolute constant $c_0 > 0$, then

$$ce^{-Cxk/(1-\beta)} \leqslant h_k \leqslant Ce^{-cxk/(1-\beta)}, \qquad k \geqslant 0, \tag{36}$$

for absolute constants $c, C > 0$.

Proof. The characteristic polynomial is $p(r) = r^2 - (1+\beta)(1-x)r + \beta(1-x)$. To estimate the distance of its roots from 1, set $r = 1-u$. Then

$$p(1-u) = u^2 - \gamma u + x, \qquad \gamma = 1-\beta+(1+\beta)x.$$

Thus $1-\beta \leqslant \gamma \leqslant 1-\beta+2x$. Since $\gamma \eqsim 1-\beta$ and $\gamma^2-4x \geqslant (1-4c_0)(1-\beta)^2 > 0$, the roots $u_+ \leqslant u_-$ are real and positive. Set $r_+ = 1-u_+$ and $r_- = 1-u_-$. Then

$$u_+ = \frac{2x}{\gamma + \sqrt{\gamma^2-4x}} \eqsim \frac{x}{1-\beta}, \qquad r_+ - r_- = \sqrt{\gamma^2-4x} \eqsim 1-\beta.$$

Since $u_+ \lesssim c_0$, we have $r_+ > 0$. Since $x \leqslant c_0(1-\beta)^2 < 1$, the identity $r_+ r_- = \beta(1-x) \geqslant 0$ gives $0 \leqslant r_- \leqslant r_+ < 1$. By (26),

$$h_k = r_+^k + d(r_+^k - r_-^k), \qquad 0 \leqslant d := \frac{r_-(1-r_+)}{r_+ - r_-} \lesssim \frac{x}{(1-\beta)^2} \leqslant c_0.$$

Thus $r_+^k \leqslant h_k \leqslant (1+Cc_0)r_+^k$. Since $-\log r_+ \eqsim 1-r_+ \eqsim x/(1-\beta)$, this proves the stated bounds. □

A.4 Power-law spectral sums

Coordinatewise decay bounds reduce signal learning to sums of exponentials over the spectrum. The next lemma covers both the full spectrum and the overdamped tail, so the same estimate applies to both momentum methods.

Lemma A.9 (Power-law spectral sums)**.** *Let $q, s, c > 0$. For every integer $J \geqslant 1$ and $t \geqslant 0$,*

$$\sum_{j \geqslant J} j^{-1-qs} e^{-ctj^{-q}} \eqsim (J^q + t)^{-s}. \tag{37}$$

The comparison constants depend only on q, s, c. In particular, the full sum is of order $(1+t)^{-s}$, and the sum at $t = 0$ is of order J^{-qs}.

Proof. Write $F_J(t)$ for the sum in (37). First suppose $0 \leqslant t \leqslant J^q$. The integral test gives

$$F_J(t) \leqslant \sum_{j \geqslant J} j^{-1-qs} \leqslant J^{-1-qs} + \int_J^\infty x^{-1-qs}\, dx \lesssim J^{-qs}.$$

For $J \leqslant j \leqslant 2J$, we have $tj^{-q} \leqslant 1$. Therefore,

$$F_J(t) \geqslant e^{-c} \sum_{j=J}^{2J} j^{-1-qs} \geqslant e^{-c}(J+1)(2J)^{-1-qs} \gtrsim J^{-qs}.$$

Because $J^q \leqslant J^q + t \leqslant 2J^q$, these bounds give (37) in this case.

Now suppose $t > J^q$ and put $L := \lceil t^{1/q} \rceil$. Then $J \leqslant L \leqslant 2t^{1/q}$. For $L \leqslant j \leqslant 2L$, $tj^{-q} \leqslant 1$, so

$$F_J(t) \geqslant e^{-c} \sum_{j=L}^{2L} j^{-1-qs} \geqslant e^{-c}(L+1)(2L)^{-1-qs} \gtrsim t^{-s}.$$

For the upper bound, set $f_t(x) := x^{-1-qs} e^{-ctx^{-q}}$. Its derivative has the sign of $cqtx^{-q} - (1+qs)$, so f_t increases and then decreases on $(0, \infty)$. Compare each term before the maximum with the integral on the interval to its right, and each term after the maximum with the integral on the interval to its left. At most two terms remain, each bounded by the maximum. Thus

$$F_J(t) \leqslant \sum_{j \geqslant 1} f_t(j) \leqslant \int_1^\infty f_t(x)\, dx + 2 \sup_{x \geqslant 1} f_t(x).$$

With $u = ctx^{-q}$,

$$\int_1^\infty f_t(x)\, dx = \frac{(ct)^{-s}}{q} \int_0^{ct} u^{s-1} e^{-u}\, du \leqslant \frac{\Gamma(s)}{q(ct)^s}.$$

The maximum over $x > 0$ is $O(t^{-s-1/q})$, which is $O(t^{-s})$ because $t > J^q \geqslant 1$. Thus $F_J(t) \lesssim t^{-s}$. Since $t < J^q + t < 2t$, the claimed comparison follows. □

For a truncated sum, nonnegativity and the full-sum case $J = 1$ give

$$\sum_{j=1}^{J} j^{-1-qs} e^{-ctj^{-q}} \leqslant \sum_{j \geqslant 1} j^{-1-qs} e^{-ctj^{-q}} \lesssim (1+t)^{-s}, \qquad t \geqslant 0. \tag{38}$$

This is the bound used for the Nesterov underdamped signal learning term in Proposition C.14. No restriction on t is needed for this upper bound. A matching lower bound would require $t \lesssim J^q$: for fixed J, the truncated sum decays exponentially once t/J^q is large, whereas the full sum retains its power-law tail.

A.5 A CONVOLUTION INEQUALITY

The following convolution bound is used to control the multiplicative noise in Appendix C.2 and C.3. The index shift accommodates the different gradient evaluation points of the two methods.

Lemma A.10 (Boundedness of a convolution inequality)**.** *Let (x_k), (u_k), (y_k), and (w_k) be nonnegative sequences, with $U := \sup_k u_k < \infty$, and let $v \geqslant 0$. Suppose*

$$x_k \leqslant u_k + \sum_{\tau=0}^{k-1} w_\tau \big(v + y_{k-1-\tau}\big), \qquad k \geqslant 0,$$

where the sum is empty for $k = 0$. Suppose there exist a constant $L > 0$ and a fixed index shift $\iota \in \{0, 1\}$ such that $y_m \leqslant L x_{m+\iota}$ for all $m \geqslant 0$. If $L \sum_{\tau \geqslant 0} w_\tau \leqslant 1/2$, then

$$\sup_{k \geqslant 0} x_k \leqslant 2U + \frac{v}{L}.$$

Proof. Set $W := \sum_{\tau \geqslant 0} w_\tau$ and $X_n := \max_{0 \leqslant k \leqslant n} x_k$. Since $k - 1 - \tau + \iota \leqslant n$ for $1 \leqslant k \leqslant n$, we have $X_n \leqslant U + vW + LWX_n$. Thus $X_n \leqslant 2(U + vW) \leqslant 2U + v/L$, which proves the claim. □

B RISK STABILITY ANALYSIS

B.1 RENEWAL FORMULATION OF RISK STABILITY

We first recall a general boundedness criterion for renewal equations (Feller, 1991). Consider a nonnegative sequence $(x_k)_{k\geqslant 0}$ satisfying

$$x_k = a_k + \sum_{j=0}^{k-1} K_{k-1-j} x_j,$$

where $a_k \geqslant 0$ is the input and $K = (K_k)_{k\geqslant 0}$ is the renewal kernel. Its total mass $\|K\|_1 := \sum_{k\geqslant 0} K_k$ measures the overall feedback strength of the recursion.

Lemma B.1 (Boundedness of a renewal equation)**.** *Suppose* $0 < \underline{a} := \inf_{k\geqslant 0} a_k \leqslant \sup_{k\geqslant 0} a_k =: \bar{a} < \infty$. *Then*

$$\sup_{k\geqslant 0} x_k < \infty \quad \Longleftrightarrow \quad \|K\|_1 < 1.$$

Proof. Suppose first that $\|K\|_1 < 1$, and let $M_k := \max_{0\leqslant j\leqslant k} x_j$. Then $x_k \leqslant \bar{a} + \|K\|_1 M_{k-1}$, which implies

$$\sup_{k\geqslant 0} x_k \leqslant \max\left\{x_0, \frac{\bar{a}}{1-\|K\|_1}\right\} < \infty.$$

Conversely, suppose that (x_k) is bounded and let $L := \liminf_{k\to\infty} x_k$. For every fixed N, nonnegativity gives

$$x_k \geqslant \underline{a} + \sum_{m=0}^{N} K_m x_{k-1-m}.$$

Taking $\liminf_{k\to\infty}$ yields $L \geqslant \underline{a} + L\sum_{m=0}^{N} K_m$. Letting $N \to \infty$, boundedness is impossible if $\sum_{m\geqslant 0} K_m \geqslant 1$. □

For each method $\mathcal{A} \in \{\mathrm{sgd}, \mathrm{polyak}, \mathrm{nesterov}\}$, we will derive a nonnegative risk moment $\ell_k^{(\mathcal{A})}$, equal to $2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)]$ for SGD and Polyak and comparable to it for Nesterov, satisfying

$$\ell_k^{(\mathcal{A})} = h_k^{(\mathcal{A})} + \sum_{j=0}^{k-1} K_{k-1-j}^{(\mathcal{A})} \big(\ell_j^{(\mathcal{A})} + \sigma^2\big). \tag{39}$$

Setting $x_k = \ell_k^{(\mathcal{A})} + \sigma^2$ and $a_k = h_k^{(\mathcal{A})} + \sigma^2$ reduces (39) to the renewal equation above. Hence, once the coordinate-wise dynamics are stable and $h_k^{(\mathcal{A})}$ is bounded, Lemma B.1 gives the global feedback condition

$$\kappa_{\mathcal{A}}(\eta, \beta, B) := \|K^{(\mathcal{A})}\|_1 = \sum_{k\geqslant 0} K_k^{(\mathcal{A})} < 1.$$

The kernel $K^{(\mathcal{A})}$ describes how perturbations propagate through the optimizer dynamics and depends on the method and hyperparameters (η, β, B), but not on the label-noise variance σ^2. Additive label noise enters as the constant external forcing σ^2, whereas multiplicative noise feeds the current risk back through $K^{(\mathcal{A})}$. In Appendix B.2–B.4, we derive $K^{(\mathcal{A})}$, compute $\kappa_{\mathcal{A}}$, and combine $\kappa_{\mathcal{A}} < 1$ with the corresponding coordinate-wise stability condition to obtain the critical learning rates.

B.2 RISK STABILITY OF SGD

The SGD error recursion, including label noise, is

$$\mathbf{e}_{k+1} = (\mathbf{I} - \eta\widehat{\mathbf{H}}_k)\mathbf{e}_k + \eta\widehat{\boldsymbol{\zeta}}_k. \tag{40}$$

Theorem B.2 (Risk stability of SGD)**.** *Let the Hessian be a positive trace-class operator with spectrum*

$$\mathbf{H} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots), \qquad \lambda_1 = \lambda_{\max} \geqslant \lambda_2 \geqslant \cdots > 0,$$

and suppose $\sigma > 0$. Then SGD is risk stable if and only if

$$\eta < \frac{2B}{(B+1)\lambda_{\max}} \tag{41}$$

and

$$\kappa_{\mathrm{sgd}}(\eta) := \sum_{i \geqslant 1} \frac{\eta\lambda_i}{2B - (B+1)\eta\lambda_i} < 1. \tag{42}$$

Proof. Let $\{\boldsymbol{\varphi}_i\}_{i \geqslant 1}$ denote the orthonormal eigenbasis of $\mathbf{H}$. We work with its spectral coordinates. For every finite-risk initialization, the relevant risk moment is

$$\ell_k := \sum_{i \geqslant 1} \lambda_i \mathbb{E}\big[|\langle \boldsymbol{\varphi}_i, \mathbf{e}_k \rangle|^2\big] = 2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)].$$

In particular, $\ell_0 < \infty$. The coordinate identities below may be justified first on finite spectral truncations and then passed to the limit by monotone convergence. Expanding (40), using the vanishing cross terms, and applying Lemma A.1 gives, for every $i \geqslant 1$,

$$\mathbb{E}\big[|\langle \boldsymbol{\varphi}_i, \mathbf{e}_{k+1} \rangle|^2\big] = (1 - \eta\lambda_i)^2 \mathbb{E}\big[|\langle \boldsymbol{\varphi}_i, \mathbf{e}_k \rangle|^2\big] + \frac{\eta^2}{B}\left(\lambda_i^2 \mathbb{E}\big[|\langle \boldsymbol{\varphi}_i, \mathbf{e}_k \rangle|^2\big] + \lambda_i \ell_k + \sigma^2 \lambda_i\right). \tag{43}$$

Define the coordinate variance by

$$u_{i,k} := \mathbb{E}\big[|\langle \boldsymbol{\varphi}_i, \mathbf{e}_k \rangle|^2\big].$$

With this notation, (43) yields

$$u_{i,k+1} = a_i u_{i,k} + \frac{\eta^2}{B}\lambda_i\big(\ell_k + \sigma^2\big), \qquad a_i := (1 - \eta\lambda_i)^2 + \frac{\eta^2}{B}\lambda_i^2. \tag{44}$$

We first determine the local condition. Since

$$1 - a_i = \eta\lambda_i \left[2 - \left(1 + \frac{1}{B}\right)\eta\lambda_i\right],$$

the coordinatewise contraction condition

$$a_i < 1 \qquad \text{for every } i$$

is equivalent to (41). If this condition fails, then (44) contains an eigendirection satisfying

$$u_{i,k+1} \geqslant u_{i,k} + \frac{\eta^2\sigma^2}{B}\lambda_i,$$

and therefore the risk is unbounded. Hence the local condition is necessary.

Assume henceforth that (41) holds. Iterating (44), multiplying the i-th equation by λ_i, and summing over all eigendirections gives

$$\ell_k = h_k + \sum_{s=0}^{k-1} K_{k-1-s}\big(\ell_s + \sigma^2\big), \tag{45}$$

where

$$h_k := \sum_{i \geqslant 1} \lambda_i a_i^k u_{i,0}, \qquad K_m := \frac{\eta^2}{B}\sum_{i \geqslant 1} \lambda_i^2 a_i^m.$$

The local condition gives

$$0 \leqslant a_i < 1, \qquad 0 \leqslant h_k \leqslant \ell_0.$$

Moreover,

$$\sum_{m \geqslant 0} K_m = \frac{\eta^2}{B}\sum_{i \geqslant 1} \frac{\lambda_i^2}{1 - a_i} = \sum_{i \geqslant 1} \frac{\eta\lambda_i}{2B - (B+1)\eta\lambda_i} = \kappa_{\mathrm{sgd}}(\eta).$$

Lemma B.1 therefore gives

$$\sup_{k\geqslant 0} \ell_k < \infty \quad \Longleftrightarrow \quad \kappa_{\mathrm{sgd}}(\eta) < 1,$$

which completes the proof.

□

Corollary B.3 (Critical learning rate of SGD)**.** *There exists a unique critical learning rate satisfying*

$$0 < \eta_{\mathrm{sgd}}^{\mathrm{crit}} < \frac{2B}{(B+1)\lambda_{\max}}$$

and

$$\sum_{i\geqslant 1} \frac{\eta_{\mathrm{sgd}}^{\mathrm{crit}}\lambda_i}{2B - (B+1)\eta_{\mathrm{sgd}}^{\mathrm{crit}}\lambda_i} = 1. \tag{46}$$

The SGD iteration is risk stable if and only if

$$0 < \eta < \eta_{\mathrm{sgd}}^{\mathrm{crit}}.$$

Moreover,

$$\frac{2B}{(B+1)\lambda_{\max} + \mathrm{Tr}(\mathbf{H})} \leqslant \eta_{\mathrm{sgd}}^{\mathrm{crit}} \leqslant \min\left\{\frac{2B}{(B+1)\lambda_{\max}}, \frac{2B}{\mathrm{Tr}(\mathbf{H})}\right\}. \tag{47}$$

Consequently, under the fixed-exponent power-law assumptions $\lambda_i \eqsim i^{-q}, q > 1$, we have

$$\eta_{\mathrm{sgd}}^{\mathrm{crit}} \eqsim 1.$$

Proof. The renewal mass $\kappa_{\mathrm{sgd}}(\eta)$ is continuous and strictly increasing on the local stability interval. It vanishes at $\eta = 0$ and diverges as η approaches the local boundary, because the denominator of the leading eigendirection tends to zero. Hence there is a unique solution of $\kappa_{\mathrm{sgd}}(\eta) = 1$, which gives the stated critical learning rate and stability interval.

At the critical point, every denominator in (46) is at most $2B$. Therefore $1 \geqslant \frac{\eta_{\mathrm{sgd}}^{\mathrm{crit}}}{2B}\mathrm{Tr}(\mathbf{H})$, which yields

$$\eta_{\mathrm{sgd}}^{\mathrm{crit}} \leqslant \frac{2B}{\mathrm{Tr}(\mathbf{H})}.$$

The local condition gives the other upper bound in (47). Conversely, every denominator is at least

$$2B - (B+1)\eta_{\mathrm{sgd}}^{\mathrm{crit}}\lambda_{\max}.$$

Using the critical equation then gives

$$1 \leqslant \frac{\eta_{\mathrm{sgd}}^{\mathrm{crit}}\mathrm{Tr}(\mathbf{H})}{2B - (B+1)\eta_{\mathrm{sgd}}^{\mathrm{crit}}\lambda_{\max}},$$

and rearranging yields the lower bound in (47).

Finally, under $\lambda_i \eqsim i^{-q}$ with $q > 1$, we have $\lambda_{\max} \eqsim 1$ and $\mathrm{Tr}(\mathbf{H}) \eqsim 1$. The two-sided bounds therefore imply $\eta_{\mathrm{sgd}}^{\mathrm{crit}} \eqsim 1$. □

B.3 Risk stability of Polyak

The Polyak error recursion, including label noise, is

$$\mathbf{e}_{k+1} = (\mathbf{I} - \eta\widehat{\mathbf{H}}_k)\mathbf{e}_k + \beta(\mathbf{e}_k - \mathbf{e}_{k-1}) + \eta\widehat{\boldsymbol{\zeta}}_k. \tag{48}$$

Writing $\mathbf{D} = (1+\beta)\mathbf{I} - \eta\mathbf{H}$, Equation (48) becomes

$$\mathbf{e}_{k+1} = \mathbf{D}\mathbf{e}_k - \beta\mathbf{e}_{k-1} - \eta(\widehat{\mathbf{H}}_k - \mathbf{H})\mathbf{e}_k + \eta\widehat{\boldsymbol{\zeta}}_k.$$

Theorem B.4 (Risk stability of Polyak)**.** *Let the Hessian be a positive trace-class operator and assume*

$$\mathbf{H} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots), \qquad \lambda_1 = \lambda_{\max} \geqslant \lambda_2 \geqslant \cdots > 0, \qquad 0 \leqslant \beta < 1, \qquad \sigma > 0.$$

Then Polyak momentum is risk stable if and only if

$$\eta < \frac{2B(1-\beta^2)}{\lambda_{\max}\left[B(1-\beta)+(1+\beta)\right]} \tag{49}$$

and

$$\kappa_{\mathrm{polyak}}(\eta,\beta) := \sum_{i\geqslant 1} \frac{\eta\lambda_i(1+\beta)}{2B(1-\beta^2)-\eta\lambda_i\left[B(1-\beta)+(1+\beta)\right]} < 1. \tag{50}$$

Proof. Define the coordinate second moments

$$\begin{aligned} \mathbf{U}_k &:= \mathbb{E}[\mathbf{e}_k \otimes \mathbf{e}_k], \qquad \mathbf{P}_k := \mathbb{E}[\mathbf{e}_k \otimes \mathbf{e}_{k-1}], \\ \mathbf{Q}_k &:= \mathbb{E}[\mathbf{e}_{k-1} \otimes \mathbf{e}_k], \quad \mathbf{R}_k := \mathbb{E}[\mathbf{e}_{k-1} \otimes \mathbf{e}_{k-1}]. \end{aligned}$$

The risk moment relevant to risk stability is

$$\ell_k := \mathrm{Tr}(\mathbf{H}\mathbf{U}_k) = 2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)].$$

Expanding (48), using the vanishing cross terms, and applying Lemma A.1 gives

$$\begin{cases} \mathbf{U}_{k+1} = \mathbf{D}\mathbf{U}_k\mathbf{D} - \beta\mathbf{D}\mathbf{P}_k - \beta\mathbf{Q}_k\mathbf{D} + \beta^2\mathbf{R}_k \\ \qquad\qquad + \dfrac{\eta^2}{B}\left(\mathbf{H}\mathbf{U}_k\mathbf{H} + \mathbf{H}\mathrm{Tr}(\mathbf{H}\mathbf{U}_k) + \sigma^2\mathbf{H}\right), \\ \mathbf{P}_{k+1} = \mathbf{D}\mathbf{U}_k - \beta\mathbf{Q}_k, \\ \mathbf{Q}_{k+1} = \mathbf{U}_k\mathbf{D} - \beta\mathbf{P}_k, \\ \mathbf{R}_{k+1} = \mathbf{U}_k. \end{cases} \tag{51}$$

In the eigenbasis of the Hessian, define

$$u_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{U}_k\boldsymbol{\varphi}_i\rangle, \qquad v_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{P}_k\boldsymbol{\varphi}_i\rangle, \qquad r_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{R}_k\boldsymbol{\varphi}_i\rangle.$$

The diagonal entries of the two cross-covariance operators coincide because $\mathbf{Q}_k = \mathbf{P}_k^*$. Set

$$d_i := 1 + \beta - \eta\lambda_i, \qquad \mathbf{X}_{i,k} := (u_{i,k}, v_{i,k}, r_{i,k})^\top, \qquad \mathbf{b} := (1,0,0)^\top,$$

and

$$\mathbf{M}_i := \begin{pmatrix} d_i^2 + \dfrac{\eta^2}{B}\lambda_i^2 & -2\beta d_i & \beta^2 \\ d_i & -\beta & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Taking the diagonal entries of (51) yields

$$\mathbf{X}_{i,k+1} = \mathbf{M}_i\mathbf{X}_{i,k} + \frac{\eta^2}{B}\lambda_i\mathbf{b}\left(\ell_k + \sigma^2\right). \tag{52}$$

After removing the global risk-coupling term, the characteristic polynomial of the isolated block is

$$p_i(\mu) = \mu^3 + \left(\beta - d_i^2 - \frac{\eta^2}{B}\lambda_i^2\right)\mu^2 + \beta\left(d_i^2 - \frac{\eta^2}{B}\lambda_i^2 - \beta\right)\mu - \beta^3.$$

The Jury stability criterion shows that all roots lie strictly inside the unit disk if and only if

$$2B(1-\beta^2) - \eta\lambda_i\left[B(1-\beta)+(1+\beta)\right] > 0.$$

Requiring this inequality in every eigendirection gives (49). If the local condition fails, the isolated second-moment system has a mode on or outside the unit circle. Because the label noise forcing enters its current-error coordinate directly and has strictly positive variance, the corresponding coordinate risk is unbounded. Thus, the local condition is necessary.

Assume henceforth that the local condition holds, and define $g_{i,m} := \mathbf{b}^\top \mathbf{M}_i^m \mathbf{b}$. This quantity is nonnegative. Iterating (52), taking the first component, multiplying by the corresponding eigenvalue, and summing over all eigendirections gives

$$\ell_k = h_k + \sum_{s=0}^{k-1} K_{k-1-s}\big(\ell_s + \sigma^2\big), \tag{53}$$

where

$$h_k := \sum_{i\geqslant 1} \lambda_i \mathbf{b}^\top \mathbf{M}_i^k \mathbf{X}_{i,0}, \qquad K_m := \frac{\eta^2}{B} \sum_{i\geqslant 1} \lambda_i^2 g_{i,m}.$$

The strict local condition gives the uniform power bound $\sup_{i,k} \|\mathbf{M}_i^k\| < \infty$. Positivity of the initial coordinate covariance also gives $|v_{i,0}| \leqslant \frac{u_{i,0}+r_{i,0}}{2}$. Together, these estimates imply

$$0 \leqslant h_k \leqslant C \sum_{i\geqslant 1} \lambda_i (u_{i,0} + r_{i,0}) < \infty.$$

A direct calculation gives

$$\mathbf{b}^\top (\mathbf{I} - \mathbf{M}_i)^{-1} = \frac{B\big(1+\beta, -2\beta d_i, \beta^2(1+\beta)\big)}{\eta\lambda_i \left\{2B(1-\beta^2) - \eta\lambda_i \left[B(1-\beta) + (1+\beta)\right]\right\}}. \tag{54}$$

Consequently,

$$\sum_{m\geqslant 0} K_m = \frac{\eta^2}{B} \sum_{i\geqslant 1} \lambda_i^2 \mathbf{b}^\top (\mathbf{I} - \mathbf{M}_i)^{-1} \mathbf{b} = \kappa_{\text{polyak}}(\eta, \beta).$$

Lemma B.1 therefore gives

$$\sup_{k\geqslant 0} \ell_k < \infty \quad \Longleftrightarrow \quad \kappa_{\text{polyak}}(\eta, \beta) < 1,$$

which completes the proof.

□

Corollary B.5 (Critical learning rate of Polyak)**.** *There exists a unique critical learning rate satisfying*

$$0 < \eta_{\text{polyak}}^{\text{crit}} < \frac{2B(1-\beta^2)}{\lambda_{\max}\left[B(1-\beta) + (1+\beta)\right]}$$

and the critical equation

$$\sum_{i\geqslant 1} \frac{\eta_{\text{polyak}}^{\text{crit}} \lambda_i (1+\beta)}{2B(1-\beta^2) - \eta_{\text{polyak}}^{\text{crit}} \lambda_i \left[B(1-\beta) + (1+\beta)\right]} = 1. \tag{55}$$

The Polyak iteration is risk stable if and only if

$$0 < \eta < \eta_{\text{polyak}}^{\text{crit}}.$$

Moreover,

$$\begin{aligned} \frac{2B(1-\beta^2)}{\lambda_{\max}\left[B(1-\beta) + (1+\beta)\right] + (1+\beta)\text{Tr}(\mathbf{H})} &\leqslant \eta_{\text{polyak}}^{\text{crit}}, \\ \eta_{\text{polyak}}^{\text{crit}} &\leqslant \min\left\{ \frac{2B(1-\beta^2)}{\lambda_{\max}\left[B(1-\beta) + (1+\beta)\right]}, \frac{2B(1-\beta)}{\text{Tr}(\mathbf{H})} \right\}. \end{aligned} \tag{56}$$

Under the power-law assumption $\lambda_i \eqsim i^{-q}, q > 1$, when $B \gg 1, 1-\beta \ll 1$, we have

$$\eta_{\text{polyak}}^{\text{crit}} \eqsim \min\{1, B(1-\beta)\}.$$

Proof. The renewal mass $\kappa_{\text{polyak}}(\eta, \beta)$ is continuous and strictly increasing throughout the local stability interval. It vanishes at the origin and diverges at the local boundary. Theorem B.4 therefore gives a unique critical learning rate and the stated stability interval.

At the critical point, every denominator in (55) is at most $2B(1-\beta^2)$ and hence

$$\eta^{\mathrm{crit}}_{\mathrm{polyak}} \leqslant \frac{2B(1-\beta)}{\mathrm{Tr}(\mathbf{H})},$$

while the local condition gives the other upper bound in (56). Conversely, every denominator is at least $2B(1-\beta^2) - \eta^{\mathrm{crit}}_{\mathrm{polyak}}\lambda_{\max}\left[B(1-\beta)+(1+\beta)\right]$. Substitution into the critical equation gives

$$1 \leqslant \frac{\eta^{\mathrm{crit}}_{\mathrm{polyak}}(1+\beta)\mathrm{Tr}(\mathbf{H})}{2B(1-\beta^2) - \eta^{\mathrm{crit}}_{\mathrm{polyak}}\lambda_{\max}\left[B(1-\beta)+(1+\beta)\right]}.$$

Rearranging this inequality gives the lower bound in (56).

Finally, the power-law spectrum and the high-momentum regime give

$$\lambda_{\max} \eqsim 1, \qquad \mathrm{Tr}(\mathbf{H}) \eqsim 1, \qquad 1-\beta^2 \eqsim 1-\beta.$$

Consequently, both sides of (56) scale as

$$\frac{B(1-\beta)}{B(1-\beta)+1} \eqsim \min\{1, B(1-\beta)\}.$$

□

B.4 Risk stability of Nesterov

Define the look-ahead error

$$\mathbf{e}^{\mathrm{la}}_k := (1+\beta)\mathbf{e}_k - \beta\mathbf{e}_{k-1}.$$

The Nesterov error recursion, including label noise, is

$$\mathbf{e}_{k+1} = (\mathbf{I} - \eta\widehat{\mathbf{H}}_k)\mathbf{e}^{\mathrm{la}}_k + \eta\widehat{\boldsymbol{\zeta}}_k. \tag{57}$$

With the deterministic transition $\mathbf{D} = \mathbf{I} - \eta\mathbf{H}$, the recursion becomes

$$\mathbf{e}_{k+1} = \mathbf{D}\mathbf{e}^{\mathrm{la}}_k - \eta(\widehat{\mathbf{H}}_k - \mathbf{H})\mathbf{e}^{\mathrm{la}}_k + \eta\widehat{\boldsymbol{\zeta}}_k.$$

Theorem B.6 (Risk stability of Nesterov)**.** *Let the Hessian be a positive trace-class operator and assume*

$$\mathbf{H} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots), \qquad \lambda_1 = \lambda_{\max} \geqslant \lambda_2 \geqslant \cdots > 0, \qquad 0 \leqslant \beta < 1, \qquad \sigma > 0.$$

Define

$$\Delta_{\mathrm{nesterov}}(x) := 2B(1-\beta^2) + \left[4B\beta^2 + (B-1)\beta - (B+1)\right]x - (B+1)\beta(1+2\beta)x^2.$$

Then Nesterov momentum is risk stable if and only if

$$\Delta_{\mathrm{nesterov}}(\eta\lambda_{\max}) > 0 \tag{58}$$

and

$$\kappa_{\mathrm{nesterov}}(\eta,\beta) := \sum_{i\geqslant 1} \frac{\eta\lambda_i\left[1+\beta+\beta(1+2\beta)\eta\lambda_i\right]}{\Delta_{\mathrm{nesterov}}(\eta\lambda_i)} < 1. \tag{59}$$

Proof. Define

$$\mathbf{U}_k := \mathbb{E}[\mathbf{e}_k \otimes \mathbf{e}_k], \qquad \mathbf{V}_k := \mathbb{E}[\mathbf{e}^{\mathrm{la}}_k \otimes \mathbf{e}^{\mathrm{la}}_k], \qquad \mathbf{P}_k := \mathbb{E}[\mathbf{e}_k \otimes \mathbf{e}_{k-1}],$$
$$\mathbf{Q}_k := \mathbb{E}[\mathbf{e}_{k-1} \otimes \mathbf{e}_k], \qquad \mathbf{R}_k := \mathbb{E}[\mathbf{e}_{k-1} \otimes \mathbf{e}_{k-1}].$$

Expanding (57), using the vanishing cross terms, and applying Lemma A.1 gives

$$\begin{cases} \mathbf{U}_{k+1} = \mathbf{D}\mathbf{V}_k\mathbf{D} + \dfrac{\eta^2}{B}\left(\mathbf{H}\mathbf{V}_k\mathbf{H} + \mathbf{H}\mathrm{Tr}(\mathbf{H}\mathbf{V}_k) + \sigma^2\mathbf{H}\right), \\ \mathbf{V}_k = (1+\beta)^2\mathbf{U}_k - \beta(1+\beta)(\mathbf{P}_k + \mathbf{Q}_k) + \beta^2\mathbf{R}_k, \\ \mathbf{P}_{k+1} = \mathbf{D}\big((1+\beta)\mathbf{U}_k - \beta\mathbf{Q}_k\big), \\ \mathbf{Q}_{k+1} = \big((1+\beta)\mathbf{U}_k - \beta\mathbf{P}_k\big)\mathbf{D}, \\ \mathbf{R}_{k+1} = \mathbf{U}_k. \end{cases} \tag{60}$$

The renewal equation below naturally controls the look-ahead loss moment

$$\ell_k^{\mathrm{la}} := \mathrm{Tr}(\mathbf{H}\mathbf{V}_k) = \mathbb{E}\|\mathbf{e}_k^{\mathrm{la}}\|_{\mathbf{H}}^2.$$

This is equivalent to controlling the risk at the actual iterate. Define its loss moment by

$$\ell_k := \mathrm{Tr}(\mathbf{H}\mathbf{U}_k) = 2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)].$$

Then

$$\ell_k^{\mathrm{la}} \leqslant 2(1+\beta)^2\ell_k + 2\beta^2\ell_{k-1}. \tag{61}$$

Conversely, since $\mathbf{D}$ commutes with $\mathbf{H}$ and $\mathbf{H}$ is trace class, the first equation in (60) gives

$$\ell_{k+1} \leqslant C_{\eta,\beta,\mathbf{H},B}\big(\ell_k^{\mathrm{la}} + \sigma^2\big). \tag{62}$$

Therefore

$$\sup_k \ell_k^{\mathrm{la}} < \infty \quad \Longleftrightarrow \quad \sup_k \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] < \infty.$$

Let

$$u_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{U}_k\boldsymbol{\varphi}_i\rangle, \qquad p_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{P}_k\boldsymbol{\varphi}_i\rangle, \qquad r_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{R}_k\boldsymbol{\varphi}_i\rangle,$$

and

$$v_{i,k} := \langle \boldsymbol{\varphi}_i, \mathbf{V}_k\boldsymbol{\varphi}_i\rangle.$$

The cross-covariance operators satisfy

$$\mathbf{Q}_k = \mathbf{P}_k^*.$$

Therefore,

$$v_{i,k} = (1+\beta)^2 u_{i,k} - 2\beta(1+\beta)p_{i,k} + \beta^2 r_{i,k}.$$

Set

$$d_i := 1-\eta\lambda_i, \qquad a_i := d_i^2 + \frac{\eta^2}{B}\lambda_i^2, \qquad \mathbf{X}_{i,k} := (u_{i,k}, p_{i,k}, r_{i,k})^\top,$$

$$\mathbf{b} := (1,0,0)^\top, \qquad \mathbf{c} := \big((1+\beta)^2, -2\beta(1+\beta), \beta^2\big)^\top,$$

and

$$\mathbf{M}_i := \begin{pmatrix} a_i(1+\beta)^2 & -2a_i\beta(1+\beta) & a_i\beta^2 \\ d_i(1+\beta) & -d_i\beta & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Then

$$v_{i,k} = \mathbf{c}^\top\mathbf{X}_{i,k}, \qquad \ell_k^{\mathrm{la}} = \sum_{i\geqslant 1}\lambda_i\mathbf{c}^\top\mathbf{X}_{i,k},$$

and the diagonal part of (60) becomes

$$\mathbf{X}_{i,k+1} = \mathbf{M}_i\mathbf{X}_{i,k} + \frac{\eta^2}{B}\lambda_i\mathbf{b}\big(\ell_k^{\mathrm{la}} + \sigma^2\big). \tag{63}$$

The characteristic polynomial of the isolated block is

$$p_i(\mu) = \mu^3 - \left[a_i(1+\beta)^2 - d_i\beta\right]\mu^2 + a_i\beta\left[d_i(1+\beta)^2 - \beta\right]\mu - a_i d_i\beta^3.$$

The Jury stability criterion shows that all roots lie strictly inside the unit disk if and only if

$$\Delta_{\mathrm{nesterov}}(\eta\lambda_i) > 0.$$

The polynomial is a concave quadratic satisfying

$$\Delta_{\mathrm{nesterov}}(0) = 2B(1-\beta^2) > 0,$$

and it has exactly one positive root. Imposing the coordinate condition in every eigendirection is therefore equivalent to (58). If it fails, the nondegenerate label-noise forcing excites a nonstable isolated mode, and the corresponding coordinate risk is unbounded.

Assume henceforth that the local condition holds. Iterating (63), multiplying the i-th coordinate contribution by λ_i, and summing over all eigendirections gives

$$\ell_k^{\mathrm{la}} = h_k + \sum_{s=0}^{k-1} K_{k-1-s}\big(\ell_s^{\mathrm{la}} + \sigma^2\big), \tag{64}$$

where

$$h_k := \sum_{i\geqslant 1} \lambda_i \mathbf{c}^\top \mathbf{M}_i^k \mathbf{X}_{i,0}, \qquad K_m := \frac{\eta^2}{B} \sum_{i\geqslant 1} \lambda_i^2 \mathbf{c}^\top \mathbf{M}_i^m \mathbf{b}.$$

Both sequences are nonnegative: their summands are look-ahead variances generated by isolated coordinate systems. The local blocks are uniformly power bounded, and positivity of the initial coordinate covariance gives

$$0 \leqslant h_k \leqslant C \sum_{i\geqslant 1} \lambda_i (u_{i,0} + r_{i,0}) < \infty.$$

A direct calculation gives

$$\mathbf{c}^\top (\mathbf{I} - \mathbf{M}_i)^{-1} = \frac{B\big(1+\beta+\beta(1+2\beta)\eta\lambda_i, -2\beta(1+\beta), \beta^2[1+\beta(1-\eta\lambda_i)]\big)}{\eta\lambda_i \Delta_{\text{nesterov}}(\eta\lambda_i)}. \tag{65}$$

Its first entry yields

$$\sum_{m\geqslant 0} K_m = \frac{\eta^2}{B} \sum_{i\geqslant 1} \lambda_i^2 \mathbf{c}^\top (\mathbf{I} - \mathbf{M}_i)^{-1} \mathbf{b} = \kappa_{\text{nesterov}}(\eta, \beta).$$

Lemma B.1, together with (61)– (62), gives

$$\sup_{k\geqslant 0} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] < \infty \quad \Longleftrightarrow \quad \kappa_{\text{nesterov}}(\eta, \beta) < 1,$$

which completes the proof.

□

Corollary B.7 (Critical learning rate of Nesterov)**.** *There exists a unique critical learning rate satisfying*

$$\kappa_{\text{nesterov}}\big(\eta^{\text{crit}}_{\text{nesterov}}, \beta\big) = 1$$

before the local stability boundary. The Nesterov iteration is risk stable if and only if

$$0 < \eta < \eta^{\text{crit}}_{\text{nesterov}}.$$

Under the power-law and high-momentum assumptions $\lambda_i \eqsim i^{-q}, q > 1, B \gg 1, 1-\beta \ll 1,$

$$\eta^{\text{crit}}_{\text{nesterov}} \eqsim \min\{1, B^q(1-\beta)\}. \tag{66}$$

Proof. First consider the local condition. The polynomial is concave and satisfies

$$\Delta_{\text{nesterov}}(0) > 0.$$

It has a unique positive root, denoted by $x_+(B,\beta)$, and hence the local condition is equivalent to

$$\eta\lambda_{\max} < x_+(B,\beta).$$

In the high-momentum regime,

$$B \gg 1, \qquad \beta \to 1^-,$$

the root remains of constant order. Indeed,

$$x_+(B,\beta) \eqsim 1, \qquad x_+(B,1) = \frac{4B-2}{3(B+1)} \longrightarrow \frac{4}{3}.$$

Together with the power-law normalization $\lambda_{\max} \eqsim 1$, this gives the order-one local ceiling

$$\eta \lesssim 1.$$

We next consider the global renewal mass $\kappa_{\text{nesterov}}(\eta,\beta)$. Each summand is continuous and strictly increasing up to the local boundary, and the leading eigendirection diverges at that boundary. Hence Theorem B.6 gives a unique solution of the critical equation and the stated stability interval.

On every fixed subinterval strictly inside the local stability region,

$$\kappa_{\text{nesterov}}(\eta,\beta) \eqsim \frac{1}{B} \sum_{i\geqslant 1} \frac{\eta\lambda_i}{1-\beta+\eta\lambda_i}.$$

For the power-law spectrum, the spectral sum satisfies

$$\sum_{i\geqslant 1}\frac{\eta\lambda_i}{1-\beta+\eta\lambda_i}\eqsim\begin{cases}\dfrac{\eta}{1-\beta}, & \eta\lesssim 1-\beta,\\ \left(\dfrac{\eta}{1-\beta}\right)^{1/q}, & \eta\gtrsim 1-\beta.\end{cases}$$

At a global critical point, the first regime would require $\frac{\eta}{1-\beta}\eqsim B$, contradicting $\frac{\eta}{1-\beta}\lesssim 1$ when the batch size is large. Therefore the global critical point lies in the second regime, where

$$\left(\frac{\eta}{1-\beta}\right)^{1/q}\eqsim B.$$

Thus the global condition permits learning rates up to

$$\eta\eqsim B^q(1-\beta).$$

Combining this scale with the order-one local ceiling yields (66). □

C Functional scaling laws for risk dynamics

We follow the functional scaling law framework (Li et al., 2025) and show that the risk dynamics of a general second-order iteration can likewise be characterized by two macroscopic functions: signal learning S and noise forgetting $\mathcal{K}$. We then specialize the framework to Polyak and Nesterov in Appendices C.2 and C.3. For each method, we bound the signal term S, characterize the label-noise contribution N, and control the multiplicative-noise contribution M; combining the three yields the corresponding risk scaling law. The SGD result is recalled in Appendix C.4.

Throughout, $\boldsymbol{\theta}_{-1}=\boldsymbol{\theta}_0=\mathbf{0}$ for momentum methods, and $\boldsymbol{\theta}_0=\mathbf{0}$ for SGD.

C.1 Functional scaling law for second-order iterations

Consider a second-order stochastic iteration with constant learning rate η and fixed coefficients β,γ:

$$\begin{aligned}\boldsymbol{v}_k&=\boldsymbol{\theta}_k+\gamma(\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}),\\ \boldsymbol{\theta}_{k+1}&=\boldsymbol{\theta}_k+\beta(\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1})-\eta\mathbf{g}(\boldsymbol{v}_k;\mathcal{B}_k).\end{aligned}\tag{67}$$

Here $\boldsymbol{v}_k$ is the gradient evaluation point, and $\mathbf{g}(\boldsymbol{v}_k;\mathcal{B}_k)=\mathbf{H}(\boldsymbol{v}_k-\boldsymbol{\theta}^*)+\boldsymbol{\xi}_k(\boldsymbol{v}_k)$, where $\boldsymbol{\xi}_k(\boldsymbol{v}_k)$ is the mini-batch gradient noise. In the eigenbasis of $\mathbf{H}$, set $e_k(j)=\theta_k(j)-\theta_j^*$ and $\zeta_k(j)=[\boldsymbol{\xi}_k(\boldsymbol{v}_k)]_j$. Subtracting $\boldsymbol{\theta}^*$ from the update and collecting the two error terms gives

$$e_{k+1}(j)+a_je_k(j)+b_je_{k-1}(j)=-\eta\zeta_k(j),\tag{68}$$

where $a_j=\eta\lambda_j(1+\gamma)-(1+\beta)$ and $b_j=\beta-\eta\lambda_j\gamma$. Thus $\zeta_k(j)$ is exactly the gradient noise in coordinate j of the stochastic update. Let $\mathscr{F}_k$ be the history before mini-batch k is sampled. We assume that this noise satisfies

$$\mathbb{E}[\zeta_k(j)\mid\mathscr{F}_k]=0,\qquad\mathbb{E}[\zeta_k(j)^2\mid\mathscr{F}_k]\eqsim\frac{\lambda_j}{B}(\mathcal{J}_k+\sigma^2),\tag{69}$$

where $\mathcal{J}_k=\mathcal{E}(\boldsymbol{v}_k)$ is the excess risk at the gradient evaluation point, which depends on the algorithm: Polyak uses $\gamma=0$, so $\mathcal{J}_k=\mathcal{E}(\boldsymbol{\theta}_k)$, whereas Nesterov uses $\gamma=\beta$, so $\mathcal{J}_k=\mathcal{E}(\boldsymbol{\theta}_k^{\mathrm{la}})$. Lemma A.3 provides these conditional noise bounds. To derive the scaling law, we need to estimate the expected excess risk $R_k:=\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)]=\frac{1}{2}\sum_j\lambda_j\mathbb{E}[e_k(j)^2]$.

Let $e_k^{\det}(j)$ be the error sequence obtained by setting the gradient noise $\zeta_k(j)$ to zero in (68), so that

$$e_{k+1}^{\det}(j)+a_je_k^{\det}(j)+b_je_{k-1}^{\det}(j)=0,$$

with $e_{-1}^{\det}(j)=e_{-1}(j)$ and $e_0^{\det}(j)=e_0(j)$ (under our zero initialization, both initial errors equal $-\theta_j^*$). Let $G_k(j)$ satisfy the same noise-free recursion, but with $G_{-1}(j)=0$ and $G_0(j)=1$. Lemma A.5 gives

$$e_k(j)=e_k^{\det}(j)-\eta\sum_{\tau=0}^{k-1}G_\tau(j)\zeta_{k-1-\tau}(j).\tag{70}$$

Following the framework of functional scaling laws (Li et al., 2025), we introduce the signal-learning function and the noise-forgetting kernel by

$$S(k) := \frac{1}{2}\sum_{j\geqslant 1}\lambda_j (e_k^{\det}(j))^2, \qquad \mathcal{K}(\tau) := \sum_{j\geqslant 1}\lambda_j^2 G_\tau(j)^2. \tag{71}$$

Theorem C.1 (Functional scaling law for second-order iterations)**.** *Under (69), the expected excess risk satisfies*

$$R_k \eqsim S(k) + \frac{\eta^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}(\tau)\big(\sigma^2 + \mathbb{E}[\mathcal{J}_{k-1-\tau}]\big). \tag{72}$$

Proof. By (69), $\mathbb{E}[\zeta_m(j)] = 0$ and, for $m < n$, $\mathbb{E}[\zeta_m(j)\zeta_n(j)] = \mathbb{E}[\zeta_m(j)\mathbb{E}[\zeta_n(j) \mid \mathscr{F}_n]] = 0$. Squaring (70) and taking expectations gives

$$\mathbb{E}[e_k(j)^2] = (e_k^{\det}(j))^2 + \eta^2\sum_{\tau=0}^{k-1} G_\tau(j)^2\mathbb{E}[\zeta_{k-1-\tau}(j)^2]. \tag{73}$$

We multiply both sides by $\lambda_j/2$ and sum over j. The noise bound (69) then yields the result. □

The two noise contributions have different sources. Label noise is of scale σ^2, which is independent of the current error, whereas multiplicative noise depends on the risk at the gradient evaluation point. Write

$$N(k) := \frac{\eta^2\sigma^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}(\tau), \qquad M(k) := \frac{\eta^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}(\tau)\mathbb{E}[\mathcal{J}_{k-1-\tau}]. \tag{74}$$

The scaling law then takes the form

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim S(k) + N(k) + M(k). \tag{75}$$

C.2 Scaling law for Polyak (Proof of Theorem 5.1)

C.2.1 Polyak update

The Polyak update (4) is a special case of (67) with $\gamma = 0$, so $\boldsymbol{v}_k = \boldsymbol{\theta}_k$. The coefficients in Appendix C.1 are therefore $a_j = -(1 + \beta - \eta\lambda_j)$ and $b_j = \beta$, and the error recursion becomes

$$e_{k+1}(j) = (1 + \beta - \eta\lambda_j)e_k(j) - \beta e_{k-1}(j) - \eta\xi_k(j). \tag{76}$$

Here $\xi_k(j) = [\boldsymbol{\xi}_k(\boldsymbol{\theta}_k)]_j$ is the mini-batch gradient noise at $\boldsymbol{\theta}_k$, so $\zeta_k(j) = \xi_k(j)$. By Lemma A.3, the noise satisfies (69) with $\mathcal{J}_k = \mathcal{E}(\boldsymbol{\theta}_k)$. Theorem C.1 gives

$$R_k \eqsim S_{\text{polyak}}(k) + N_{\text{polyak}}(k) + M_{\text{polyak}}(k). \tag{77}$$

Here

$$\begin{aligned} S_{\text{polyak}}(k) &:= \frac{1}{2}\sum_{j\geqslant 1}\lambda_j (e_k^{\det}(j))^2, \qquad \mathcal{K}_{\text{polyak}}(\tau) := \sum_{j\geqslant 1}\lambda_j^2 G_\tau(j)^2, \\ N_{\text{polyak}}(k) &:= \frac{\eta^2\sigma^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\text{polyak}}(\tau), \\ M_{\text{polyak}}(k) &:= \frac{\eta^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\text{polyak}}(\tau)R_{k-1-\tau}, \end{aligned} \tag{78}$$

where $e_k^{\det}(j)$ and $G_k(j)$ are the noise-free error and Green function associated with (76), with the initial conditions specified in Appendix C.1. In the following we estimate the signal learning S_{polyak}, label noise N_{polyak}, and multiplicative noise M_{polyak}.

C.2.2 Classification of damping regimes

These estimates depend on $e_k^{\det}(j)$ and $G_k(j)$, which satisfy the same second-order recursion obtained from (76) by setting the noise to zero. We use the results in Appendix A.3 to bound these sequences through the roots of the characteristic equation. We first determine the damping threshold. The characteristic polynomial is $r^2 - (1+\beta-\eta\lambda_j)r + \beta$, and its discriminant factors as

$$\Delta_j = (\eta\lambda_j - (1-\sqrt{\beta})^2)(\eta\lambda_j - (1+\sqrt{\beta})^2).$$

Suppose $\eta\lambda_1 < 1$. The damping threshold is $(1-\sqrt{\beta})^2 \eqsim (1-\beta)^2$. Let $J_{\max}$ be the largest index satisfying $\eta\lambda_j \geqslant (1-\sqrt{\beta})^2$, i.e.,

$$J_{\max} := \max\{j \geqslant 1 : \eta\lambda_j \geqslant (1-\sqrt{\beta})^2\}.$$

If no such index exists, we set $J_{\max} = 1$. Since $\lambda_j \eqsim j^{-q}$,

$$J_{\max} \eqsim \max\left\{1, \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}\right\}. \tag{79}$$

The transition between the two regimes occurs at $\eta \eqsim (1-\beta)^2$:

- **Fully overdamped regime:** $\eta \lesssim (1-\beta)^2$. Every coordinate is overdamped.
- **Mixed-damping regime:** $\eta \gtrsim (1-\beta)^2$. The leading coordinates $j \leqslant J_{\max}$ are underdamped, while the tail $j > J_{\max}$ is overdamped.

Here $\eta \lesssim (1-\beta)^2$ means $\eta\lambda_1 \leqslant c_0(1-\beta)^2$ for a sufficiently small constant $c_0 > 0$, so that Lemma A.7 applies to every coordinate.

C.2.3 Estimates of signal learning $S_{\mathrm{polyak}}(k)$

We estimate $S_{\mathrm{polyak}}(k)$ by summing the weighted squared noise-free coordinate errors. We first treat the fully overdamped regime, then separate the overdamped and underdamped contributions in the mixed-damping regime.

Fully overdamped regime. In this regime, all coordinates are overdamped, and the effective learning rate is $\eta/(1-\beta)$. The following proposition shows that $S_{\mathrm{polyak}}(k)$ stays of constant order until $k \eqsim (1-\beta)/\eta$, then follows a power-law decay.

Proposition C.2 (Signal learning for Polyak in the fully overdamped regime)**.** *Under Assumptions 3.1 and 3.3, if $\eta \lesssim (1-\beta)^2$, then*

$$S_{\mathrm{polyak}}(k) \eqsim \begin{cases} 1, & k \lesssim \dfrac{1-\beta}{\eta}, \\ \left(\dfrac{1-\beta}{\eta k}\right)^s, & k \gtrsim \dfrac{1-\beta}{\eta}. \end{cases} \tag{80}$$

Proof. Since the noise-free recursion is linear and $e_{-1}^{\det}(j) = e_0^{\det}(j) = -\theta_j^*$, we write $e_k^{\det}(j) = -h_k(j)\theta_j^*$, where $h_k(j)$ describes how the initial error evolves in coordinate j. By (76), it satisfies

$$h_{k+1}(j) = (1+\beta-\eta\lambda_j)h_k(j) - \beta h_{k-1}(j), \qquad h_{-1}(j) = h_0(j) = 1.$$

Applying Lemma A.7 with $x = \eta\lambda_j$ gives

$$ce^{-C\eta\lambda_j k/(1-\beta)} \leqslant h_k(j) \leqslant Ce^{-c\eta\lambda_j k/(1-\beta)}, \qquad j \geqslant 1, \quad k \geqslant 0.$$

Since $S_{\mathrm{polyak}}(k) = \frac{1}{2}\sum_j \lambda_j(\theta_j^*)^2 h_k(j)^2$ and $\lambda_j(\theta_j^*)^2 \eqsim j^{-1-qs}$, applying Lemma A.9 with $J = 1$ to both exponential bounds yields

$$S_{\mathrm{polyak}}(k) \eqsim \left(1 + \frac{\eta k}{1-\beta}\right)^{-s}.$$

This proves the proposition. □

Mixed-damping regime. When $\eta \gtrsim (1-\beta)^2$, we enter the mixed-damping regime. In this regime, the crossover index $J_{\max}$ separates the underdamped leading coordinates $j \leqslant J_{\max}$ from the overdamped tail $j > J_{\max}$. Accordingly, we write $S_{\mathrm{polyak}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{under}}(k)$, where

$$S_{\mathrm{over}}(k) := \frac{1}{2} \sum_{j > J_{\max}} \lambda_j \big(e_k^{\mathrm{det}}(j)\big)^2,$$
$$S_{\mathrm{under}}(k) := \frac{1}{2} \sum_{1 \leqslant j \leqslant J_{\max}} \lambda_j \big(e_k^{\mathrm{det}}(j)\big)^2.$$

The next two propositions estimate these contributions separately.

Proposition C.3 (Overdamped signal learning for Polyak in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume $\eta \gtrsim (1-\beta)^2$. Then the overdamped contribution satisfies*

$$S_{\mathrm{over}}(k) \eqsim \begin{cases} \left(\dfrac{(1-\beta)^2}{\eta}\right)^s, & k \lesssim (1-\beta)^{-1}, \\ \left(\dfrac{1-\beta}{\eta k}\right)^s, & k \gtrsim (1-\beta)^{-1}. \end{cases} \tag{81}$$

Proof. Write $e_k^{\mathrm{det}}(j) = -h_k(j)\theta_j^*$. Lemma A.7 applies when $\eta\lambda_j \leqslant c_0(1-\beta)^2$. To separate these coordinates from those near the damping threshold, set

$$J_0 := \min\{j > J_{\max} : \eta\lambda_j \leqslant c_0(1-\beta)^2\}.$$

Then

$$J_0 \eqsim J_{\max} \eqsim \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}.$$

The coordinate bound and Lemma A.9 give

$$\frac{1}{2} \sum_{j \geqslant J_0} \lambda_j (\theta_j^*)^2 h_k(j)^2 \eqsim \left(J_0^q + \frac{\eta k}{(1-\beta)}\right)^{-s} \eqsim J_{\max}^{-qs}(1 + (1-\beta)k)^{-s}.$$

For $J_{\max} < j < J_0$, we have $\eta\lambda_j \eqsim (1-\beta)^2$. The positive gaps $1 - r_{j,\pm}$ have sum $(1-\beta) + \eta\lambda_j \eqsim (1-\beta)$ and product $\eta\lambda_j \eqsim (1-\beta)^2$, so both are of order $(1-\beta)$. The roots have positive sum and nonnegative product β, hence $0 \leqslant r_{j,-} \leqslant r_{j,+} \leqslant 1 - c(1-\beta)$. Equation (26) therefore gives

$$|h_k(j)| \lesssim (1 + (1-\beta)k)e^{-c(1-\beta)k}.$$

Consequently,

$$\frac{1}{2} \sum_{J_{\max} < j < J_0} \lambda_j (\theta_j^*)^2 h_k(j)^2 \lesssim J_{\max}^{-qs}(1 + (1-\beta)k)^2 e^{-2c(1-\beta)k} \lesssim J_{\max}^{-qs}(1 + (1-\beta)k)^{-s}.$$

Combining the two parts and substituting the order of $J_{\max}$ yields

$$S_{\mathrm{over}}(k) \eqsim \left(\frac{(1-\beta)^2}{\eta}\right)^s (1 + (1-\beta)k)^{-s}.$$

The two time ranges follow by comparing k with $(1-\beta)^{-1}$. This proves the proposition. □

We next estimate the underdamped contribution. The following proposition gives an exponentially decaying upper bound for this contribution.

Proposition C.4 (Underdamped signal learning for Polyak in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume $\eta \gtrsim (1-\beta)^2$ and $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ for sufficiently small $c_{\mathrm{st}} > 0$. Then there exists a nonnegative separated-underdamped contribution $\mathbb{T}_P(k)$ such that*

$$\mathbb{T}_P(k) \leqslant S_{\mathrm{under}}(k) \lesssim \mathbb{T}_P(k) + \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1 + (1-\beta)k\big)^{-s}, \qquad k \geqslant 0, \tag{82}$$

and

$$0 \leqslant \mathbb{T}_P(k) \lesssim \beta^k, \qquad k \geqslant 0. \tag{83}$$

Proof. Recall that $e_k^{\mathrm{det}}(j) = -h_k(j)\theta_j^*$, with $h_{-1}(j) = h_0(j) = 1$. For each underdamped coordinate, the characteristic roots are

$$r_{j,\pm} = \sqrt{\beta}e^{\pm i\omega_j}, \qquad \cos\omega_j = \frac{1+\beta-\eta\lambda_j}{2\sqrt{\beta}}.$$

The initial values give the common representation

$$h_k(j) = \beta^{k/2}\big[\cos(k\omega_j) + A_j\sin(k\omega_j)\big], \qquad A_j = \frac{1-\beta-\eta\lambda_j}{2\sqrt{\beta}\sin\omega_j}.$$

At the damping threshold $\eta\lambda_j = (1-\sqrt{\beta})^2$, the expression is understood by continuity. Away from this threshold, A_j is uniformly bounded, so we can bound the sine and cosine terms separately. Near the threshold, $\sin\omega_j$ can approach zero and A_j is not uniformly bounded. We therefore estimate $A_j\sin(k\omega_j)$ together, using the root formula (26), which remains valid when the roots coincide. This distinction motivates the following partition. Fix a constant $C_{\mathrm{mom}} > 2$ and set

$$\mathcal{I}_{\mathrm{cross}} := \{j \leqslant J_{\max} : \eta\lambda_j \leqslant C_{\mathrm{mom}}(1-\beta)^2\},$$
$$\mathcal{I}_{\mathrm{sep}} := \{j \leqslant J_{\max} : \eta\lambda_j > C_{\mathrm{mom}}(1-\beta)^2\}.$$

Accordingly,

$$S_{\mathrm{under}}(k) = S_{\mathrm{cross}}(k) + \mathbb{T}_P(k),$$

where

$$S_{\mathrm{cross}}(k) := \frac{1}{2}\sum_{j\in\mathcal{I}_{\mathrm{cross}}}\lambda_j(\theta_j^*)^2h_k(j)^2, \qquad \mathbb{T}_P(k) := \frac{1}{2}\sum_{j\in\mathcal{I}_{\mathrm{sep}}}\lambda_j(\theta_j^*)^2h_k(j)^2.$$

On $\mathcal{I}_{\mathrm{cross}}$, $|r_{j,-}(1-r_{j,+})|^2 = \beta\eta\lambda_j$. The divided-difference bound in (26) gives

$$|h_k(j)| \leqslant \big(1+k\sqrt{\eta\lambda_j}\big)\beta^{k/2} \lesssim \big(1+(1-\beta)k\big)\beta^{k/2},$$

including at a repeated root. To sum this coordinate bound, note that $\eta\lambda_j \leqslant C_{\mathrm{mom}}(1-\beta)^2$ and $\lambda_j \asymp j^{-q}$ imply

$$j \gtrsim \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}$$

on $\mathcal{I}_{\mathrm{cross}}$. Thus the power-law weights give

$$\sum_{j\in\mathcal{I}_{\mathrm{cross}}}\lambda_j(\theta_j^*)^2 \lesssim \sum_{j\gtrsim(\eta/(1-\beta)^2)^{1/q}} j^{-1-qs} \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s.$$

Substituting both bounds into the definition of S_{cross} yields

$$S_{\mathrm{cross}}(k) \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s\big(1+(1-\beta)k\big)^2\beta^k.$$

Since $\beta^k \leqslant e^{-(1-\beta)k}$ and

$$\sup_{t\geqslant 0}(1+t)^{s+2}e^{-t} < \infty,$$

we obtain

$$\big(1+(1-\beta)k\big)^2\beta^k \lesssim \big(1+(1-\beta)k\big)^{-s}.$$

Therefore,

$$S_{\mathrm{cross}}(k) \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s\big(1+(1-\beta)k\big)^{-s} \asymp S_{\mathrm{over}}(k), \tag{84}$$

where the last equivalence follows from Proposition C.3.

It remains to control the separated underdamped modes. For $j \in \mathcal{I}_{\mathrm{sep}}$,

$$4\beta\sin^2\omega_j = \big(\eta\lambda_j - (1-\sqrt{\beta})^2\big)\big((1+\sqrt{\beta})^2 - \eta\lambda_j\big) \gtrsim \eta\lambda_j,$$

and

$$(1-\beta-\eta\lambda_j)^2 \lesssim \eta\lambda_j.$$

Hence $A_j^2 \lesssim 1$, and therefore $h_k(j)^2 \lesssim \beta^k$ on $\mathcal{I}_{\mathrm{sep}}$. Since $\sum_{j\geqslant 1}\lambda_j(\theta_j^*)^2 \lesssim 1$,

$$0 \leqslant \mathbb{T}_P(k) \lesssim \beta^k\sum_{j\in\mathcal{I}_{\mathrm{sep}}}\lambda_j(\theta_j^*)^2 \lesssim \beta^k,$$

which proves (83). Since $S_{\mathrm{under}}(k) = S_{\mathrm{cross}}(k) + \mathbb{T}_P(k)$, the bound (84) gives (82). □

Corollary C.5 (Total signal learning for Polyak in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume* $\eta \gtrsim (1-\beta)^2$ *and* $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ *for sufficiently small* $c_{\mathrm{st}} > 0$*. Then*

$$S_{\mathrm{polyak}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{under}}(k) \eqsim \mathbb{T}_P(k) + \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s}, \qquad k \geqslant 0. \tag{85}$$

Proof. By Proposition C.4,

$$S_{\mathrm{under}}(k) = \mathbb{T}_P(k) + S_{\mathrm{cross}}(k), \qquad S_{\mathrm{cross}}(k) \lesssim S_{\mathrm{over}}(k).$$

Since $S_{\mathrm{over}}(k)$, $S_{\mathrm{cross}}(k)$, and $\mathbb{T}_P(k)$ are nonnegative,

$$S_{\mathrm{over}}(k) + \mathbb{T}_P(k) \leqslant S_{\mathrm{polyak}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{cross}}(k) + \mathbb{T}_P(k) \lesssim S_{\mathrm{over}}(k) + \mathbb{T}_P(k).$$

Proposition C.3 gives

$$S_{\mathrm{over}}(k) \eqsim \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s},$$

which proves the corollary. □

C.2.4 Estimates of label noise

Recall from (78) that

$$N_{\mathrm{polyak}}(k) = \frac{\eta^2\sigma^2}{B}\sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{polyak}}(\tau), \qquad \mathcal{K}_{\mathrm{polyak}}(\tau) = \sum_{j\geqslant 1} \lambda_j^2 \big(G_\tau^{\mathrm{polyak}}(j)\big)^2.$$

Here $G_\tau^{\mathrm{polyak}}(j)$ is the Green function associated with (76). The following proposition establishes a property of the forgetting kernel that we use to estimate $N_{\mathrm{polyak}}(k)$.

Proposition C.6 (A property of the forgetting kernel for Polyak)**.** *Assume* $\eta\lambda_1 \leqslant c_{\mathrm{st}} < 1$*. There is a universal constant* $C > 0$ *such that, whenever*

$$k \geqslant C \max\left\{\frac{1}{1-\beta}, \frac{1-\beta}{\eta}\right\},$$

one has

$$\sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{polyak}}(\tau) \eqsim \frac{1}{\eta(1-\beta)}. \tag{86}$$

Proof. The learning-rate condition ensures that both characteristic roots have modulus less than one. Applying Lemma A.6 with $a_j = -(1+\beta-\eta\lambda_j)$ and $b_j = \beta$ from (76) gives

$$\sum_{\tau=0}^{\infty} \big(G_\tau^{\mathrm{polyak}}(j)\big)^2 = \frac{1+\beta}{(1-\beta)\eta\lambda_j\,(2(1+\beta)-\eta\lambda_j)}. \tag{87}$$

Consequently,

$$\sum_{\tau=0}^{\infty} \mathcal{K}_{\mathrm{polyak}}(\tau) = \frac{1}{\eta}\sum_{j\geqslant 1} \frac{(1+\beta)\lambda_j}{(1-\beta)\,(2(1+\beta)-\eta\lambda_j)} \eqsim \frac{1}{\eta(1-\beta)}. \tag{88}$$

The last equivalence uses the small-step bound $\eta\lambda_1 \leqslant c_{\mathrm{st}} < 1$ and $\sum_{j\geqslant 1}\lambda_j \eqsim 1$. This gives the upper bound in (86). For the lower bound, choose a fixed J_* such that

$$\sum_{j=1}^{J_*} \lambda_j \geqslant \frac{1}{2}\sum_{j\geqslant 1}\lambda_j.$$

For $j \leqslant J_*$, set $\nu_j := \min\{1-\beta, \eta\lambda_j/(1-\beta)\}$ and $z_j := 1+\varepsilon\nu_j$, where $\varepsilon > 0$ is a sufficiently small absolute constant. Since $\nu_j \leqslant 1-\beta$ and $(1-\beta)\nu_j \leqslant \eta\lambda_j$, we have $\sqrt{z_j}\,|r_{j,\pm}| < 1$, $1-\beta z_j \eqsim 1-\beta$, and $(1+\beta z_j)^2 - a_j^2 z_j \eqsim \eta\lambda_j$. The weighted energy in Lemma A.6 is therefore comparable to the

total energy in (87). Since $\log z_j \asymp \nu_j$, the tail bound in that lemma gives constants $C_G, c_G > 0$ such that

$$\sum_{\tau=k}^{\infty} \left(G_\tau^{\mathrm{polyak}}(j)\right)^2 \leqslant C_G \exp\left\{-c_G k \min\left(1-\beta, \frac{\eta\lambda_j}{1-\beta}\right)\right\} \sum_{\tau=0}^{\infty} \left(G_\tau^{\mathrm{polyak}}(j)\right)^2. \tag{89}$$

For $j \leqslant J_*$,

$$\min\left(1-\beta, \frac{\eta\lambda_j}{1-\beta}\right) \gtrsim \min\left(1-\beta, \frac{\eta}{1-\beta}\right).$$

The time bound makes the tail in (89) at most a fixed fraction of the full energy. Sum over $j \leqslant J_*$ and use (87) to obtain (86). □

Multiplying the kernel sum by $\eta^2\sigma^2/B$ gives the label-noise contribution.

Corollary C.7 (Label noise for Polyak)**.** *Assume $\eta\lambda_1 \leqslant c_{\mathrm{st}} < 1$. There is a constant $C > 0$ such that, if*

$$k \geqslant C \max\left\{\frac{1}{1-\beta}, \frac{1-\beta}{\eta}\right\},$$

then

$$N_{\mathrm{polyak}}(k) \asymp \frac{\sigma^2\eta}{B(1-\beta)}. \tag{90}$$

Proof. This follows immediately from the definition of $N_{\mathrm{polyak}}(k)$ and Proposition C.6. □

C.2.5 Control of multiplicative noise

Recall from (78) that

$$M_{\mathrm{polyak}}(k) = \frac{\eta^2}{B} \sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{polyak}}(\tau) R_{k-1-\tau}.$$

In this appendix, we control this term and show that

$$M_{\mathrm{polyak}}(k) \lesssim N_{\mathrm{polyak}}(k).$$

We first establish a uniform risk bound using Lemma A.10 and the kernel estimate established above. The following proposition gives the resulting control of multiplicative noise.

Proposition C.8 (Control of multiplicative noise for Polyak)**.** *Suppose $\sigma > 0$ and $\eta \leqslant c\min\{1, B(1-\beta)\}$ for sufficiently small $c > 0$. There is a constant $M < \infty$, independent of k, η, β, B, such that*

$$\sup_{k \geqslant 0} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \leqslant M.$$

Moreover,

$$M_{\mathrm{polyak}}(k) \lesssim N_{\mathrm{polyak}}(k).$$

Proof. The signal learning bounds above give $\sup_k S_{\mathrm{polyak}}(k) \leqslant S_*$, uniformly over the admissible parameters. By (88),

$$\kappa := \frac{\eta^2}{B} \sum_{\tau \geqslant 0} \mathcal{K}_{\mathrm{polyak}}(\tau) \lesssim \frac{\eta}{B(1-\beta)} \leqslant c.$$

Let $A \geqslant 1$ satisfy

$$R_k \leqslant A\big(S_{\mathrm{polyak}}(k) + N_{\mathrm{polyak}}(k) + M_{\mathrm{polyak}}(k)\big).$$

Set $x_k = y_k = R_k$, $u_k = AS_{\mathrm{polyak}}(k)$, and $w_\tau = A\eta^2\mathcal{K}_{\mathrm{polyak}}(\tau)/B$. Take $v = \sigma^2$, $L = 1$, and $\iota = 0$. The risk inequality above has the form required by Lemma A.10, with $y_m \leqslant Lx_{m+\iota}$ and $U := \sup_k u_k \leqslant AS_*$. By these definitions,

$$L\sum_{\tau \geqslant 0} w_\tau = A\frac{\eta^2}{B} \sum_{\tau \geqslant 0} \mathcal{K}_{\mathrm{polyak}}(\tau) = A\kappa.$$

Since $\kappa \lesssim c$, choosing c sufficiently small ensures $A\kappa \leqslant 1/2$, as required by the lemma. Applying Lemma A.10 now gives

$$\sup_k R_k = \sup_k x_k \leqslant 2U + \frac{v}{L} \leqslant 2AS_* + \sigma^2 =: M.$$

Hence

$$M_{\mathrm{polyak}}(k) \leqslant \frac{\eta^2 M}{B} \sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{polyak}}(\tau) = \frac{M}{\sigma^2} N_{\mathrm{polyak}}(k).$$

□

The multiplicative-noise bound gives a simplified form of the scaling law.

Corollary C.9 (Scaling law of simplified form for Polyak)**.** *Under the assumptions of Proposition C.8,*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim S_{\mathrm{polyak}}(k) + N_{\mathrm{polyak}}(k). \tag{91}$$

Proof. By Proposition C.8, $M_{\mathrm{polyak}}(k) \lesssim N_{\mathrm{polyak}}(k)$. Substituting this bound into (77) and using nonnegativity proves the claim. □

C.2.6 Completion of the proof

We now combine the signal learning and noise estimates to establish the following scaling laws in the two damping regimes. The constant $c > 0$ is sufficiently small to ensure $\eta\lambda_1 \leqslant c_{\mathrm{st}}$, and $C > 0$ is sufficiently large to satisfy the time conditions of the preceding estimates.

Fully overdamped regime. Here the overdamped signal learning estimate combines directly with the label noise bound, since the multiplicative noise is controlled by the latter.

Theorem C.10 (Scaling law for Polyak in the fully overdamped regime)**.** *Under Assumptions 3.1 and 3.3, suppose $\sigma > 0$, assume the Polyak stability condition $\eta \leqslant c\min\{1, B(1-\beta)\}$, and consider the fully overdamped regime $\eta \lesssim (1-\beta)^2$. For k satisfying*

$$k \geqslant C\frac{1-\beta}{\eta},$$

the expected excess risk satisfies

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \left(\frac{1-\beta}{\eta k}\right)^s + \frac{\sigma^2\eta}{B(1-\beta)}.$$

Proof. By Corollary C.9, $R_k \eqsim S_{\mathrm{polyak}}(k) + N_{\mathrm{polyak}}(k)$. Proposition C.2 gives $S_{\mathrm{polyak}}(k) \eqsim (1 + \eta k/(1-\beta))^{-s}$. Since $k \geqslant C(1-\beta)/\eta$, this is comparable to $((1-\beta)/(\eta k))^s$. The stated learning-rate and time conditions also allow us to apply Corollary C.7. Thus $N_{\mathrm{polyak}}(k) \eqsim \sigma^2\eta/[B(1-\beta)]$. Substituting these estimates proves the theorem. □

Mixed-damping regime. Here we retain the underdamped transient and combine the overdamped tail estimate with the label noise bound.

Theorem C.11 (Scaling law for Polyak in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, suppose $\sigma > 0$, assume the Polyak stability condition $\eta \leqslant c\min\{1, B(1-\beta)\}$, and consider the mixed-damping regime $\eta \gtrsim (1-\beta)^2$. For k satisfying*

$$k \geqslant \frac{C}{1-\beta},$$

the expected excess risk satisfies

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \mathbb{T}_P(k) + \left(\frac{1-\beta}{\eta k}\right)^s + \frac{\sigma^2\eta}{B(1-\beta)}. \tag{92}$$

Here $\mathbb{T}_P(k)$ is the separated underdamped transient defined in Proposition C.4 and satisfies

$$0 \leqslant \mathbb{T}_P(k) \lesssim \beta^k.$$

Proof. By Corollary C.9, $R_k \eqsim S_{\mathrm{polyak}}(k) + N_{\mathrm{polyak}}(k)$. Corollary C.5 gives

$$S_{\mathrm{polyak}}(k) \eqsim \mathbb{T}_P(k) + \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s}.$$

Since $k \geqslant C/(1-\beta)$, the second term is comparable to $((1-\beta)/(\eta k))^s$. The learning-rate and time conditions also imply the assumptions of Corollary C.7, which supplies the stated label noise term. Combining these estimates proves the theorem. □

C.3 Scaling law for Nesterov (Proof of Theorem 5.2)

C.3.1 Nesterov update

The Nesterov update is a special case of (67) with $\gamma = \beta$, so $\boldsymbol{v}_k = \boldsymbol{\theta}_k^{\mathrm{la}} = (1+\beta)\boldsymbol{\theta}_k - \beta\boldsymbol{\theta}_{k-1}$. The coefficients in Appendix C.1 are therefore $a_j = -(1+\beta)(1-\eta\lambda_j)$ and $b_j = \beta(1-\eta\lambda_j)$, and the error recursion becomes

$$e_{k+1}(j) = (1+\beta)(1-\eta\lambda_j)e_k(j) - \beta(1-\eta\lambda_j)e_{k-1}(j) - \eta\xi_k^{\mathrm{la}}(j). \tag{93}$$

Here $\xi_k^{\mathrm{la}}(j) = [\boldsymbol{\xi}_k(\boldsymbol{\theta}_k^{\mathrm{la}})]_j$ is the mini-batch gradient noise at $\boldsymbol{\theta}_k^{\mathrm{la}}$, so $\zeta_k(j) = \xi_k^{\mathrm{la}}(j)$. Let $\mathcal{L}_k := \mathcal{E}(\boldsymbol{\theta}_k^{\mathrm{la}})$ be the excess risk at that point. By Lemma A.3, the noise satisfies (69) with $\mathcal{J}_k = \mathcal{L}_k$. Theorem C.1 gives

$$R_k \eqsim S_{\mathrm{nesterov}}(k) + N_{\mathrm{nesterov}}(k) + M_{\mathrm{nesterov}}(k). \tag{94}$$

Here

$$\begin{aligned}
S_{\mathrm{nesterov}}(k) &:= \frac{1}{2}\sum_{j\geqslant 1}\lambda_j (e_k^{\mathrm{det}}(j))^2, \qquad \mathcal{K}_{\mathrm{nesterov}}(\tau) := \sum_{j\geqslant 1}\lambda_j^2 G_\tau(j)^2,\\
N_{\mathrm{nesterov}}(k) &:= \frac{\eta^2\sigma^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\mathrm{nesterov}}(\tau),\\
M_{\mathrm{nesterov}}(k) &:= \frac{\eta^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\mathrm{nesterov}}(\tau)\mathbb{E}[\mathcal{L}_{k-1-\tau}],
\end{aligned} \tag{95}$$

where $e_k^{\mathrm{det}}(j)$ and $G_k(j)$ are the noise-free error and Green function associated with (93), with the initial conditions specified in Appendix C.1. In the following we estimate the signal learning S_{nesterov}, label noise N_{nesterov}, and multiplicative noise M_{nesterov}.

C.3.2 Classification of damping regimes

These estimates depend on $e_k^{\mathrm{det}}(j)$ and $G_k(j)$, which satisfy the same second-order recursion obtained from (93) by setting the noise to zero. We use the results in Appendix A.3 to bound these sequences through the roots of the characteristic equation. We first determine the damping threshold. The characteristic polynomial is $r^2 - (1+\beta)(1-\eta\lambda_j)r + \beta(1-\eta\lambda_j)$, with discriminant

$$\Delta_j = (1-\eta\lambda_j)\big((1+\beta)^2(1-\eta\lambda_j) - 4\beta\big).$$

Suppose $\eta\lambda_1 < 1$. The damping threshold is $(1-\beta)^2/(1+\beta)^2 \eqsim (1-\beta)^2$. Let $J_{\max}$ be the largest index satisfying $\eta\lambda_j \geqslant (1-\beta)^2/(1+\beta)^2$, i.e.,

$$J_{\max} := \max\{j \geqslant 1 : \eta\lambda_j \geqslant (1-\beta)^2/(1+\beta)^2\}.$$

If no such index exists, we set $J_{\max} = 1$. Since $\lambda_j \eqsim j^{-q}$,

$$J_{\max} \eqsim \max\left\{1, \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}\right\}. \tag{96}$$

The transition between the two regimes occurs at $\eta \eqsim (1-\beta)^2$:

- **Fully overdamped regime:** $\eta \lesssim (1-\beta)^2$. Every coordinate is overdamped.
- **Mixed-damping regime:** $\eta \gtrsim (1-\beta)^2$. The leading coordinates $j \leqslant J_{\max}$ are underdamped, while the tail $j > J_{\max}$ is overdamped.

Here $\eta \lesssim (1-\beta)^2$ means $\eta\lambda_1 \leqslant c_0(1-\beta)^2$ for a sufficiently small constant $c_0 > 0$, so that Lemma A.8 applies to every coordinate.

C.3.3 ESTIMATES OF SIGNAL LEARNING $S_{\mathrm{nesterov}}(k)$

We estimate $S_{\mathrm{nesterov}}(k)$ by summing the weighted squared noise-free coordinate errors. We first treat the fully overdamped regime, then separate the overdamped and underdamped contributions in the mixed-damping regime.

Fully overdamped regime. In this regime, all coordinates are overdamped, and the effective learning rate is $\eta/(1-\beta)$. The following proposition shows that $S_{\mathrm{nesterov}}(k)$ stays of constant order until $k \eqsim (1-\beta)/\eta$, then follows a power-law decay.

Proposition C.12 (Signal learning for Nesterov in the fully overdamped regime)**.** *Under Assumptions 3.1 and 3.3, if $\eta \lesssim (1-\beta)^2$, then*

$$S_{\mathrm{nesterov}}(k) \eqsim \begin{cases} 1, & k \lesssim \dfrac{1-\beta}{\eta}, \\ \left(\dfrac{1-\beta}{\eta k}\right)^s, & k \gtrsim \dfrac{1-\beta}{\eta}. \end{cases} \tag{97}$$

Proof. Since the noise-free recursion is linear and $e_{-1}^{\mathrm{det}}(j) = e_0^{\mathrm{det}}(j) = -\theta_j^*$, we write $e_k^{\mathrm{det}}(j) = -h_k(j)\theta_j^*$, where $h_k(j)$ describes how the initial error evolves in coordinate j. By (93), it satisfies

$$h_{k+1}(j) = (1+\beta)(1-\eta\lambda_j)h_k(j) - \beta(1-\eta\lambda_j)h_{k-1}(j), \qquad h_{-1}(j) = h_0(j) = 1.$$

Applying Lemma A.8 with $x = \eta\lambda_j$ gives

$$ce^{-C\eta\lambda_j k/(1-\beta)} \leqslant h_k(j) \leqslant Ce^{-c\eta\lambda_j k/(1-\beta)}, \qquad j \geqslant 1, \quad k \geqslant 0.$$

Since $S_{\mathrm{nesterov}}(k) = \frac{1}{2}\sum_j \lambda_j(\theta_j^*)^2 h_k(j)^2$ and $\lambda_j(\theta_j^*)^2 \eqsim j^{-1-qs}$, applying Lemma A.9 with $J = 1$ to both exponential bounds yields

$$S_{\mathrm{nesterov}}(k) \eqsim \left(1 + \frac{\eta k}{1-\beta}\right)^{-s}.$$

This proves the proposition. □

Mixed-damping regime. When $\eta \gtrsim (1-\beta)^2$, we enter the mixed-damping regime. In this regime, the crossover index $J_{\max}$ separates the underdamped leading coordinates $j \leqslant J_{\max}$ from the overdamped tail $j > J_{\max}$. Accordingly, we write $S_{\mathrm{nesterov}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{under}}(k)$, where

$$S_{\mathrm{over}}(k) := \frac{1}{2}\sum_{j > J_{\max}} \lambda_j \big(e_k^{\mathrm{det}}(j)\big)^2,$$

$$S_{\mathrm{under}}(k) := \frac{1}{2}\sum_{1 \leqslant j \leqslant J_{\max}} \lambda_j \big(e_k^{\mathrm{det}}(j)\big)^2.$$

The next two propositions estimate these contributions separately.

Proposition C.13 (Overdamped signal learning for Nesterov in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume $\eta \gtrsim (1-\beta)^2$. Then the overdamped contribution satisfies*

$$S_{\mathrm{over}}(k) \eqsim \begin{cases} \left(\dfrac{(1-\beta)^2}{\eta}\right)^s, & k \lesssim (1-\beta)^{-1}, \\ \left(\dfrac{1-\beta}{\eta k}\right)^s, & k \gtrsim (1-\beta)^{-1}. \end{cases} \tag{98}$$

Proof. Write $e_k^{\mathrm{det}}(j) = -h_k(j)\theta_j^*$. Lemma A.8 applies when $\eta\lambda_j \leqslant c_0(1-\beta)^2$. To separate these coordinates from those near the damping threshold, set

$$J_0 := \min\{j > J_{\max} : \eta\lambda_j \leqslant c_0(1-\beta)^2\}.$$

Then

$$J_0 \eqsim J_{\max} \eqsim \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}.$$

The coordinate bound and Lemma A.9 give

$$\frac{1}{2}\sum_{j\geqslant J_0}\lambda_j(\theta_j^*)^2h_k(j)^2 \eqsim \left(J_0^q+\frac{\eta k}{(1-\beta)}\right)^{-s} \eqsim J_{\max}^{-qs}(1+(1-\beta)k)^{-s}.$$

For $J_{\max}<j<J_0$, we have $\eta\lambda_j\eqsim(1-\beta)^2$. The positive gaps $1-r_{j,\pm}$ have sum $(1-\beta)+(1+\beta)\eta\lambda_j\eqsim(1-\beta)$ and product $\eta\lambda_j\eqsim(1-\beta)^2$, so both are of order $(1-\beta)$. The roots have positive sum and nonnegative product $\beta(1-\eta\lambda_j)$, hence $0\leqslant r_{j,-}\leqslant r_{j,+}\leqslant 1-c(1-\beta)$. Equation (26) therefore gives

$$|h_k(j)|\lesssim(1+(1-\beta)k)e^{-c(1-\beta)k}.$$

Consequently,

$$\frac{1}{2}\sum_{J_{\max}<j<J_0}\lambda_j(\theta_j^*)^2h_k(j)^2\lesssim J_{\max}^{-qs}(1+(1-\beta)k)^2e^{-2c(1-\beta)k}\lesssim J_{\max}^{-qs}(1+(1-\beta)k)^{-s}.$$

Combining the two parts and substituting the order of $J_{\max}$ yields

$$S_{\mathrm{over}}(k)\eqsim\left(\frac{(1-\beta)^2}{\eta}\right)^s(1+(1-\beta)k)^{-s}.$$

The two time ranges follow by comparing k with $(1-\beta)^{-1}$. This proves the proposition. □

We next estimate the underdamped contribution. The following proposition gives an exponentially decaying upper bound for this contribution.

Proposition C.14 (Underdamped signal learning for Nesterov in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume $\eta\gtrsim(1-\beta)^2$ and $\eta\lambda_1\leqslant c_{\mathrm{st}}$ for sufficiently small $c_{\mathrm{st}}>0$. Then there exists a nonnegative separated-underdamped contribution $\mathbb{T}_N(k)$ such that*

$$\mathbb{T}_N(k)\leqslant S_{\mathrm{under}}(k)\lesssim\mathbb{T}_N(k)+\left(\frac{(1-\beta)^2}{\eta}\right)^s\big(1+(1-\beta)k\big)^{-s},\qquad k\geqslant 0,\tag{99}$$

and

$$0\leqslant\mathbb{T}_N(k)\lesssim\beta^k(1+\eta k)^{-s},\qquad k\geqslant 0.\tag{100}$$

In particular, for $k\geqslant 1$,

$$0\leqslant\mathbb{T}_N(k)\lesssim\min\{1,(\eta k)^{-s}\}\beta^k.\tag{101}$$

Proof. Let

$$u_c:=\frac{(1-\beta)^2}{(1+\beta)^2},\qquad p_j:=\beta(1-\eta\lambda_j).$$

Recall that $e_k^{\det}(j)=-h_k(j)\theta_j^*$, with $h_{-1}(j)=h_0(j)=1$. For each underdamped coordinate, the characteristic roots are

$$r_{j,\pm}=\sqrt{p_j}e^{\pm i\omega_j},\qquad\cos\omega_j=\frac{(1+\beta)\sqrt{1-\eta\lambda_j}}{2\sqrt{\beta}}.$$

The initial values give the common representation

$$h_k(j)=p_j^{k/2}\big[\cos(k\omega_j)+A_j\sin(k\omega_j)\big],\qquad A_j=\frac{(1-\beta)\sqrt{1-\eta\lambda_j}}{2\sqrt{\beta}\sin\omega_j}.$$

At the damping threshold $\eta\lambda_j=u_c$, the expression is understood by continuity. Away from this threshold, A_j is uniformly bounded, so we can bound the sine and cosine terms separately. Near the threshold, $\sin\omega_j$ can approach zero and A_j is not uniformly bounded. We therefore estimate $A_j\sin(k\omega_j)$ together, using the root formula (26), which remains valid when the roots coincide. This distinction motivates the following partition. Fix a constant $C_{\mathrm{mom}}>2$ and set

$$\mathcal{I}_{\mathrm{cross}}:=\{j\leqslant J_{\max}:\eta\lambda_j\leqslant C_{\mathrm{mom}}(1-\beta)^2\},$$
$$\mathcal{I}_{\mathrm{sep}}:=\{j\leqslant J_{\max}:\eta\lambda_j>C_{\mathrm{mom}}(1-\beta)^2\}.$$

Accordingly,

$$S_{\text{under}}(k) = S_{\text{cross}}(k) + \mathbb{T}_N(k),$$

where

$$S_{\text{cross}}(k) := \frac{1}{2}\sum_{j\in\mathcal{I}_{\text{cross}}} \lambda_j(\theta_j^*)^2 h_k(j)^2, \qquad \mathbb{T}_N(k) := \frac{1}{2}\sum_{j\in\mathcal{I}_{\text{sep}}} \lambda_j(\theta_j^*)^2 h_k(j)^2.$$

On $\mathcal{I}_{\text{cross}}$, $|r_{j,-}(1-r_{j,+})|^2 = p_j\eta\lambda_j$, so (26) gives

$$|h_k(j)| \leqslant \big(1 + k\sqrt{\eta\lambda_j}\big)p_j^{k/2} \lesssim \big(1+(1-\beta)k\big)p_j^{k/2}.$$

Since $\eta\lambda_j \geqslant u_c$,

$$p_j \leqslant p_{\text{nesterov}}, \qquad p_{\text{nesterov}} := \beta(1-u_c) = \frac{4\beta^2}{(1+\beta)^2}.$$

To sum the coordinate bound, note that $\eta\lambda_j \leqslant C_{\text{mom}}(1-\beta)^2$ and $\lambda_j \eqsim j^{-q}$ imply

$$j \gtrsim \left(\frac{\eta}{(1-\beta)^2}\right)^{1/q}$$

on $\mathcal{I}_{\text{cross}}$. Thus

$$\sum_{j\in\mathcal{I}_{\text{cross}}} \lambda_j(\theta_j^*)^2 \lesssim \sum_{j\gtrsim(\eta/(1-\beta)^2)^{1/q}} j^{-1-qs} \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s.$$

Substituting the coordinate and weight bounds into the definition of S_{cross} gives

$$S_{\text{cross}}(k) \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^2 p_{\text{nesterov}}^k.$$

Moreover,

$$1 - p_{\text{nesterov}} = \frac{(1-\beta)(1+3\beta)}{(1+\beta)^2} \eqsim 1-\beta.$$

Hence $p_{\text{nesterov}}^k \leqslant e^{-c(1-\beta)k}$ for an absolute constant $c>0$. Since

$$\sup_{t\geqslant 0}(1+t)^{s+2}e^{-ct} < \infty,$$

we obtain

$$\big(1+(1-\beta)k\big)^2 p_{\text{nesterov}}^k \lesssim \big(1+(1-\beta)k\big)^{-s}.$$

Therefore,

$$S_{\text{cross}}(k) \lesssim \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s} \eqsim S_{\text{over}}(k), \tag{102}$$

where the last equivalence follows from Proposition C.13.

It remains to control the separated underdamped modes. For $j\in\mathcal{I}_{\text{sep}}$,

$$4\beta\sin^2\omega_j = (1+\beta)^2(\eta\lambda_j - u_c), \qquad A_j^2 \lesssim \frac{(1-\beta)^2}{\eta\lambda_j} \lesssim 1.$$

Thus $h_k(j)^2 \lesssim p_j^k = \beta^k(1-\eta\lambda_j)^k$ on $\mathcal{I}_{\text{sep}}$. Substituting this bound into $\mathbb{T}_N(k)$ and using $(1-\eta\lambda_j)^k \leqslant e^{-\eta k\lambda_j}$, we obtain

$$0 \leqslant \mathbb{T}_N(k) \lesssim \beta^k \sum_{j\in\mathcal{I}_{\text{sep}}} j^{-1-qs} e^{-c\eta k j^{-q}} \lesssim \beta^k(1+\eta k)^{-s}, \qquad k\geqslant 0,$$

where the last step is the all-time upper bound (38). This proves (100); for $k\geqslant 1$, the equivalence

$$(1+\eta k)^{-s} \eqsim \min\{1, (\eta k)^{-s}\}$$

gives (101). Since $S_{\text{under}}(k) = S_{\text{cross}}(k) + \mathbb{T}_N(k)$, the bound (102) gives (99). □

Corollary C.15 (Total signal learning for Nesterov in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, assume* $\eta \gtrsim (1-\beta)^2$ *and* $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ *for sufficiently small* $c_{\mathrm{st}} > 0$*. Then*

$$S_{\mathrm{nesterov}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{under}}(k) \eqsim \mathbb{T}_N(k) + \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s}, \qquad k \geqslant 0. \tag{103}$$

Proof. By Proposition C.14,

$$S_{\mathrm{under}}(k) = \mathbb{T}_N(k) + S_{\mathrm{cross}}(k), \qquad S_{\mathrm{cross}}(k) \lesssim S_{\mathrm{over}}(k).$$

Since $S_{\mathrm{over}}(k)$, $S_{\mathrm{cross}}(k)$, and $\mathbb{T}_N(k)$ are nonnegative,

$$S_{\mathrm{over}}(k) + \mathbb{T}_N(k) \leqslant S_{\mathrm{nesterov}}(k) = S_{\mathrm{over}}(k) + S_{\mathrm{cross}}(k) + \mathbb{T}_N(k) \lesssim S_{\mathrm{over}}(k) + \mathbb{T}_N(k).$$

Proposition C.13 gives

$$S_{\mathrm{over}}(k) \eqsim \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1+(1-\beta)k\big)^{-s},$$

which proves the corollary. □

C.3.4 Estimates of label noise

Recall from (95) that

$$N_{\mathrm{nesterov}}(k) = \frac{\eta^2\sigma^2}{B}\sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{nesterov}}(\tau), \qquad \mathcal{K}_{\mathrm{nesterov}}(\tau) = \sum_{j\geqslant 1} \lambda_j^2 \big(G_\tau^{\mathrm{nesterov}}(j)\big)^2.$$

Here $G_\tau^{\mathrm{nesterov}}(j)$ is the Green function associated with (93). The following proposition establishes a property of the forgetting kernel that we use to estimate $N_{\mathrm{nesterov}}(k)$.

Proposition C.16 (A property of the forgetting kernel for Nesterov)**.** *Assume* $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ *for sufficiently small universal* $c_{\mathrm{st}} > 0$*. There is a universal constant* $C > 0$ *such that, whenever*

$$k \geqslant C \max\left\{\frac{1}{1-\beta}, \frac{1-\beta}{\eta}\right\},$$

one has

$$\sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{nesterov}}(\tau) \eqsim \frac{1}{\eta^2}\min\left\{\left(\frac{\eta}{1-\beta}\right)^{1/q}, \frac{\eta}{1-\beta}\right\}. \tag{104}$$

Proof. The learning-rate condition ensures that both characteristic roots have modulus less than one. Applying Lemma A.6 with $a_j = -(1+\beta)(1-\eta\lambda_j)$ and $b_j = \beta(1-\eta\lambda_j)$ from (93) gives

$$\sum_{\tau=0}^{\infty} \big(G_\tau^{\mathrm{nesterov}}(j)\big)^2 = \frac{1+\beta(1-\eta\lambda_j)}{\big(1-\beta(1-\eta\lambda_j)\big)\,\eta\lambda_j\,[1+(1+2\beta)(1-\eta\lambda_j)]}. \tag{105}$$

If $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ for sufficiently small universal c_{st}, the numerator and final bracket are bounded above and below by positive constants. Also, $1-\beta(1-\eta\lambda_j) \eqsim 1-\beta+\eta\lambda_j$. Hence

$$\sum_{\tau=0}^{\infty} \big(G_\tau^{\mathrm{nesterov}}(j)\big)^2 \eqsim \frac{1}{\eta\lambda_j(1-\beta+\eta\lambda_j)}. \tag{106}$$

Write $\eta_\beta = \eta/(1-\beta)$. Then

$$\eta^2 \sum_{\tau=0}^{\infty} \mathcal{K}_{\mathrm{nesterov}}(\tau) \eqsim \eta \sum_{j\geqslant 1} \frac{\lambda_j}{1-\beta+\eta\lambda_j} = \sum_{j\geqslant 1} \frac{\lambda_j}{\lambda_j + \eta_\beta^{-1}}. \tag{107}$$

If $\eta_\beta \leqslant 1$, then $\lambda_j \leqslant O(1)$ and $\frac{\lambda_j}{\lambda_j+\eta_\beta^{-1}} \eqsim \eta_\beta\lambda_j$, so

$$\sum_{j\geqslant 1} \frac{\lambda_j}{\lambda_j+\eta_\beta^{-1}} \eqsim \eta_\beta \sum_{j\geqslant 1}\lambda_j \eqsim \eta_\beta.$$

If $\eta_\beta \geqslant 1$, take $J_* := \lceil \eta_\beta^{1/q} \rceil$. The spectral decay gives

$$\lambda_{J_*} \eqsim \eta_\beta^{-1}, \qquad J_* \eqsim \eta_\beta^{1/q}.$$

For $j \leqslant J_*$, the summand is bounded above and below by constants, contributing $\eqsim J_*$. For $j > J_*$, it is $\eqsim \eta_\beta\lambda_j$, yielding

$$\eta_\beta \sum_{j>J_*} \lambda_j \eqsim \eta_\beta J_*^{1-q} \eqsim \eta_\beta^{1/q}.$$

Hence

$$\eta^2 \sum_{\tau=0}^{\infty} \mathcal{K}_{\mathrm{nesterov}}(\tau) \eqsim \min\left\{\eta_\beta^{1/q}, \eta_\beta\right\}. \tag{108}$$

This gives the upper bound in (104). Put $E_j := \sum_{\tau\geqslant 0}(G_\tau^{\mathrm{nesterov}}(j))^2$ and $\nu_j := \min\{1-\beta, \eta_\beta\lambda_j\}$. Write $d_j := 1-\beta(1-\eta\lambda_j)$, and set $z_j := 1+\varepsilon\nu_j$ for a sufficiently small absolute constant $\varepsilon > 0$. Since $\nu_j \leqslant d_j$ and $d_j\nu_j \leqslant 2\eta\lambda_j$, we have $\sqrt{z_j}\,|r_{j,\pm}| < 1$, $1-b_jz_j \eqsim d_j$, and $(1+b_jz_j)^2 - a_j^2z_j \eqsim \eta\lambda_j$. The weighted energy in Lemma A.6 is therefore comparable to E_j. Since $\log z_j \eqsim \nu_j$, the tail bound in that lemma gives constants $C, c > 0$ such that

$$\sum_{\tau=k}^{\infty} \left(G_\tau^{\mathrm{nesterov}}(j)\right)^2 \leqslant C\exp[-ck\min\{1-\beta, \eta_\beta\lambda_j\}] \sum_{\tau=0}^{\infty} \left(G_\tau^{\mathrm{nesterov}}(j)\right)^2. \tag{109}$$

The polynomial factor at a repeated root is absorbed by decreasing c.

If $\eta_\beta \leqslant 1$, fix J_* with $\sum_{j\leqslant J_*}\lambda_j \geqslant \frac{1}{2}\sum_{j\geqslant 1}\lambda_j$. By (106), the block $A = \{1, \dots, J_*\}$ carries a fixed fraction of $\sum_j \lambda_j^2 E_j$, and $\nu_j \gtrsim \min\{1-\beta, \eta_\beta\}$ on A.

If $\eta_\beta \geqslant 1$, take $A = \{j : \lambda_j \geqslant c_*\eta_\beta^{-1}\}$, with fixed sufficiently small $c_* > 0$. Then $|A| \eqsim \eta_\beta^{1/q}$, $\lambda_j^2 E_j \eqsim \eta^{-2}$ on A, and $\nu_j \gtrsim 1-\beta$. By (108), this block again carries a fixed fraction of the total mass.

Thus the stated lower bound on k, with a sufficiently large constant, makes the tail in (109) at most $E_j/2$ for $j \in A$. Consequently,

$$\sum_{\tau=0}^{k-1} \mathcal{K}_{\mathrm{nesterov}}(\tau) \geqslant \frac{1}{2}\sum_{j\in A}\lambda_j^2 E_j \gtrsim \sum_{\tau=0}^{\infty} \mathcal{K}_{\mathrm{nesterov}}(\tau),$$

and (108) gives (104). □

Multiplying the kernel sum by $\eta^2\sigma^2/B$ gives the label-noise contribution.

Corollary C.17 (Label noise for Nesterov)**.** *Assume $\eta\lambda_1 \leqslant c_{\mathrm{st}}$ for sufficiently small universal c_{st}. There is a constant $C > 0$ such that, if*

$$k \geqslant C\max\left\{\frac{1}{1-\beta}, \frac{1-\beta}{\eta}\right\},$$

then

$$N_{\mathrm{nesterov}}(k) \eqsim \frac{\sigma^2}{B}\min\left\{\left(\frac{\eta}{1-\beta}\right)^{1/q}, \frac{\eta}{1-\beta}\right\}. \tag{110}$$

Proof. This follows immediately from the definition of $N_{\mathrm{nesterov}}(k)$ and Proposition C.16. □

C.3.5 CONTROL OF MULTIPLICATIVE NOISE

Recall from (95) that

$$M_{\mathrm{nesterov}}(k) = \frac{\eta^2}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\mathrm{nesterov}}(\tau)\mathbb{E}[\mathcal{L}_{k-1-\tau}],$$

where $\mathcal{L}_m = \mathcal{E}(\boldsymbol{\theta}_m^{\mathrm{la}})$ is the risk at the gradient evaluation point. In this appendix, we control this term and show that

$$M_{\mathrm{nesterov}}(k) \lesssim N_{\mathrm{nesterov}}(k).$$

We first use the following lemma to compare $\mathbb{E}[\mathcal{L}_{k-1-\tau}]$ with $R_{k-\tau}$.

Lemma C.18 (Evaluation risk for Nesterov). *Under the small-step condition $\eta\lambda_1 \leqslant c_{\mathrm{st}} < 1$, there exist constants $c_{\mathrm{la}}, C_{\mathrm{la}} > 0$, independent of k, η, β, B, such that*

$$c_{\mathrm{la}}\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_{k+1})] - C_{\mathrm{la}}\frac{\eta^2\sigma^2}{B} \leqslant \mathbb{E}[\mathcal{L}_k] \leqslant C_{\mathrm{la}}\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_{k+1})]. \tag{111}$$

Proof. The stochastic Nesterov recursion at the look-ahead point is

$$\mathbf{e}_{k+1} = (\mathbf{I} - \eta\mathbf{H})\mathbf{e}_k^{\mathrm{la}} - \eta\boldsymbol{\xi}_k(\boldsymbol{\theta}_k^{\mathrm{la}}).$$

Conditioning on $\mathscr{F}_k$, the cross term vanishes, so

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_{k+1})] = \frac{1}{2}\mathbb{E}\big[\|(\mathbf{I} - \eta\mathbf{H})\mathbf{e}_k^{\mathrm{la}}\|_{\mathbf{H}}^2\big] + \frac{\eta^2}{2}\mathbb{E}\|\boldsymbol{\xi}_k(\boldsymbol{\theta}_k^{\mathrm{la}})\|_{\mathbf{H}}^2. \tag{112}$$

Because $1 - c_{\mathrm{st}} \leqslant 1 - \eta\lambda_j \leqslant 1$, the first term is comparable to $\mathbb{E}[\mathcal{L}_k]$. Moreover, Lemma A.3 gives

$$\mathbb{E}\|\boldsymbol{\xi}_k(\boldsymbol{\theta}_k^{\mathrm{la}})\|_{\mathbf{H}}^2 \lesssim \frac{\mathrm{Tr}(\mathbf{H}^2)}{B}\left(\mathbb{E}[\mathcal{L}_k] + \sigma^2\right).$$

The assumptions give $\eta^2\mathrm{Tr}(\mathbf{H}^2)/B \leqslant c_{\mathrm{st}}^2\mathrm{Tr}(\mathbf{H}^2)/\lambda_1^2$, so this coefficient is uniformly bounded. Hence there are constants $c_{\mathrm{cmp}}, C_0, C_1 > 0$ such that

$$c_{\mathrm{cmp}}\mathbb{E}[\mathcal{L}_k] \leqslant \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_{k+1})] \leqslant C_0\mathbb{E}[\mathcal{L}_k] + C_1\frac{\eta^2\sigma^2}{B}.$$

Rearranging these two inequalities gives (111). □

This comparison allows us to apply Lemma A.10 to the risk recursion. Together with the kernel estimate established above, it yields a uniform risk bound and hence the following control of multiplicative noise.

Proposition C.19 (Control of multiplicative noise for Nesterov). *Suppose $\sigma > 0$ and $\eta \leqslant c\min\{1, B^q(1-\beta)\}$ for sufficiently small $c > 0$. There is a constant $M < \infty$, independent of k, η, β, B, such that*

$$\sup_{k\geqslant 0}\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \leqslant M.$$

Moreover,

$$M_{\mathrm{nesterov}}(k) \lesssim N_{\mathrm{nesterov}}(k).$$

Proof. The signal learning bounds give $\sup_k S_{\mathrm{nesterov}}(k) \leqslant S_*$. By (108),

$$\kappa := \frac{\eta^2}{B}\sum_{\tau\geqslant 0}\mathcal{K}_{\mathrm{nesterov}}(\tau) \lesssim \frac{1}{B}\min\left\{\left(\frac{\eta}{1-\beta}\right)^{1/q}, \frac{\eta}{1-\beta}\right\} \leqslant \frac{1}{B}\left(\frac{\eta}{1-\beta}\right)^{1/q} \leqslant c^{1/q}.$$

Lemma C.18 gives $\mathbb{E}[\mathcal{L}_m] \leqslant C_{\mathrm{la}}R_{m+1}$. Let $A \geqslant 1$ satisfy

$$R_k \leqslant A\big(S_{\mathrm{nesterov}}(k) + N_{\mathrm{nesterov}}(k) + M_{\mathrm{nesterov}}(k)\big).$$

Set $x_k = R_k$, $y_k = \mathbb{E}[\mathcal{L}_k]$, $u_k = AS_{\mathrm{nesterov}}(k)$, and $w_\tau = A\eta^2\mathcal{K}_{\mathrm{nesterov}}(\tau)/B$. Take $v = \sigma^2$, $L = C_{\mathrm{la}}$, and $\iota = 1$. The risk inequality above has the form required by Lemma A.10, with $y_m \leqslant Lx_{m+\iota}$ and $U := \sup_k u_k \leqslant AS_*$. By these definitions,

$$L\sum_{\tau\geqslant 0} w_\tau = AC_{\mathrm{la}}\frac{\eta^2}{B}\sum_{\tau\geqslant 0}\mathcal{K}_{\mathrm{nesterov}}(\tau) = AC_{\mathrm{la}}\kappa.$$

Since $\kappa \lesssim c^{1/q}$, choosing c sufficiently small ensures $AC_{\mathrm{la}}\kappa \leqslant 1/2$, as required by the lemma. Applying Lemma A.10 now gives

$$\sup_k R_k = \sup_k x_k \leqslant 2U + \frac{v}{L} \leqslant 2AS_* + \frac{\sigma^2}{C_{\mathrm{la}}} =: M.$$

Therefore,

$$M_{\mathrm{nesterov}}(k) \leqslant \frac{\eta^2 C_{\mathrm{la}} M}{B}\sum_{\tau=0}^{k-1}\mathcal{K}_{\mathrm{nesterov}}(\tau) = \frac{C_{\mathrm{la}}M}{\sigma^2}N_{\mathrm{nesterov}}(k).$$

□

The multiplicative-noise bound gives a simplified form of the scaling law.

Corollary C.20 (Scaling law of simplified form for Nesterov)**.** *Under the assumptions of Proposition C.19,*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim S_{\mathrm{nesterov}}(k) + N_{\mathrm{nesterov}}(k). \tag{113}$$

Proof. By Proposition C.19, $M_{\mathrm{nesterov}}(k) \lesssim N_{\mathrm{nesterov}}(k)$. Substituting this bound into (94) and using nonnegativity proves the claim. □

C.3.6 Completion of the proof

We now combine the signal learning and noise estimates to establish the following scaling laws in the two damping regimes. The constant $c > 0$ is sufficiently small to ensure $\eta\lambda_1 \leqslant c_{\mathrm{st}}$, and $C > 0$ is sufficiently large to satisfy the time conditions of the preceding estimates.

Fully overdamped regime. Here the overdamped signal learning estimate combines directly with the label noise bound, since the multiplicative noise is controlled by the latter.

Theorem C.21 (Scaling law for Nesterov in the fully overdamped regime)**.** *Under Assumptions 3.1 and 3.3, suppose $\sigma > 0$, assume the Nesterov stability condition $\eta \leqslant c\min\{1, B^q(1-\beta)\}$, and consider the fully overdamped regime $\eta \lesssim (1-\beta)^2$. For k satisfying*

$$k \geqslant C\frac{1-\beta}{\eta},$$

the expected excess risk satisfies

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \left(\frac{1-\beta}{\eta k}\right)^s + \frac{\sigma^2\eta}{B(1-\beta)}.$$

Proof. By Corollary C.20, $R_k \eqsim S_{\mathrm{nesterov}}(k) + N_{\mathrm{nesterov}}(k)$. Proposition C.12 gives $S_{\mathrm{nesterov}}(k) \eqsim (1+\eta k/(1-\beta))^{-s}$. Since $k \geqslant C(1-\beta)/\eta$, this is comparable to $((1-\beta)/(\eta k))^s$. The stated learning-rate and time conditions also allow us to apply Corollary C.17. The fully overdamped condition and $\lambda_1 \eqsim 1$ imply $\eta/(1-\beta) \lesssim 1$, so the minimum in that corollary is comparable to $\eta/(1-\beta)$. Thus $N_{\mathrm{nesterov}}(k) \eqsim \sigma^2\eta/[B(1-\beta)]$. Substituting these estimates proves the theorem. □

Mixed-damping regime. Here we retain the underdamped transient and combine the overdamped tail estimate with the label noise bound.

Theorem C.22 (Scaling law for Nesterov in the mixed-damping regime)**.** *Under Assumptions 3.1 and 3.3, suppose $\sigma > 0$, assume the Nesterov stability condition $\eta \leqslant c\min\{1, B^q(1-\beta)\}$, and consider the mixed-damping regime $\eta \gtrsim (1-\beta)^2$. For k satisfying*

$$k \geqslant \frac{C}{1-\beta},$$

the expected excess risk satisfies

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim \mathbb{T}_N(k) + \left(\frac{1-\beta}{\eta k}\right)^s + \frac{\sigma^2}{B}\min\left\{\left(\frac{\eta}{1-\beta}\right)^{1/q}, \frac{\eta}{1-\beta}\right\}. \tag{114}$$

Here $\mathbb{T}_N(k)$ is the separated underdamped transient defined in Proposition C.14 and satisfies

$$0 \leqslant \mathbb{T}_N(k) \lesssim \min\{1, (\eta k)^{-s}\}\beta^k.$$

Proof. By Corollary C.20, $R_k \eqsim S_{\mathrm{nesterov}}(k) + N_{\mathrm{nesterov}}(k)$. Corollary C.15 gives

$$S_{\mathrm{nesterov}}(k) \eqsim \mathbb{T}_N(k) + \left(\frac{(1-\beta)^2}{\eta}\right)^s \big(1 + (1-\beta)k\big)^{-s}.$$

Since $k \geqslant C/(1-\beta)$, the second term is comparable to $((1-\beta)/(\eta k))^s$. The learning-rate and time conditions also imply the assumptions of Corollary C.17, which supplies the stated label noise term. Combining these estimates proves the theorem. □

C.4 Scaling law for SGD

We recall the SGD scaling law, which is the special case $M = \infty$ of Li et al. (2025, Theorem 5.2). Below, we give an alternative proof by specializing the Polyak result to $\beta = 0$. It provides the baseline for comparing the signal learning rate and noise floor of the two momentum methods.

Theorem C.23 (SGD scaling law (Li et al., 2025))**.** *Under Assumptions 3.1 and 3.3, if the stability condition $\eta \lesssim 1$ holds, then for $k \gtrsim 1/\eta$,*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim (\eta k)^{-s} + \frac{\eta\sigma^2}{B}. \tag{115}$$

Proof. Since Polyak momentum reduces to SGD when $\beta = 0$, the result follows by setting $\beta = 0$ in Theorem C.10. □

D Optimal data scaling across batch-size regimes

This appendix proves the batch-size-dependent scaling laws stated in Theorem 6.3. Throughout this appendix, we work in the noisy setting $\sigma > 0$ considered in Section 6 and impose the data-budget constraint

$$B \eqsim D^b, \qquad N = \frac{D}{B} \eqsim D^{1-b}, \qquad 0 \leqslant b < 1.$$

Thus, the batch exponent b is prescribed, and only the remaining hyperparameters are optimized. We use the transition exponents defined in Section 6:

$$b_1 = \frac{s}{s+1}, \qquad b_2 = \frac{2s+1/q}{2s+1+1/q} = \frac{2sq+1}{2sq+q+1}, \qquad b_3 = \frac{2s+1}{2s+2}.$$

For the momentum methods, we parameterize the momentum gap and learning rate by $1-\beta \eqsim D^{-m}$ and $\eta \eqsim D^{-c}$, where $m, c \geqslant 0$.

D.1 Data scaling for SGD

For SGD, write $\eta \eqsim D^{-c}$, where $c \geqslant 0$. By the SGD risk scaling law in Section 5,

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim (\eta N)^{-s} + \frac{\sigma^2 \eta}{B} \eqsim D^{-s(1-b-c)} + D^{-(b+c)}.$$

Therefore, for a fixed batch exponent b, the achievable decay exponent is $\min\{s(1-b-c), b+c\}$. If $0 \leqslant b < b_1$, the two terms can be balanced by taking $c = b_1 - b \geqslant 0$. Since $b_1 = s/(s+1)$, this gives

$$s(1-b-c) = b + c = b_1 = \frac{s}{s+1}.$$

Hence $r_{\mathrm{sgd}}(b) = s/(s+1)$ for $0 \leqslant b < b_1$.

If $b \geqslant b_1$, the unconstrained balancing choice would require $c = b_1 - b \leqslant 0$. Because the learning rate is not allowed to grow polynomially with D, we must have $c \geqslant 0$. Moreover, $b + c \geqslant b \geqslant b_1$, so the signal learning term is the slower-decaying term. It is optimized by taking $c = 0$, which yields $r_{\mathrm{sgd}}(b) = s(1-b)$ for $b_1 \leqslant b < 1$.

Consequently,

$$r_{\mathrm{sgd}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_1, \\ s(1-b), & b_1 \leqslant b < 1. \end{cases}$$

Thus, b_1 is precisely the largest batch exponent for which SGD preserves the data-optimal rate $s/(s+1)$.

D.2 Data scaling for Polyak

We next prove the Polyak part of Theorem 6.3. We retain $N \eqsim D^{1-b}$, $1-\beta \eqsim D^{-m}$, and $\eta \eqsim D^{-c}$, where $m, c \geqslant 0$. We require $b + m < 1$. Proposition C.4 gives

$$\mathbb{T}_P(N) \lesssim \beta^N \leqslant \exp\bigl(-(1-\beta)N\bigr) = \exp\bigl(-\Theta(D^{1-b-m})\bigr),$$

which is smaller than any polynomial power of D. The stability condition $\eta \lesssim B(1-\beta)$ becomes $b + c - m \geqslant 0$. Moreover,

$$\eta_\beta N = \frac{\eta N}{1-\beta} \eqsim D^{1-b-c+m}, \qquad \frac{\eta_\beta}{B} = \frac{\eta}{B(1-\beta)} \eqsim D^{-(b+c-m)}.$$

The damping threshold $\eta \eqsim (1-\beta)^2$ corresponds to $c = 2m$. We consider the fully overdamped and mixed-damping regimes separately.

Fully overdamped regime. If $c \geqslant 2m$, Theorem 5.1 gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim (\eta_\beta N)^{-s} + \frac{\sigma^2 \eta_\beta}{B} \eqsim D^{-s(1-b-c+m)} + D^{-(b+c-m)}.$$

Therefore, the polynomial decay exponent in the fully overdamped regime is $\min\{s(1-b-c+m), b+c-m\}$.

Mixed-damping regime. If $c \leqslant 2m$, Theorem 5.1 gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim \mathbb{T}_P(N) + (\eta_\beta N)^{-s} + \frac{\sigma^2 \eta_\beta}{B}.$$

Since $b + m < 1$, Proposition C.4 implies

$$\mathbb{T}_P(N) \lesssim \beta^N \leqslant \exp\bigl(-(1-\beta)N\bigr) = \exp\bigl(-\Theta(D^{1-b-m})\bigr),$$

which is smaller than any polynomial power of D. Hence

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim D^{-s(1-b-c+m)} + D^{-(b+c-m)},$$

and the polynomial decay exponent is again $\min\{s(1-b-c+m), b+c-m\}$. Thus, although the two regimes have different transient dynamics, they lead to the same polynomial optimization problem. The signal learning and additive-noise powers are balanced when

$$s(1-b-c+m) = b+c-m \quad \Longleftrightarrow \quad b+c-m = \frac{s}{s+1} = b_1.$$

If $0 \leqslant b < b_1$, this balance is attained by taking $m = 0$ and $c = b_1 - b$. Since $c \geqslant 2m$, this choice lies in the fully overdamped regime and yields $r_{\text{polyak}}(b) = s/(s+1)$.

If $b_1 \leqslant b < b_3$, the same balance is attained by taking $c = 0$ and $m = b - b_1$. For $b > b_1$, this choice satisfies $c \leqslant 2m$ and therefore lies in the mixed-damping regime. Moreover, $b+c-m = b_1$, while the transient condition becomes $b+m = 2b-b_1 < 1$. Since $(1+b_1)/2 = (2s+1)/(2s+2) = b_3$, this condition holds precisely when $b < b_3$. Hence $r_{\text{polyak}}(b) = s/(s+1)$ for $0 \leqslant b < b_3$.

Now suppose that $b \geqslant b_3$. Since $c \geqslant 0$ and $m < 1-b$, every admissible choice satisfies $b+c-m > 2b-1 \geqslant b_1$, so the signal learning term is the slower-decaying term and $s(1-b-c+m) < 2s(1-b)$. This upper bound is approached by taking $c = 0$ and $m \uparrow 1-b$, which lies in the mixed-damping regime and gives

$$s(1-b-c+m) \uparrow 2s(1-b), \qquad b+c-m \downarrow 2b-1.$$

Moreover, $2b-1 \geqslant 2s(1-b)$ holds exactly when $b \geqslant b_3$, so the signal learning term is indeed dominant. Consequently, $r_{\text{polyak}}(b) = 2s(1-b)$ for $b_3 \leqslant b < 1$, where the endpoint value is understood in the supremal sense represented by the $o(1)$ term in Definition 6.2. Combining the two regimes gives

$$r_{\text{polyak}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_3, \\ 2s(1-b), & b_3 \leqslant b < 1. \end{cases}$$

Thus, Polyak preserves the optimal data-scaling exponent $s/(s+1)$ up to the critical batch exponent b_3. Compared with SGD, it extends the critical batch size from D^{b_1} to D^{b_3}, while leaving the optimal small-batch exponent unchanged.

D.3 Data scaling for Nesterov

We finally prove the Nesterov part of Theorem 6.3. We retain $B \eqsim D^b$, $N \eqsim D^{1-b}$, $1-\beta \eqsim D^{-m}$, and $\eta \eqsim D^{-c}$, where $m, c \geqslant 0$. We require $b+m < 1$. Proposition C.14 gives

$$\mathbb{T}_N(N) \lesssim \min\{1, (\eta N)^{-s}\}\beta^N \leqslant \beta^N \leqslant \exp\big(-(1-\beta)N\big) = \exp\big(-\Theta(D^{1-b-m})\big),$$

which is smaller than any polynomial power of D. The stability condition $\eta \lesssim B^q(1-\beta)$ becomes $qb+c-m \geqslant 0$, while the damping threshold $\eta \eqsim (1-\beta)^2$ corresponds to $c = 2m$. We optimize the fully overdamped and mixed-damping regimes separately and then compare their optimal decay exponents.

Fully overdamped regime. If $c \geqslant 2m$, Theorem 5.2 gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim (\eta_\beta N)^{-s} + \frac{\sigma^2 \eta_\beta}{B} \eqsim D^{-s(1-b-c+m)} + D^{-(b+c-m)}.$$

Thus, the decay exponent is $\min\{s(1-b-c+m), b+c-m\}$. Since $c \geqslant 2m$ implies $c-m \geqslant 0$, we have $b+c-m \geqslant b$. If $0 \leqslant b < b_1$, the signal learning and additive-noise powers can be balanced by taking $m = 0$ and $c = b_1 - b$, which gives $b+c-m = b_1$ and decay exponent $s/(s+1)$. If $b \geqslant b_1$, then $b+c-m \geqslant b \geqslant b_1$, so the signal learning term is limiting. Its exponent is maximized by taking $m = c = 0$, which gives $s(1-b)$. Therefore, the optimal exponent in the fully overdamped regime is

$$r^{\text{full}}_{\text{nesterov}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_1, \\ s(1-b), & b_1 \leqslant b < 1. \end{cases}$$

In particular, the fully overdamped Nesterov dynamics have the same batch-size-dependent scaling exponent as SGD.

Mixed-damping regime. If $c \leqslant 2m$, Theorem 5.2 gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim \mathbb{T}_N(N) + (\eta_\beta N)^{-s} + \frac{\sigma^2}{B} \min\left\{\eta_\beta, \eta_\beta^{1/q}\right\}.$$

Since $b + m < 1$, Proposition C.14 implies

$$\mathbb{T}_N(N) \lesssim \min\{1, (\eta N)^{-s}\}\beta^N \leqslant \beta^N \leqslant \exp\big(-(1-\beta)N\big) = \exp\big(-\Theta(D^{1-b-m})\big),$$

which is smaller than any polynomial power of D. Hence

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \eqsim D^{-s(1-b-c+m)} + \frac{\sigma^2}{B} \min\left\{\frac{\eta}{1-\beta}, \left(\frac{\eta}{1-\beta}\right)^{1/q}\right\}.$$

The signal learning term has decay power $s(1-b-c+m)$. Since $\eta/(1-\beta) \eqsim D^{m-c}$, the additive-noise decay power is

$$\begin{cases} b + c - m, & c \geqslant m, \\ b + \dfrac{c-m}{q}, & c \leqslant m. \end{cases}$$

We optimize these two branches separately.

The branch $c \geqslant m$. In this branch, the decay exponent is

$$\min\{s(1-b-c+m), b+c-m\}.$$

Because $c - m \geqslant 0$, the same argument as in the fully overdamped regime applies. For $0 \leqslant b < b_1$, the balance $b + c - m = b_1$ is feasible in the mixed-damping regime. For example, one may take $m = b_1 - b$ and $c = 2(b_1 - b)$, for which $b + m = b_1 < 1$. For $b \geqslant b_1$, the optimal choice satisfies $c - m = 0$ and gives exponent $s(1-b)$. Therefore,

$$r_{\text{nesterov}}^{\text{mixed, c}\geqslant\text{m}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_1, \\ s(1-b), & b_1 \leqslant b < 1. \end{cases}$$

The branch $c \leqslant m$. In this branch, the decay exponent is

$$\min\left\{s(1-b-c+m),\, b + \frac{c-m}{q}\right\}.$$

If $0 \leqslant b < b_1$, then at $m - c = 0$ we have $b < s(1-b)$. Increasing $m - c$ increases the signal learning power but decreases the additive-noise power, so the optimal choice is $m - c = 0$, which gives decay exponent b. If $b \geqslant b_1$, balancing the signal learning and additive-noise powers gives

$$s(1-b-c+m) = b + \frac{c-m}{q}, \qquad m - c = \frac{q\big((s+1)b - s\big)}{sq+1}.$$

For a fixed value of $m - c$, the least restrictive choice for the transient condition is $c = 0$, so we take

$$c = 0, \qquad m = \frac{q\big((s+1)b - s\big)}{sq+1}.$$

This choice satisfies the stability condition because

$$qb + c - m = \frac{qs\big(1 + (q-1)b\big)}{sq+1} > 0.$$

The signal learning and additive-noise powers are then equal to

$$s(1-b+m) = b - \frac{m}{q} = \frac{s\big(1+(q-1)b\big)}{sq+1}.$$

The transient condition holds precisely when

$$b + m < 1 \iff b < \frac{2sq+1}{2sq+q+1} = \frac{2s+1/q}{2s+1+1/q} = b_2.$$

Therefore, for $b_1 \leqslant b < b_2$, the optimal exponent on the branch $c \leqslant m$ is

$$\frac{s\big(1+(q-1)b\big)}{sq+1}.$$

If $b \geqslant b_2$, the preceding balance is no longer compatible with $b+m<1$. Since $c \geqslant 0$ and $m < 1-b$, every admissible choice satisfies

$$s(1-b-c+m) < 2s(1-b).$$

This upper bound is approached by taking $c = 0$ and $m \uparrow 1-b$. Under this choice,

$$s(1-b+m) \uparrow 2s(1-b), \qquad b - \frac{m}{q} \to \frac{(q+1)b-1}{q}.$$

At $b = b_2$, the two limiting powers agree, while for $b > b_2$,

$$\frac{(q+1)b-1}{q} > 2s(1-b).$$

Thus, the signal learning term is limiting and the supremal decay exponent is $2s(1-b)$. Consequently,

$$r^{\text{mixed, c}\leqslant\text{m}}_{\text{nesterov}}(b) = \begin{cases} b, & 0 \leqslant b < b_1, \\ \dfrac{s\big(1+(q-1)b\big)}{sq+1}, & b_1 \leqslant b < b_2, \\ 2s(1-b), & b_2 \leqslant b < 1. \end{cases}$$

Taking the larger exponent between the two additive-noise branches gives the optimal exponent in the mixed-damping regime:

$$r^{\text{mixed}}_{\text{nesterov}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_1, \\ \dfrac{s\big(1+(q-1)b\big)}{sq+1}, & b_1 \leqslant b < b_2, \\ 2s(1-b), & b_2 \leqslant b < 1. \end{cases}$$

Indeed, the branch $c \geqslant m$ is optimal for $b < b_1$, whereas the branch $c \leqslant m$ is optimal for $b > b_1$. Finally, optimizing over the fully overdamped and mixed-damping regimes gives

$$r_{\text{nesterov}}(b) = \max\left\{r^{\text{full}}_{\text{nesterov}}(b), r^{\text{mixed}}_{\text{nesterov}}(b)\right\}.$$

For $0 \leqslant b < b_1$, the two regimes attain the same exponent $s/(s+1)$. For $b_1 < b < b_2$, the mixed-damping regime exponent $s(1+(q-1)b)/(sq+1)$ is strictly larger than the fully overdamped exponent $s(1-b)$. For $b \geqslant b_2$, the mixed-damping regime exponent $2s(1-b)$ is also larger than the fully overdamped exponent $s(1-b)$. We therefore obtain

$$r_{\text{nesterov}}(b) = \begin{cases} \dfrac{s}{s+1}, & 0 \leqslant b < b_1, \\ \dfrac{s\big(1+(q-1)b\big)}{sq+1}, & b_1 \leqslant b < b_2, \\ 2s(1-b), & b_2 \leqslant b < 1. \end{cases}$$

Thus, Nesterov agrees with SGD and Polyak in the small-batch regime, improves its data-scaling exponent once b exceeds b_1, and reaches its maximal exponent at $b = b_2$. For $b \geqslant b_2$, its exponent becomes $2s(1-b)$, which coincides with the Polyak exponent in the ultra-large-batch regime.

E Experimental details and additional results

This appendix provides experimental details and additional results for Section 7. In Section E.1, we first present the PLK experiments corresponding to Section 7.1, including the stability boundaries, risk dynamics, and batch-size scaling results. We then turn to neural-network experiments: we provide the full setup for the ImageNet-1K experiment in Section E.2 and report an additional CIFAR-10 experiment that examines whether the predicted stability behavior also appears in neural-network training.

E.1 Power-law kernel regression

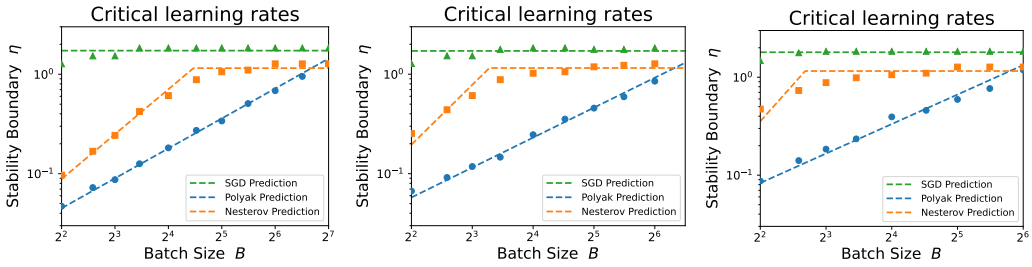
Our theoretical analysis is formulated in the infinite-dimensional PLK setting. For numerical experiments, we truncate the system to the first $d = 4096$ eigendirections, so that, for example, $\mathbf{H} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d\times d}$. Unless otherwise stated, all experiments use $d = 4096$.

E.1.1 Risk-stability boundaries

Simulation setup. We run with $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0 = 0$ for $N = 2000$ iterations. A run is classified as stable if its average risk over the final 20 iterations is below 1. It is classified as unstable if the risk blows up or exceeds 10^{12}. For each setting and random seed, the empirical stability boundary is the largest stable learning rate on the search grid. We repeat each scan over five independent seeds and report the mean boundary.

- **Stability versus batch size.** For Figure 2 (left), we use $(s, q) = (0.3, 1.5)$, fix $1 - \beta = 2^{-6}$ for the momentum methods, and scan $B \in \mathrm{round}\{2^a\}, a \in \{2, 2.5, \ldots, 7\}$ and 50 logarithmically spaced learning rates between 2^{-10} and 8. The dashed curves are post-hoc fits to the predicted scalings: a constant for SGD, $B(1-\beta)$ for Polyak, and $B^q(1-\beta)$ followed by a constant plateau for Nesterov. The dotted line marks the split used for the Nesterov fit.
- **Stability versus momentum damping.** For Figure 2 (middle), we use $(s, q) = (0.3, 1.5)$, $\sigma = 0$, and $B = 4$, vary $1-\beta \in \{2^{-8}, 2^{-7.5}, 2^{-7}, \ldots, 2^{-3.5}\}$, and scan 50 logarithmically spaced learning rates between 2^{-10} and 8. Since SGD is independent of β, its boundary is constant across the damping grid. The dashed curves are post-hoc fits to the predicted form $\eta^{\mathrm{crit}} = \min\{L, c(1-\beta)\}$ for Polyak and Nesterov, and to a constant for SGD.

Additional results. Figure 4 shows additional results for different choices of (s, q).



(a) $s = 0.5, q = 1.5$ (b) $s = 0.2, q = 2$ (c) $s = 0.5, q = 2.5$

Figure 4: Additional stability boundaries across (s, q).

E.1.2 Risk-dynamics scaling laws

Simulation setup. We simulate PLK regression initialized with $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0 = 0$. For Figure 2 (right), we set $(s, q) = (0.3, 4)$, $\sigma = 1$, $B = 32$, $\eta = 0.03$, and $\beta = 0.98$. Each method is run for 2×10^7 iterations. Let $\overline{\mathcal{E}}(k)$ denote the population excess risk averaged over the 100 runs.

Scaling-law fits. Theorems 5.1 and 5.2 determine the functional forms of the risk dynamics up to multiplicative constants. We therefore introduce $\Theta = (C_1, C_2, C_3) \in \mathbb{R}^3_+$ and fit the following ansatz separately for each method. For Polyak,

$$F_{\mathrm{polyak}}(k; \Theta) = C_1\beta^k + C_2(\eta_\beta k)^{-s} + C_3\frac{\sigma^2}{B}\eta_\beta, \qquad \eta_\beta := \frac{\eta}{1-\beta},$$

while for Nesterov,

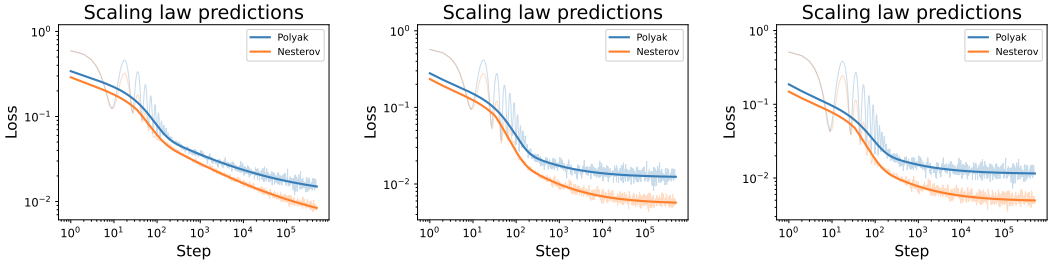
$$F_{\mathrm{nesterov}}(k; \Theta) = C_1 \min\{1, (\eta k)^{-s}\}\beta^k + C_2(\eta_\beta k)^{-s} + C_3\frac{\sigma^2}{B}\min\{\eta_\beta, \eta_\beta^{1/q}\}.$$

The first terms use the transient envelopes from Theorems 5.1 and 5.2, so the fits test the predicted scaling forms rather than their multiplicative constants. For each method, the coefficients are fitted by minimizing the mean squared relative error:

$$\min_{\Theta \in \mathbb{R}^3_+} \frac{1}{|\mathcal{I}_{\text{fit}}|} \sum_{k \in \mathcal{I}_{\text{fit}}} \left(\frac{F(k;\Theta) - \overline{\mathcal{E}}(k)}{\overline{\mathcal{E}}(k)} \right)^2, \qquad \mathcal{I}_{\text{fit}} := \{k \geqslant 1 : \eta_\beta k \geqslant 1\},$$

where $F = F_{\text{polyak}}$ or F_{nesterov} for the corresponding method.

Additional results. We repeat the same experiment for $(s, q) \in \{(0.3, 6), (0.5, 4), (0.5, 6)\}$, keeping $\sigma = 1$, $B = 32$, $\eta = 0.03$, and $\beta = 0.98$. Each run uses 5×10^5 iterations, and the reported risk is averaged over 100 independent runs. Figure 5 shows that the predicted scaling forms consistently track the empirical risk dynamics across these choices of (s, q).



(a) $(s, q) = (0.3, 6)$ (b) $(s, q) = (0.5, 4)$ (c) $(s, q) = (0.5, 6)$

Figure 5: **Additional validation of the risk-dynamics scaling laws for Polyak and Nesterov.** Thin translucent lines show empirical excess risks, while thick solid curves show the fitted scaling-law predictions. All experiments use $\sigma = 1$, $B = 32$, $\eta = 0.03$, and $\beta = 0.98$, run for 5×10^5 iterations, and report averages over 100 independent runs.

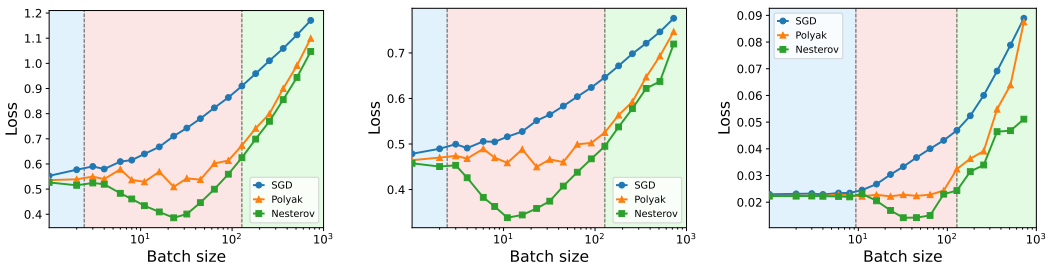
E.1.3 Batch-size phase diagram at a fixed data budget

We simulate PLK regression with $(s, q) = (0.3, 4)$ and $\sigma = 0.1$, with all methods initialized at $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0 = 0$. We fix the total data budget at $D = 2^{14}$. For each batch size, we search locally around the hyperparameter scalings predicted by Appendix D. Each candidate is evaluated over 50 independent runs. To reduce sensitivity to oscillations of the last iterate, we use the excess risk averaged over the final 1% of iterations as a numerical proxy for the final-step excess risk. Let $W = \max\{1, \lceil 0.01N \rceil\}$. We score each candidate by the mean excess risk over the final W iterations and across runs:

$$\overline{\mathcal{E}} = \frac{1}{50} \sum_{r=1}^{50} \frac{1}{W} \sum_{k=N-W+1}^{N} \mathcal{E}(\boldsymbol{\theta}_k^{(r)}).$$

For each method and batch size, Figure 1 (right) reports the minimum $\overline{\mathcal{E}}$ over the corresponding search grid.

Figure 6 shows additional results on the batch-size phase diagram across (s, q).



(a) $s = 0.1, q = 2$ (b) $s = 0.1, q = 4$ (c) $s = 0.3, q = 6$

Figure 6: Additional experiments on the batch-size phase diagram under different (s, q).

E.1.4 DATA-SCALING EXPONENTS ACROSS BATCH-SIZE REGIMES

In Figure 3, we simulate PLK regression with $(s, q) = (0.3, 4)$, $\sigma = 0.1$, and $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0 = 0$, using data budgets $D \in \{2000, 4000, 8000, 16000, 32000, 64000, 128000\}$. For a prescribed batch-size exponent b, we set $B = \mathrm{round}(C_B D^b)$ and $N = \lfloor D/B \rfloor$, where C_B is an order-one constant. For each method and D, we tune the multiplicative constants around the learning-rate and momentum scalings prescribed by Appendix D, and report the best stable configuration. Each run is evaluated by averaging the risk over the final 1% of optimization steps, and each point is averaged over 10 independent runs.

We evaluate one representative batch-size exponent from each of the three regimes.

- **Small batch.** We take $b_{\mathrm{small}} = 0$. For all three methods, we search around $\eta \eqsim D^{-s/(s+1)}$. For Polyak and Nesterov, we additionally use $1 - \beta \eqsim 1$.
- **Large batch.** We take $b_{\mathrm{large}} = b_2 - \delta$, where $b_2 = \frac{2sq+1}{2sq+q+1}$ and $\delta > 0$ is small. SGD uses $\eta \eqsim 1$. For Polyak, we search around $\eta \eqsim 1$ and $1 - \beta \eqsim D^{-(b_{\mathrm{large}} - s/(s+1))}$. For Nesterov, we search around $\eta \eqsim 1$ and
$$1 - \beta \eqsim D^{-m_{\mathrm{nesterov}}}, \qquad m_{\mathrm{nesterov}} = b_{\mathrm{large}} - \frac{sq - (q-1)b_{\mathrm{large}}}{sq + 1}.$$
- **Ultra-large batch.** We choose $b_3 < b_{\mathrm{ultra}} < 1$. SGD uses $\eta \eqsim 1$, while for both momentum methods we search around $\eta \eqsim 1$ and $1 - \beta \eqsim D^{-(1 - b_{\mathrm{ultra}} - \delta)}$, with $\delta > 0$ chosen so that $N(1 - \beta) \to \infty$.

Additional results. Figures 7 and 8 show additional data-scaling results across (s, q).

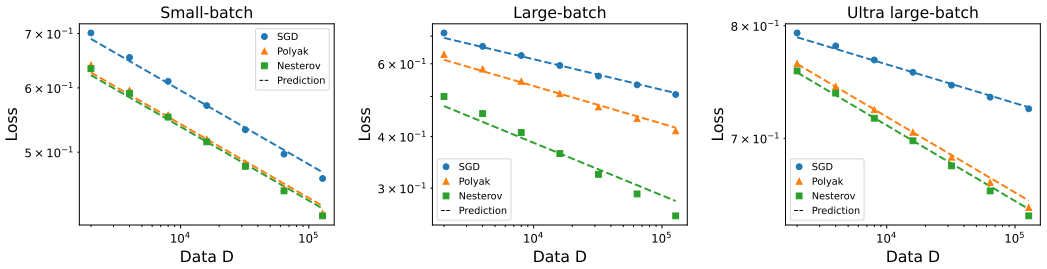


Figure 7: Data scaling with $(s, q) = (0.1, 4)$.

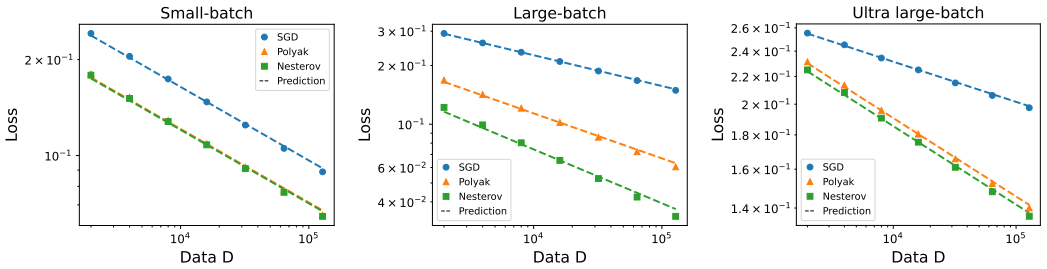


Figure 8: Data scaling with $(s, q) = (0.3, 2)$.

E.2 NEURAL-NETWORK EXPERIMENTS

E.2.1 IMAGENET EXPERIMENTS

We train ResNet-18 from scratch on ImageNet-1K, which contains 1,281,167 training images and 50,000 validation images across 1,000 classes. We replace batch normalization with group normalization to avoid dependence on batch statistics when varying the batch size.

We consider $B \in \{2^0, 2^1, \ldots, 2^{12}\}$. For each method and batch size, we fix the total sample budget at $D = 3 \times 1{,}281{,}167$, equivalent to three training epochs, and perform $N = \lfloor D/B \rfloor$ optimizer

updates. We train using cross-entropy loss with a constant learning rate and momentum factor, without learning-rate warmup, weight decay, or label smoothing. All runs use BF16 mixed precision and the same random seed. For $B > 128$, we accumulate gradients over micro-batches of size 128 to obtain the prescribed batch size.

For each method and batch size, we perform a grid search over the learning rate and, for Polyak and Nesterov, the momentum factor. We evaluate each final model on the full validation set. Figure 9 reports the lowest validation cross-entropy among the completed stable configurations for each method and batch size.

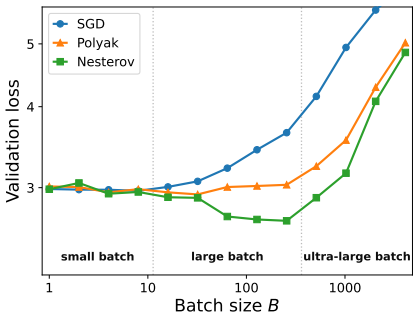


Figure 9: Batch-size dependence on ImageNet-1K under a fixed sample budget. Each point reports the lowest final validation cross-entropy among the evaluated stable hyperparameter configurations for the corresponding method and batch size.

E.2.2 CIFAR-10 EXPERIMENTS

We further test whether the stability behavior predicted by the PLK analysis in Section 4 also appears in neural-network training. We train ResNet-18 on CIFAR-10 using cross-entropy loss, replacing batch normalization with group normalization and adapting the initial convolution to 32×32 inputs. Each run starts from a random initialization and performs $N = 2000$ updates with constant learning rates and momentum factors, without warmup or weight decay. For the batch-size scan, we fix $1 - \beta = 2^{-6}$ and vary $B \in \{2^0, 2^1, \ldots, 2^8\}$. For the damping scan, we fix $B = 4$ and vary $1 - \beta \in \{2^{-6}, 2^{-5.5}, \ldots, 2^{-2}\}$. Learning rates are scanned logarithmically from 2^{-8} to 8, with additional learning rates and random seeds near the observed transitions.

We evaluate loss on a fixed subset of 1024 training images every 50 updates. A completed run is stable if its average loss over the final ten evaluations is at most twice its initial loss and 1.25 times the average over the preceding ten evaluations. Runs terminated by numerical divergence or safety limits are classified as unstable.

Figure 10 shows the results. We can see that the SGD boundary becomes approximately constant beyond the smallest batch sizes, while the momentum boundaries increase with B before saturation. At fixed $B = 4$, both momentum boundaries initially increase with $1-\beta$ and then approach plateaus, with Nesterov saturating earlier. These observations are qualitatively consistent with the growth and saturation predicted by the theory, while the precise scaling may vary in neural network training.

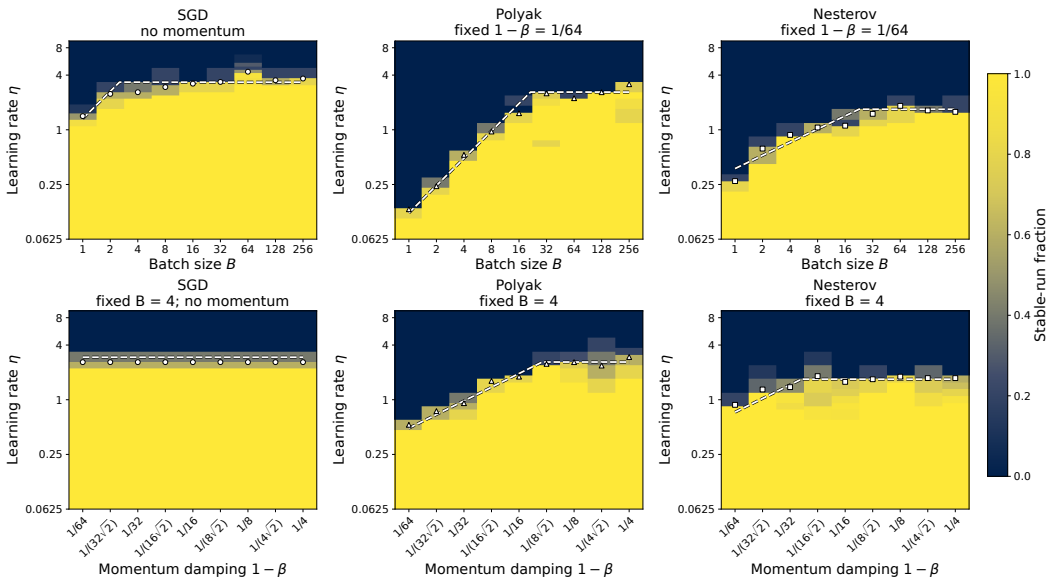


Figure 10: Empirical risk stability of ResNet-18 on CIFAR-10 over $N = 2000$ updates. **Top:** varying B, with $1 - \beta = 2^{-6}$ for the momentum methods. **Bottom:** varying $1 - \beta$ at fixed $B = 4$. Colors show stable-run fractions, white markers indicate estimated 50% stability boundaries, and dashed curves show fitted relations.