When and How SignSGD Outperforms SGD: A Functional Scaling Law Perspective

Zilin Wang[1,*] **Binghui Li**[1,*] **Jia-Nan Wang**[1,*] **Lean Wang**[1,*] **Jinbo Wang**[1,*] **Lei Wu**[1,2,†]

[1]Peking University [2]AI for Science Institute, Beijing

Abstract

We study how the sign operator interacts with gradient noise to shape the learning dynamics and scaling behavior of sign-based optimizers. In power-law kernel regression with diagonal features, we establish a functional scaling law (FSL) for SignSGD under general learning rate schedules. The FSL describes the learning dynamics through two processes: signal learning and noise forgetting. The key mechanism is noise-induced preconditioning: gradient noise smooths the sign operator, inducing an effective preconditioned SGD. Noise-curvature alignment enables this preconditioning to transform the effective Hessian from $\mathbf{H}$ to $\mathbf{H}^{1/2}$. The resulting dynamics exhibit faster signal learning but slower noise forgetting than SGD. This tradeoff yields a regime-dependent comparison with SGD: (i) SignSGD achieves improved data efficiency in the hard-task regime; (ii) both optimizers achieve the same data efficiency in the easy-task regime, but SignSGD maintains this efficiency at larger batch sizes, requiring fewer optimization steps for a fixed data budget. Power-law kernel regression experiments support the proposed mechanism and predicted scaling behavior. LLM pre-training experiments further suggest that the SignSGD FSL better captures the batch-size dependence of loss dynamics for sign-based optimizers than the original SGD FSL.

1 Introduction

Sign-based optimizers play an increasingly prominent role in modern large-scale training. SignSGD [4] and Lion [8] apply the sign operator coordinate-wise, while Muon [21, 33] uses the matrix sign operator. Related sign-like behavior also arises in RMSprop [16] and Adam [23, 35] through coordinate-wise second-moment normalization [3]. Together, these methods motivate understanding the sign operator as a basic mechanism in modern optimization.

Existing theoretical analyses of these sign-based optimizers have largely focused on deterministic full-batch dynamics [24, 25, 26, 29, 62] or convergence guarantees for convex and nonconvex objectives under general smoothness assumptions [9, 13, 44, 46, 50]. These analyses establish important optimization properties, but do not fully explain how the sign operator interacts with gradient noise or adequately quantify how this interaction shapes loss dynamics and scaling behavior.

In this work, we study SignSGD in power-law kernel (PLK) regression with diagonal features. We extend the functional scaling law (FSL) framework of Li et al. [27] to characterize how sign operator reshapes signal learning and noise forgetting under general learning rate (LR) schedules. Our analysis shows that SignSGD acts as an effective preconditioned SGD method, with the preconditioner determined by the structure of gradient noise. Diagonal features and noise-curvature alignment are important in this square-root spectral preconditioning, accelerating signal learning in low-curvature

*Equal contribution. Emails: wangzilin@stu.pku.edu.cn, libinghui@pku.edu.cn, jiananwang25@stu.pku.edu.cn, lean@pku.edu.cn, wangjinbo@stu.pku.edu.cn.

†Corresponding author. Email: leiwu@math.pku.edu.cn.

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

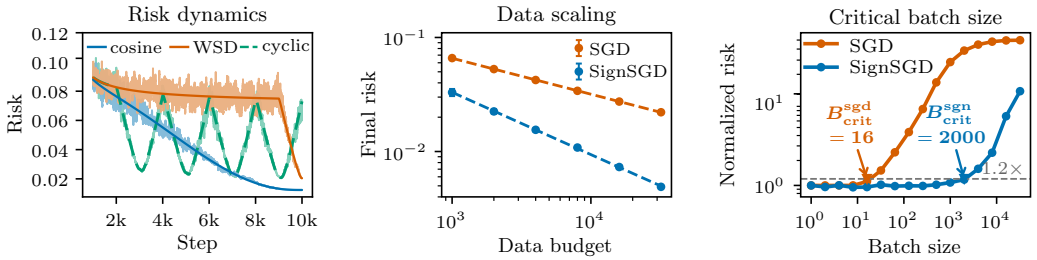


Figure 1: **SignSGD FSL accurately captures risk dynamics and predicts scaling behavior. (Left)** FSL predictions track SignSGD risk trajectories under cosine, WSD, and cyclic LR schedules. **(Middle)** In the hard-task regime, SignSGD achieves a larger data-scaling exponent than SGD, $\gamma_{\text{sgn}} > \gamma_{\text{sgd}}$ in $\mathcal{E}(D) \eqsim D^{-\gamma}$, indicating improved data efficiency. Dashed lines denote power-law fits. **(Right)** In the easy-task regime, SignSGD improves step efficiency through a larger CBS, requiring fewer optimization steps at a fixed data budget while preserving small-batch performance. Risk is normalized by each optimizer's $B = 1$ baseline. Marked CBS estimates use a 20% risk-increase tolerance.

directions. The accompanying change in update noise, however, leads to slower noise forgetting. The FSL captures these effects by expressing the risk at step k as

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \eqsim S(t_k) + \sum_{j=0}^{k-1} K(t_k - t_j)\eta_j^2.$$

Here η_j is the learning rate at step j, and $t_k = \sqrt{B}\sum_{j=0}^{k-1}\eta_j$ is the **intrinsic time** of SignSGD with batch size B. Signal-learning function $S(t_k)$ describes how the target signal is learned, and noise-forgetting kernel $K(t_k - t_j)$ measures the remaining effect of noise injected at step j. $S(\cdot)$ and $K(\cdot)$ follow the power laws: $S(t) \asymp t^{-2s}$, $K(t) \asymp t^{-(2-2/q)}$, where $s > 0$ and $q > 1$ are task-dependent parameters. Figure 1 (left) shows that this FSL captures the risk dynamics of SignSGD. Leveraging this FSL framework, we furthermore identify two advantages of SignSGD over SGD:

- **Improved data efficiency** (Section 4.1). In the hard-task regime, SignSGD achieves a faster data-scaling rate than SGD, as illustrated in Figure 1 (middle).
- **Improved step efficiency** (Section 4.2). In the easy-task regime, SignSGD supports a larger critical batch size (CBS) than SGD while preserving optimal data efficiency, requiring fewer optimization steps for a fixed data budget, as illustrated in Figure 1 (right).

Empirical validation. PLK regression experiments (Section 5.1) corroborate this mechanism and the resulting scaling predictions. In large language model (LLM) pre-training (Section 5.2), the SignSGD FSL captures the dependence of sign-based optimizers' loss trajectories on batch size more accurately than the original SGD FSL in Li et al. [27].

Related work. Xiao et al. [58] and Compagnoni et al. [11] develop SDE descriptions of signSGD and analyze how the sign operator modifies the drift and noise, revealing effects on preconditioning, risk dynamics, and convergence. More directly related to scaling laws, Kim et al. [22] study signSGD in PLK regression, primarily in the noiseless setting. Our work focuses on the label-noise-dominated regime, establishing an explicit FSL that characterizes signal learning and noise forgetting under general LR schedules, which makes the batch-size dependence explicit and enables comparisons of data and step efficiency between signSGD and SGD. Paquette et al. [43] study Muon and SignSGD in matrix-valued least squares and show that these sign-based optimizers can provide square-root spectral preconditioning effects. Additional related work is discussed in Appendix A.

2 Preliminaries

Notation. We use $\eqsim$ to denote equivalence up to a constant factor, and $\lesssim$ (resp. $\gtrsim$) to indicate 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.

2.1 Problem setup

Power-law kernel (PLK) regression with diagonalized features. Consider a kernel regression task in which labels are generated by $y = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta}^* \rangle + \epsilon$, where data input $\mathbf{x} \sim \mathcal{D}$ and label noise $\epsilon \sim$

$\mathcal{N}(0, \sigma^2)$ are independent. Here $\phi : \mathcal{X} \to \mathbb{R}^d$ is the feature mapping and $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the unknown target parameter. We consider the infinite-dimensional setting, i.e., $d = \infty$. We assume Gaussian features $\boldsymbol{\phi}(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{H})$, where $\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top]$ is the feature covariance. The prediction model $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta} \rangle$ is trained by minimizing the population risk $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y}[\ell(\boldsymbol{\theta}; \mathbf{x}, y)]$ under the squared loss $\ell(\boldsymbol{\theta}; \mathbf{x}, y) = \frac{1}{2}(f(\mathbf{x}; \boldsymbol{\theta}) - y)^2$. The population risk decomposes as $\mathcal{R}(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{\theta}) + \sigma^2/2$, where $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ is the excess risk.

Assumption 2.1 (Constant label noise)**.** Assume $\sigma \eqsim 1$.

Assumption 2.2 (Diagonal features)**.** Assume $\mathbf{H} = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$.

Assumption 2.3 (Power-law structures)**.** For some $q > 1$ and $s > 0$, $\lambda_j \eqsim j^{-q}$, $\lambda_j |\theta_j^*|^2 \eqsim j^{-(1+sq)}$.

Assumption 2.3 specifies the power-law source and capacity structure of the regression. Here q is the **spectral-decay exponent**, which controls the decay rate of the feature spectrum, and s is the **source exponent**, which describes how the target energy is distributed across spectral directions.

One-pass SignSGD. At each training step $k = 0, \ldots, N-1$, we draw a fresh mini-batch $\mathcal{B}_k = \{(\mathbf{x}_{k,i}, y_{k,i})\}_{i=1}^B$ of size B, with all samples drawn i.i.d. from the data distribution. Starting from $\boldsymbol{\theta}_0 = \mathbf{0}$, one-pass SignSGD performs the updates

$$\mathbf{g}_k = \frac{1}{B}\sum_{i=1}^{B} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_k; \mathbf{x}_{k,i}, y_{k,i}), \qquad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \,\mathrm{sgn}(\mathbf{g}_k), \tag{1}$$

where $\mathrm{sgn}(\cdot)$ acts coordinate-wise and $\{\eta_k\}_{k=0}^{N-1}$ is the LR schedule.

2.2 An SDE modeling for SignSGD

We extend the intrinsic-time SDE developed for SGD by Li et al. [27] to SignSGD.

Modeling gradient noise. Write $\mathbf{g}_k = \boldsymbol{\mu}_k + \boldsymbol{\xi}_k$, where $\boldsymbol{\mu}_k := \nabla\mathcal{R}(\boldsymbol{\theta}_k)$ is the population gradient and $\boldsymbol{\xi}_k$ is the gradient noise satisfying $\mathbb{E}[\boldsymbol{\xi}_k] = \mathbf{0}$. We model $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \sigma_B^2 \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a single-sample covariance and $\sigma_B = 1/\sqrt{B}$ characterizes the dependence on batch size. An extension beyond Gaussian noise is discussed in Appendix C.5.

Smoothed sign and update noise. For batch size B, define the scaled smoothed sign operator

$$\mathbf{m}_B(\mathbf{x}) := \sigma_B \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}[\mathrm{sgn}(\mathbf{x} + \sigma_B\boldsymbol{\xi})]$$

Since $\mathrm{sgn}(a\mathbf{x}) = \mathrm{sgn}(\mathbf{x})$ for $a > 0$, the Gaussian noise model gives $\mathbb{E}[\mathrm{sgn}(\mathbf{g}_k)] = \mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B$. Define the centered sign noise $\mathbf{n}_k := \mathrm{sgn}(\mathbf{g}_k) - \mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B$ and its covariance $\mathbf{Q}_B(\boldsymbol{\theta}_k) := \mathbb{E}[\mathbf{n}_k\mathbf{n}_k^\top]$. The update then becomes $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta_k[\mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B + \mathbf{n}_k]$. Its conditional mean and covariance are $-\eta_k \mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B$ and $\eta_k^2 \mathbf{Q}_B(\boldsymbol{\theta}_k)$, respectively, which specify the drift and diffusion covariance of the effective continuum model.

Intrinsic-time dynamics. To analyze the effects of the LR schedule and batch size, we introduce two key quantities: **intrinsic time** and **noise intensity**. Let z denote continuous training steps and $\eta(z)$ a continuous LR schedule satisfying $\eta(k) = \eta_k$. We define the intrinsic time of SignSGD as $t = \Gamma_B(z) := \int_0^z \frac{\eta(u)}{\sigma_B}\,\mathrm{d}u$, which is the continuous analogue of $\sigma_B^{-1}\sum_{j=0}^{k-1}\eta_j$. The corresponding noise intensity is defined by $n_B(t) := \sigma_B \eta(\Gamma_B^{-1}(t))$, which measures the noise strength injected per unit intrinsic time. With these two quantities, we show that SignSGD can be modeled by the SDE

$$\mathrm{d}\boldsymbol{\vartheta}_t = -\mathbf{m}_B(\nabla\mathcal{R}(\boldsymbol{\vartheta}_t))\,\mathrm{d}t + \sqrt{n_B(t)}\,\mathbf{Q}_B^{1/2}(\boldsymbol{\vartheta}_t)\,\mathrm{d}\mathbf{W}_t, \tag{2}$$

where $\boldsymbol{\vartheta}_t$ denotes the continuum parameter process with $\boldsymbol{\vartheta}_0 = \mathbf{0}$, and $\mathbf{W}_t$ is a standard Brownian motion. The function $n_B(t)$ controls the noise injection per unit intrinsic time, with instantaneous diffusion covariance $n_B(t)\mathbf{Q}_B(\boldsymbol{\vartheta}_t)$. Compagnoni et al. [11] derive a related SDE for constant learning rates. Our formulation incorporates a learning rate schedule and the intrinsic-time transformation $t = \Gamma_B(z)$. Appendix C.3 provides the derivation and the explicit coordinate-wise dynamics for the PLK regression model.

Remark 2.4 (Comparison with SGD). The intrinsic time of SGD is $t_{\mathrm{sgd}}(z) = \int_0^z \eta(u)\,\mathrm{d}u$ [27, 30]. For SignSGD, we include an additional factor $1/\sigma_B$ to account for the batch-size dependence of its effective drift. Section 6 explains the mechanism underlying this choice.

Assumption 2.5 (Curvature-aligned noise)**.** Assume $\boldsymbol{\Sigma} \eqsim \mathbf{H}$.

Assumption 2.5 states that the noise variance along each direction scales with the corresponding curvature, i.e., $\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} \eqsim \mathbf{v}^\top \mathbf{H} \mathbf{v}$ for all $\mathbf{v}$, with comparison constants independent of $\mathbf{v}$. Such noise–curvature alignment has been studied theoretically and observed empirically in deep learning [55]. In our PLK regression setting, the exact single-sample gradient-noise covariance at $\boldsymbol{\theta}_k$ is comparable to $(\mathcal{E}(\boldsymbol{\theta}_k) + \sigma^2)\mathbf{H}$ (see Appendix C.1). In the label-noise-dominated regime $\mathcal{E}(\boldsymbol{\theta}_k) \lesssim \sigma^2 \eqsim 1$, this covariance is therefore comparable to $\mathbf{H}$, motivating the assumption in our Gaussian noise model.

3 Functional scaling law for SignSGD

Theorem 3.1 (FSL for SignSGD)**.** *Suppose Assumptions 2.1, 2.2, 2.3, and 2.5 hold. Consider the SDE (2) with* $\sqrt{B}\sup_z \eta(z) \lesssim 1$. *For* $t \gtrsim B^{\max\{1/2,\, 1/(2s+2/q)\}}$, *it holds that*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] \eqsim S(t) + \int_0^t K(t-\tau) n_B(\tau)\, \mathrm{d}\tau, \tag{3}$$

where

$$S(t) \asymp t^{-2s}, \qquad K(t) \asymp t^{-(2-2/q)}. \tag{4}$$

Here the hidden constants are independent of B and the LR schedule.

The proof is provided in Appendix D. Section 6 explains the underlying mechanism and derives the FSL from the effective preconditioned dynamics. The FSL separates the remaining signal error $S(t)$ from the accumulated effect of update noise. Here $n_B(\tau)$ is the noise injection intensity per unit intrinsic time at time τ, while $K(t-\tau)$ measures the residual risk from that injection after elapsed time $t-\tau$. The LR schedule enters through both the intrinsic time and n_B, allowing the FSL to describe the risk trajectory under different schedules.

Li et al. [27] establish an SGD FSL with the same convolution form as (3). For SGD, the intrinsic time, noise intensity and FSL components are

$$t = \Gamma_{\mathrm{sgd}}(z) := \int_0^z \eta(u)\, \mathrm{d}u, \quad n_B(t) = \frac{\eta(\Gamma_{\mathrm{sgd}}^{-1}(t))}{B}, \quad S(t) \asymp t^{-s}, \quad K(t) \asymp t^{-(2-1/q)}.$$

Signal learning and noise forgetting. In their respective intrinsic times, SignSGD has a faster signal decay, t^{-2s} compared with t^{-s} for SGD, but a slower forgetting kernel, $t^{-(2-2/q)}$ compared with $t^{-(2-1/q)}$. Thus, SignSGD learns the target signal faster while retaining the effects of past noise for longer.

The effect of batch size. For the same training steps and LR schedule, the intrinsic times satisfy $t_{\mathrm{sgn}} = \sqrt{B}\, t_{\mathrm{sgd}}$. Unlike SGD, whose intrinsic time is independent of B, SignSGD has an essentially longer training time as B increases. A larger batch also allows more forgetting of earlier noise injections. Moreover, the noise intensity scales as $1/\sqrt{B}$ for SignSGD and $1/B$ for SGD.

Optimizer	**Intrinsic time**	**Noise intensity**	$S(t)$	$K(t)$
SGD [27]	$\int_0^z \eta(\tau)\, \mathrm{d}\tau$	$1/B$	t^{-s}	$t^{-(2-1/q)}$
SignSGD (**ours**)	$\sqrt{B}\int_0^z \eta(\tau)\, \mathrm{d}\tau$	$1/\sqrt{B}$	t^{-2s}	$t^{-(2-2/q)}$

4 When and how does SignSGD outperform SGD?

The FSL enables the comparison of scaling efficiency between SignSGD and SGD. For a data budget $D = NB$, we characterize the achievable excess risk and determine how increasing the batch size B can reduce the number of training steps N without sacrificing data efficiency. Following Li et al. [30], we consider the power-decay LR schedule

$$\eta(z) = \eta_{\mathrm{peak}} \left(1 - \frac{z}{N}\right)^p, \qquad 0 \leqslant z \leqslant N, \tag{5}$$

where $\eta_{\mathrm{peak}} > 0$ and $p > 0$ denote the peak learning rate and decay exponent, respectively. This schedule family attains the optimal data-scaling rates characterized below.

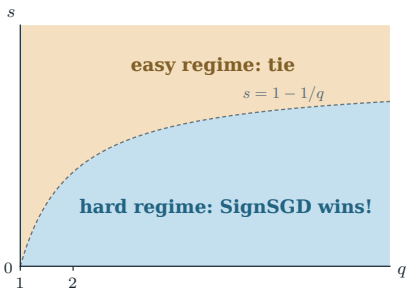
4.1 Data efficiency

We first ask: *what is the optimal final-step excess risk that SignSGD can achieve by tuning the LR schedule at a fixed data budget D?* We consider $B = 1$, with the case of $B > 1$ deferred to Section 4.2. For the power-decay schedule (5), the final expected excess risk is $\mathcal{E}_D := \mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_{\Gamma_1(D)})]$, where $\Gamma_1(D) = \int_0^D \eta(\tau)\,\mathrm{d}\tau$. The following theorem specifies choices of η_{peak} and p and the resulting data-scaling rates.

Theorem 4.1 (Data scaling of SignSGD)**.** Under the assumptions of Theorem 3.1, let $B = 1$ and use the power-decay schedule (5). Then

- If $s \geqslant \frac{1}{2} - \frac{1}{q}$, then choosing $p > \frac{q}{2} - 1$ and $\eta_{\mathrm{peak}} \eqsim D^{-\frac{2s-1+2/q}{2s+2/q}}$ yields $\mathcal{E}_D \eqsim D^{-\frac{sq}{sq+1}}$.
- If $0 < s < \frac{1}{2} - \frac{1}{q}$, then choosing $p > \frac{2s}{1-2s}$ and $\eta_{\mathrm{peak}} \eqsim 1$ yields $\mathcal{E}_D \eqsim D^{-2s}$.

Its proof is given in Appendix E.2. We show that the rates achieved in Theorem 4.1 are optimal among all admissible LR schedules (see details in Appendix E.4). Thus, the optimal data-scaling rate of SignSGD can be written as $D^{-\min\{2s,\, sq/(sq+1)\}}$. For comparison, the optimal SGD rate under its FSL is $D^{-\min\{s,\, sq/(sq+1)\}}$ [30] (Appendix E.3). Thus, these data-scaling rates partition the (s, q) space into three regimes: easy, hard, and super hard The phase diagram on the right summarizes these regimes, with boundaries at $s = 1 - 1/q$ and $s = 1/2 - 1/q$ The corresponding SGD and SignSGD rates are $D^{-\gamma_{\mathrm{sgd}}}$ and $D^{-\gamma_{\mathrm{sgn}}}$, where



$$(\gamma_{\mathrm{sgd}}, \gamma_{\mathrm{sgn}}) = \begin{cases} (\frac{sq}{sq+1}, \frac{sq}{sq+1}), & s \geqslant 1 - \frac{1}{q}, \\ (s, \frac{sq}{sq+1}), & \frac{1}{2} - \frac{1}{q} \leqslant s < 1 - \frac{1}{q}, \\ (s, 2s), & 0 < s < \frac{1}{2} - \frac{1}{q}. \end{cases}$$

Remark 4.2 (Connection to statistical lower bounds.). Classical results give the minimax rate $D^{-sq/(sq+1)}$ for the regression problem with the same spectral-decay and source conditions [7, 49]. This rate is established in the discrete setting and it matches the optimal rate in Theorem 4.1 under our continuous SDE model when $s \geqslant \frac{1}{2} - \frac{1}{q}$.

4.2 Step efficiency

At a fixed data budget $D = NB$, increasing the batch size reduces the number of training steps. We therefore ask two questions: *(i) Can we increase the batch size while preserving each optimizer's small-batch risk? (ii) If so, how large can the batch size be?* The critical batch size (CBS) characterizes the largest such batch size, with the LR schedule allowed to be retuned.

Definition 4.3 (Critical batch size)**.** Let $\mathcal{A}_{D,B}$ denote the admissible LR schedules for a given optimizer, and define $\mathcal{E}_\star(D, B) := \inf_{\eta \in \mathcal{A}_{D,B}} \mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_{\Gamma_B(D/B)})]$. For a fixed constant $\delta > 0$, we define

$$B_{\mathrm{crit}}(D) := \sup\{B \in \{1, \ldots, D\} : \mathcal{E}_\star(D, B) \leqslant (1+\delta)\,\mathcal{E}_\star(D, 1)\}, \qquad N_{\mathrm{crit}}(D) := \frac{D}{B_{\mathrm{crit}}(D)}.$$

Here Γ_B denotes the optimizer's intrinsic-time map for the chosen schedule η, and $N_{\mathrm{crit}}(D)$ is the minimum number of steps needed to preserve the optimal small-batch risk within the factor $1 + \delta$.

For SGD, admissibility requires $\eta_{\mathrm{peak}} \lesssim 1$. For SignSGD, it requires $\eta_{\mathrm{peak}}\sqrt{B} \lesssim 1$. We consider $0 < s < 3/2 - 1/q$, where the SignSGD training-time condition is compatible with the batch-size ranges below. The boundary $s = 3/2 - 1/q$ and the *super-easy regime* $s > 3/2 - 1/q$ are discussed in Appendix F. For notational convenience, we write $\alpha = s + \frac{1}{q}$.

Theorem 4.4 (Critical batch size)**.** *Suppose that the assumptions of Theorem 3.1 hold and that $0 < s < 3/2 - 1/q$. Then the critical batch sizes satisfy*

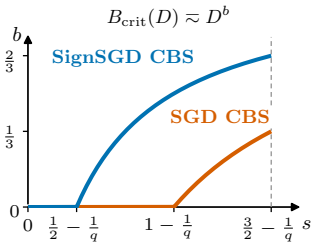
$$B_{\mathrm{crit}}^{\mathrm{sgd}}(D) \eqsim \max\left\{1, D^{1-\frac{1}{\alpha}}\right\}, \qquad B_{\mathrm{crit}}^{\mathrm{sgn}}(D) \eqsim \max\left\{1, D^{1-\frac{1}{2\alpha}}\right\}. \tag{6}$$

Peak LR scaling. The batch-size ranges in Theorem 4.4 are attained by the power-decay schedule (5) with $N = D/B$, any fixed $p > q - 1$, and peak learning rates

$$\eta^{\mathrm{sgd}}_{\mathrm{peak}} \eqsim \min\left\{B\, D^{-1+\frac{1}{\alpha}},\, 1\right\}, \qquad \eta^{\mathrm{sgn}}_{\mathrm{peak}} \eqsim \min\left\{\sqrt{B}\, D^{-1+\frac{1}{2\alpha}},\, B^{-1/2}\right\}.$$

In the easy-task regime, the linear scaling rule for SGD and the square-root scaling rule for SignSGD preserve the small-batch risk as B increases. Their admissibility constraints, $\eta^{\mathrm{sgd}}_{\mathrm{peak}} \lesssim 1$ and $\sqrt{B}\,\eta^{\mathrm{sgn}}_{\mathrm{peak}} \lesssim 1$, determine the batch-size ranges over which these rules apply.

Efficiency gains across task regimes. The figure on the right compares the CBS exponents b in $B_{\mathrm{crit}}(D) \eqsim D^b$. In the easy-task regime, $1 - 1/q \leqslant s < 3/2 - 1/q$, both optimizers attain excess risk of order $D^{-s/\alpha}$. At their respective CBSs, SGD requires $N^{\mathrm{sgd}}_{\mathrm{crit}} \eqsim D^{1/\alpha}$ updates, whereas SignSGD requires only $N^{\mathrm{sgn}}_{\mathrm{crit}} \eqsim D^{1/(2\alpha)}$. Thus, SignSGD uses a factor $D^{1/(2\alpha)}$ fewer sequential updates while maintaining the same data-scaling rate. For hard tasks, $\frac{1}{2} - \frac{1}{q} < s < 1 - \frac{1}{q}$, SignSGD achieves both a faster data-scaling rate and a CBS that grows with D, whereas SGD has $B^{\mathrm{sgd}}_{\mathrm{crit}}(D) \eqsim 1$. These CBSs preserve each optimizer's small-batch risk, so SignSGD improves both data and step efficiency. For super-hard tasks, $s \leqslant \frac{1}{2} - \frac{1}{q}$, both CBSs are of constant order, requiring $N \eqsim D$ updates to preserve their respective small-batch risks. SignSGD's advantage in this regime therefore lies in data efficiency rather than step efficiency.



5 Experiments

We evaluate the predictive performance of FSL in PLK regression and language model pre-training. The FSL for SignSGD motivates the parameterized form

$$L_\Theta(k) = L_\infty + S_k + N_k, \qquad S_k = C_S(A_S + t_k)^{-\alpha}, \qquad N_k = C_K \sum_{j=0}^{k-1} (A_K + t_k - t_j)^{-\nu} \eta_j^2, \quad (7)$$

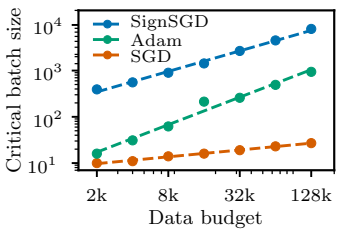
where L_∞ is the irreducible loss, S_k is the signal term, and N_k is the noise term. The trainable parameters are $\Theta = (L_\infty, C_S, A_S, \alpha, C_K, A_K, \nu)$, and $t_k = \sqrt{B}\sum_{j=0}^{k-1} \eta_j$ is the intrinsic time associated with the LR schedule $(\eta_0, \ldots, \eta_{N-1})$. Further details are provided in Appendix H.1.

5.1 PLK regression

FSL accurately captures risk dynamics. Although our FSL is derived from the continuous-time model (35), Figure 1 (left) shows that the resulting FSL also accurately describes discrete-time training. For PLK regression with $(s, q) = (0.3, 4)$, we show accurate fitting and prediction of SignSGD risk curves across cosine, WSD, and cyclic schedules.

Signal–noise decomposition and intrinsic-time collapse. Using the experiments in Figure 1 (left), we decompose the risk trajectories under the three LR schedules into the noise term N_k and signal term S_k, shown in the left and middle panels of Figure 2, respectively. Although the signal trajectories differ in training steps, they collapse onto a common curve when plotted against intrinsic time in the right panel. The FSL thus separates schedule-dependent noise effects from a shared signal-learning profile governed by intrinsic time.

Data scaling and critical batch size. We validate the data-scaling and batch-size predictions in Sections 4.1 and 4.2. On the hard task with $(s, q) = (0.3, 4)$, we tune the peak learning rate for each optimizer and data budget and compare the best final excess risk. Figure 1 (middle) shows that SignSGD achieves a faster data-scaling rate than SGD, consistent with Theorem 4.1. On the easy task with $(s, q) = (1, 4)$, we sweep batch sizes at each data budget and tune the peak learning rate separately for every optimizer and batch size. We estimate the CBS as the largest batch size that approximately preserves each optimizer's small-batch risk. The right panel shows that SignSGD has the largest CBS, followed by Adam and SGD, throughout the tested range. The power-law fits



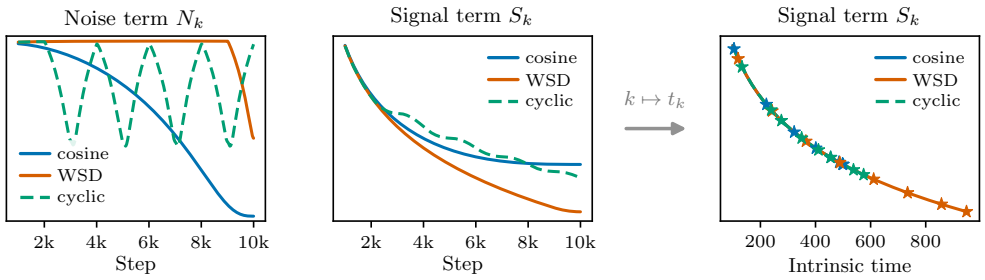


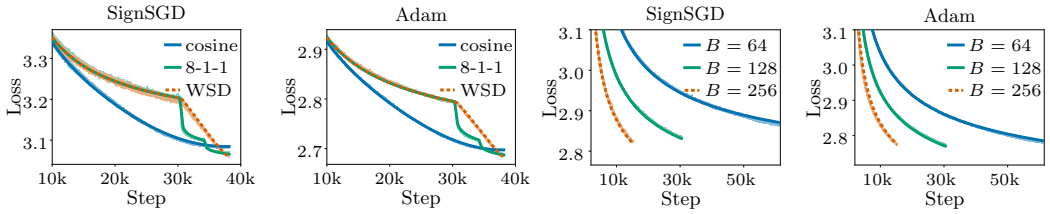
Figure 2: **FSL decomposition and intrinsic-time collapse for SignSGD.** We use the same configurations and shared FSL as in Figure 1 (left). Solid curves correspond to the cosine and WSD schedules used for fitting; dashed curves show predictions for the held-out cyclic schedule. **(Left)** Noise term N_k versus training steps. **(Middle)** Signal term S_k versus training steps. **(Right)** The signal terms collapse onto a common curve when plotted against intrinsic time.

further indicate that the CBS grows faster with the data budget for SignSGD and Adam than for SGD. These findings support the predicted step efficiency of SignSGD and its heuristic extension to Adam. Similar trends appear in language-model training: Muon exhibits greater large-batch robustness than Adam, which in turn outperforms SGD [38, 48, 51]. Since Muon applies matrix sign operator to gradient updates, these observations suggest a broader connection between sign-based normalization and batch-size efficiency.

5.2 Language model pre-training

We train a Llama-style language model [52] with approximately 0.1B parameters on 10B tokens, using a realistic token-to-parameter ratio of approximately 100. We evaluate the FSL for SignSGD and Adam under different LR schedules, including cosine, WSD [17], and the 8-1-1 schedule [12]. Additional experimental details are provided in Appendix H.

FSL accurately captures loss dynamics across optimizers. We evaluate the predictive power of FSL across different LR schedules and optimizers. Specifically, we extend the Adam experiments of Li et al. [27] to SignSGD. For each optimizer, we fit the FSL parameters using the loss curves of cosine and 8-1-1 schedules, and then use the fitted FSL to predict the loss curves under the WSD schedule. The results for SignSGD and Adam are reported in Figure 3 (a). We observe that FSL accurately captures the loss curves for both optimizers and yields accurate predictions for the held-out schedule.



(a) Cross-schedule prediction (b) Batch-size extrapolation

Figure 3: **FSL captures loss dynamics and batch-size dependence in practical language model pre-training. (a)** Cross-schedule loss prediction for SignSGD (left) and Adam (right). For each optimizer, we fit the FSL parameters Θ to the loss curves under the cosine and 8-1-1 schedules and predict the loss curve under the held-out WSD schedule. **(b)** Batch-size extrapolation for SignSGD (left) and Adam (right). With a constant learning rate, we fit Θ to the loss curves at $B = 64$ and $B = 128$ and predict the loss curve at $B = 256$. Solid and dashed FSL curves indicate fits and predictions, respectively.

FSL accurately captures batch-size dependence and enables extrapolation. A key difference between the FSL in (7) and the SGD-based FSL of Li et al. [27] lies in the batch-size scaling. We test this batch-size dependence in practical training. We train with batch sizes $B \in \{64, 128, 256\}$ and a constant learning rate. For each optimizer, we fit the FSL using the curves for $B = 64$ and $B = 128$, and then predict the held-out curve for $B = 256$. As shown in Figure 3 (b), the FSL jointly fits the two batch sizes and extrapolates well to the larger batch size for both SignSGD

and Adam. An additional experiment with cosine schedule for Adam is reported in Appendix H.3, Figure 12 (left).

6 Why does SignSGD outperform SGD?

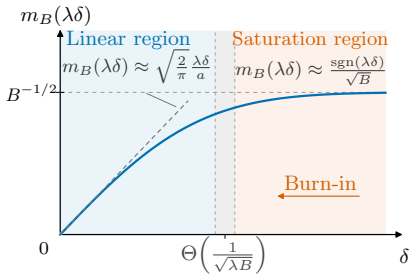
6.1 Smoothed sign and one-dimensional dynamics

Smoothed sign. We first examine the drift and diffusion in (2) in one dimension. Recall that $\sigma_B = 1/\sqrt{B}$. Let $a > 0$ denote the single-sample gradient-noise standard deviation, so the mini-batch noise standard deviation is $\sigma_B a$. The scaled smoothed sign is $m_B(x) = \sigma_B \mathbb{E}_{\xi\sim\mathcal{N}(0,a^2)}[\operatorname{sgn}(x + \sigma_B\xi)]$. Writing $c = \sqrt{2/\pi}$, its two limiting responses are

$$m_B(x) \approx \begin{cases} c\,x/a, & |x|/(\sigma_B a) \ll 1, \\ \sigma_B \operatorname{sgn}(x), & |x|/(\sigma_B a) \gg 1. \end{cases} \tag{8}$$

The variance of the corresponding sign update is $Q_B(x) = 1 - m_B(x)^2/\sigma_B^2$. Thus, the noise-dominated region has approximately linear drift and $Q_B(x) \approx 1$, whereas the signal-dominated region has saturated drift and $Q_B(x) \approx 0$.

One-dimensional dynamics. For a coordinate with curvature λ, let $\delta_t = \vartheta_t - \theta^*$, so the population gradient is $\lambda\delta_t$. Applying (8) to $x = \lambda\delta$ reveals two dynamical regimes. For $\lambda|\delta|/(\sigma_B a) \gg 1$, the response saturates at $\sigma_B \operatorname{sgn}(\delta)$ and $Q_B(\lambda\delta) \approx 0$, giving approximately deterministic motion toward zero. For $\lambda|\delta|/(\sigma_B a) \ll 1$, the response is linear, $m_B(\lambda\delta) \approx c\lambda\delta/a$, and $Q_B(\lambda\delta) \approx 1$. The local dynamics therefore have the drift of SGD preconditioned by c/a, with approximately unit sign-noise variance. The crossover occurs at $|\delta| \eqsim \sigma_B a/\lambda$. Under noise-curvature alignment $a^2 \eqsim \lambda$, this scale becomes $\sigma_B/\sqrt{\lambda} = (\lambda B)^{-1/2}$, as illustrated on the right for $\delta \geqslant 0$.



These two limits describe the saturated dynamics during burn-in and the local linear dynamics underlying the scaling regime, respectively. Substituting them into (2) gives

$$\mathrm{d}\delta_t \approx \begin{cases} -\sigma_B \operatorname{sgn}(\delta_t)\ \mathrm{d}t, & \dfrac{\lambda|\delta_t|}{\sigma_B a} \gg 1 \quad \text{(burn-in phase)}, \\ -c\,\dfrac{\lambda}{a}\delta_t\ \mathrm{d}t + \sqrt{n_B(t)}\ \mathrm{d}W_t, & \dfrac{\lambda|\delta_t|}{\sigma_B a} \ll 1 \quad \text{(local phase)}. \end{cases} \tag{9}$$

From burn-in to local phase. We refer to the initial transient before entering the linear region as *burn-in*. The first branch of (9) describes its saturated part. Under zero initialization, this approximation gives $|\delta_t| \approx |\theta^*| - \sigma_B t$ until the transition region is reached. The source condition implies $|\theta_j^*| \eqsim j^{(q-1-sq)/2}$. For easy tasks, $s + 1/q \geqslant 1$, the initial coordinate errors are uniformly bounded, giving a saturated burn-in time of at most order $\sigma_B^{-1} = \sqrt{B}$. For hard tasks, $s + 1/q < 1$, these errors can grow with the coordinate index, and controlling burn-in across the relevant spectral modes requires $t \gtrsim \sigma_B^{-1/(s+1/q)} = B^{1/(2s+2/q)}$. Together, these motivate the training-time condition $t \gtrsim B^{\max\{1/2,\,1/(2s+2/q)\}}$ in Theorem 3.1. This local analysis provides intuition for the scaling regime. The rigorous argument in Appendix D uses a Lyapunov function to establish global moment bounds for the nonlinear dynamics, allowing noise-driven excursions outside the local linear region.

6.2 Noise-induced preconditioning

From coordinate-wise normalization to preconditioning. The local linear response in the preceding subsection extends coordinate-wise to the multidimensional dynamics. Let $\boldsymbol{\delta}_t = \boldsymbol{\vartheta}_t - \boldsymbol{\theta}^*$ and $\mathbf{D} = \operatorname{diag}(\Sigma_{11}, \Sigma_{22}, \ldots)$. Linearizing the smoothed sign around the origin gives $\mathbf{m}_B(\mathbf{H}\boldsymbol{\delta}_t) \approx c\,\mathbf{D}^{-1/2}\mathbf{H}\boldsymbol{\delta}_t$. Thus, the local drift takes the form $-\mathbf{P}\nabla\mathcal{R}(\boldsymbol{\vartheta}_t)$, where $\mathbf{P} := c\,\mathbf{D}^{-1/2}$. This is the drift of preconditioned SGD: the gradient is normalized by its single-sample noise scale on each coordinate. Such normalization does not by itself guarantee acceleration; its effectiveness depends on the relation between the noise scales and the curvature.

Structural conditions for effective preconditioning. Noise-curvature alignment and diagonal features play important roles in SignSGD's scaling advantage. The former ensures that $\Sigma_{jj} \eqsim H_{jj}$, so the preconditioner compensates for coordinate-wise curvature. Diagonal features further ensure that these coordinates coincide with the eigendirections of $\mathbf{H}$. Together, they give $a_j := \sqrt{\Sigma_{jj}} \eqsim \sqrt{\lambda_j}$ and effective decay rates $\rho_j := \frac{c\lambda_j}{a_j} \eqsim \sqrt{\lambda_j}$. The effective spectrum therefore decays as $j^{-q/2}$, rather than j^{-q}, reducing the disparity between fast and slow spectral directions. Without noise-curvature alignment, noise normalization need not compensate for curvature. Without diagonal features, coordinate-wise preconditioning need not produce a square-root transformation of the Hessian spectrum.

Effective preconditioned SGD. Under our diagonal Gaussian noise model, the sign-noise coordinates are conditionally independent, and their variances approach one in the local regime. Hence $\mathbf{Q}_B \approx \mathbf{I}$, yielding

$$\mathrm{d}\boldsymbol{\delta}_t \approx -\mathbf{P}\mathbf{H}\boldsymbol{\delta}_t \, \mathrm{d}t + \sqrt{n_B(t)} \, \mathrm{d}\mathbf{W}_t, \qquad \mathbf{P} = c\,\mathbf{D}^{-1/2}. \tag{10}$$

This has the form of preconditioned SGD with normalized gradient noise. Under noise-curvature alignment and diagonal features, $\mathbf{P}\mathbf{H} = \mathrm{diag}(\rho_j)$ with $\rho_j \eqsim \sqrt{\lambda_j}$. The local dynamics therefore combine square-root spectral decay rates with approximately isotropic noise injection. These two changes determine the signal-learning and noise-forgetting profiles derived next.

Empirical evidence for the structural conditions. Figure 4 illustrates the role of these two structural ingredients at a fixed data budget. With matched total gradient noise energy and separately tuned learning rates, SignSGD attains a substantially lower final risk in the setting with diagonal features and curvature-aligned noise, while rotations (breaking diagonal structure) or isotropic noise (breaking noise-curvature alignment) reduce this advantage.

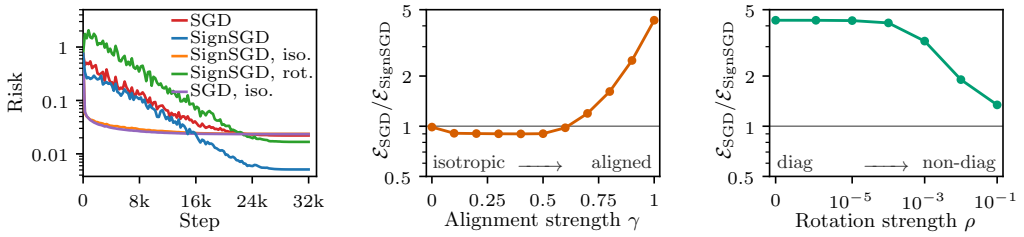


Figure 4: **Structural conditions for the SignSGD advantage. (Left)** Risk dynamics of SGD and SignSGD with curvature-aligned noise, isotropic-noise controls, and a rotated feature covariance. **(Middle)** We vary noise-curvature alignment by setting the noise covariance to $\boldsymbol{\Sigma} \propto \mathbf{H}^{\gamma}$ with $\gamma \in [0, 1]$. SignSGD shows an advantage only when γ is close to 1. **(Right)** We vary feature diagonality by rotating the feature covariance, with ρ controlling the rotation strength. As ρ increases, the covariance becomes less diagonal and the SignSGD advantage weakens. All experiments use optimally tuned learning rates.

6.3 Deriving the functional scaling law

Spectral representation. The effective dynamics (10) have diagonal drift $\mathbf{P}\mathbf{H} = \mathrm{diag}(\rho_j)$, where $\rho_j = c\lambda_j/a_j \eqsim \sqrt{\lambda_j}$. For this linear model initialized at $\boldsymbol{\delta}_0 = -\boldsymbol{\theta}^*$, the expected excess risk decomposes as

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] \approx S(t) + \int_0^t K(t-\tau) n_B(\tau) \, \mathrm{d}\tau, \tag{11}$$

where

$$S(t) = \frac{1}{2}\sum_j \lambda_j |\theta_j^*|^2 e^{-2\rho_j t}, \qquad K(t) = \frac{1}{2}\sum_j \lambda_j e^{-2\rho_j t}.$$

The signal profile measures the remaining initial error, while the forgetting kernel measures the residual risk from a unit noise injection after elapsed intrinsic time t. Both are mixtures of exponential decays with rates ρ_j, but their spectral weights differ. These expressions describe the effective linear model; they do not assume that the original dynamics are linear throughout burn-in.

From exponential modes to power laws. Each spectral mode has characteristic time $\tau_j := \rho_j^{-1} \eqsim j^{q/2}$. At time t, modes with $\tau_j \ll t$ have largely decayed, whereas those with $\tau_j \gg t$ retain substantial contributions. Therefore, we introduce the **moving frontier** j_t to characterize which modes are learned at time t. It is determined by $\tau_{j_t} \eqsim t$, and thus satisfies $j_t \eqsim t^{2/q}$. To estimate the signal profile, split the spectrum at this cutoff:

$$2S(t) = \underbrace{\sum_{j \leqslant j_t} \lambda_j |\theta_j^*|^2 e^{-2\rho_j t}}_{\text{decaying modes}} + \underbrace{\sum_{j > j_t} \lambda_j |\theta_j^*|^2 e^{-2\rho_j t}}_{\text{remaining spectral tail}}. \tag{12}$$

For $j > j_t$, we have $\rho_j t \lesssim 1$, so the exponential factor is bounded above and below by positive constants. The tail contribution is therefore $\eqsim \sum_{j>j_t} j^{-(1+sq)} \eqsim j_t^{-sq}$. For $j \leqslant j_t$, the bound $e^{-u} \lesssim u^{-p}$, with any fixed $p > 2s$, gives

$$\sum_{j \leqslant j_t} \lambda_j |\theta_j^*|^2 e^{-2\rho_j t} \lesssim t^{-p} \sum_{j \leqslant j_t} j^{-1-sq+pq/2} \eqsim t^{-p} j_t^{pq/2-sq} \eqsim j_t^{-sq}.$$

Thus, the scaling behavior is determined by j_t^{-sq}. Applying the same argument to K yields

$$S(t) \eqsim j_t^{-sq} \eqsim t^{-2s}, \qquad K(t) \eqsim j_t^{1-q} \eqsim t^{-(2-2/q)} \tag{13}$$

for sufficiently large t. Substitution into (11) recovers the FSL structure. Appendix D controls the nonlinear dynamics to establish the FSL of Theorem 3.1.

Faster signal learning, slower noise forgetting. The same spectral-cutoff argument explains the difference between SGD and SignSGD. In their respective intrinsic times, SGD has decay rates of order λ_j and cutoff $j_t \eqsim t^{1/q}$, whereas SignSGD has rates of order $\sqrt{\lambda_j}$ and cutoff $j_t \eqsim t^{2/q}$. The faster advance through the spectrum changes the signal-learning function from t^{-s} to t^{-2s}. Applying the same argument to the forgetting kernels gives decay rates $t^{-(2-1/q)}$ for SGD and $t^{-(2-2/q)}$ for SignSGD, showing that SignSGD forgets noise more slowly. Thus, the preconditioning effect accelerates signal learning while slows the noise forgetting.

The effect of batch size. The comparison above uses each method's own intrinsic time. For the same LR schedule $\eta(z)$, SGD uses $t_{\mathrm{sgd}}(z) = \int_0^z \eta(\tau)\, \mathrm{d}\tau$, whereas SignSGD uses $t_{\mathrm{sgn}}(z) = \sqrt{B}\, t_{\mathrm{sgd}}(z)$. Consequently, their signal-learning functions expressed in training steps satisfy $S_{\mathrm{sgn}}(z) \eqsim (\sqrt{B}\, t_{\mathrm{sgd}}(z))^{-2s}$ and $S_{\mathrm{sgd}}(z) \eqsim t_{\mathrm{sgd}}(z)^{-s}$. The additional $\sqrt{B}$ factor is from the local response $\mathbb{E}[\mathrm{sgn}(g_j)] \approx c\sqrt{B}\, \lambda_j \delta_j / a_j$: a larger batch size reduces gradient noise and increases the mean sign response in the local regime. Batch size also enters the noise injection intensity through $n_B(t) = \eta(\Gamma_B^{-1}(t))/\sqrt{B}$. Together, the intrinsic-time transformation and the noise injection intensity determine how batch size affects signal learning and noise accumulation in the FSL.

7 Concluding remarks

Our results show that the advantages of SignSGD depend on both task difficulty and the structure of gradient noise. The FSL makes this dependence explicit by tracking the decay of the initial error and the persistence of injected noise. Its mechanistic interpretation links the local response of the sign operator to noise-dependent preconditioning, whose effectiveness relies on diagonal features and noise-curvature alignment. The resulting balance between faster signal learning and slower noise forgetting distinguishes two benefits: better use of data on hard tasks, and fewer sequential updates at comparable data efficiency in the covered easy-task regime. Experiments corroborate these conclusions in kernel regression and show that the SignSGD FSL better describes the effect of batch size on sign-based optimizer trajectories in LLM pre-training than the original SGD FSL.

Our results motivate two directions for future work:

- **Extensions to RMSprop and Adam.** Our derivations predict the same FSL form and batch-size dependence, with supporting numerical evidence (Appendices G and H.2). A rigorous treatment would need to account for momentum, evolving second-moment estimates, and their coupling with gradient noise.
- **From coordinate-wise sign to matrix sign.** Extending the analysis to the matrix sign operator underlying Muon would clarify how matrix structure and gradient noise jointly determine effective preconditioning, signal learning, noise forgetting, and batch-size dependence.

Acknowledgments and Disclosure of Funding

Lei Wu is supported in part by the National Natural Science Foundation of China (NSFC12522120, NSFC92470122, and NSFC12288101). Binghui Li is supported by the Elite Ph.D. Program in Applied Mathematics at Peking University.

References

[1] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024. (cited on page 17)

[2] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. (cited on page 17)

[3] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413. PMLR, 2018. (cited on page 1)

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018. (cited on pages 1 and 17)

[5] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pages 4345–4382. PMLR, 2024. (cited on page 17)

[6] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8): 084002, 2025. (cited on page 17)

[7] 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 5)

[8] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Others. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023. (cited on page 1)

[9] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018. (cited on page 1)

[10] Xiaoran Chen, Binghui Li, Liwei Wang, and Pinyan Lu. When and how does muon precondition? the role of structured gradient noise. *arXiv preprint arXiv:2610.31728*, 2026. (cited on page 17)

[11] Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on pages 2, 3, and 17)

[12] DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping

Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. (cited on pages 7 and 46)

[13] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad, 2020. (cited on page 1)

[14] 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 17)

[15] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024. (cited on page 17)

[16] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012. (cited on page 1)

[17] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and Others. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. (cited on pages 7, 43, and 46)

[18] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *Advances in Neural Information Processing Systems*, 37:110246–110289, 2024. (cited on page 17)

[19] Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. *Advances in Neural Information Processing Systems*, 37:33891–33932, 2024. (cited on page 17)

[20] Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates. *arXiv preprint arXiv:2507.12091*, 2025. (cited on page 17)

[21] 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)

[22] Jihwan Kim, Dogyoon Song, and Chulhee Yun. Scaling laws of signSGD in linear regression: When does it outperform SGD? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zVwRP0ikrx. (cited on pages 2 and 17)

[23] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (cited on page 1)

[24] Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. *arXiv preprint arXiv:2505.19227*, 2025. (cited on pages 1 and 17)

[25] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023. (cited on page 1)

[26] Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *Advances in Neural Information Processing Systems*, 37:30106–30148, 2024. (cited on page 1)

[27] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. In *Advances in Neural Information Processing Systems*, volume 38, 2025. (cited on pages 1, 2, 3, 4, 7, 17, 21, and 35)

[28] Binghui Li, Jianan Wang, Jinbo Wang, Lean Wang, Zilin Wang, and Lei Wu. Scaling-law analysis of SignSGD: From feature-space linear regression to LLM pre-training. In *SciForDL, 2nd edition*, 2026. URL https://openreview.net/forum?id=KQfV97S12I. (cited on page 17)

[29] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026. (cited on pages 1 and 17)

[30] 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 3, 4, 5, 17, 33, 35, and 43)

[31] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024. (cited on page 17)

[32] Licong Lin, Jingfeng Wu, and Peter L Bartlett. Improved scaling laws in linear regression via data reuse. *arXiv preprint arXiv:2506.08415*, 2025. (cited on page 17)

[33] 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)

[34] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx. (cited on pages 43 and 46)

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. (cited on pages 1 and 46)

[36] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35:7697–7711, 2022. (cited on page 17)

[37] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022. (cited on page 17)

[38] Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla SGD works, and why gradient accumulation is wasteful. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (cited on page 7)

[39] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024. (cited on page 17)

[40] Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, and Ard Louis. An exactly solvable model for emergence and scaling laws. *arXiv preprint arXiv:2404.17563*, 2024. (cited on page 17)

[41] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, Heidelberg, 6 edition, 2003. doi: 10.1007/978-3-642-14394-6. (cited on pages 17 and 19)

[42] 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 17)

[43] Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon eclipses SignSGD. *arXiv preprint arXiv:2605.09552*, 2026. URL https://arxiv.org/abs/2605.09552. (cited on page 2)

[44] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. (cited on page 1)

[45] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020. (cited on page 17)

[46] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025. URL https://arxiv.org/abs/2505.23737. (cited on page 1)

[47] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. (cited on page 43)

[48] Teodora Srećković, Jonas Geiping, and Antonio Orvieto. Is your batch size the problem? revisiting the Adam-SGD gap in language modeling. *arXiv preprint arXiv:2506.12543*, 2025. URL https://arxiv.org/abs/2506.12543. (cited on page 7)

[49] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009. (cited on page 5)

[50] Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signSGD under weaker assumptions. In *International Conference on Machine Learning*, pages 33077–33099. PMLR, 2023. (cited on pages 1 and 17)

[51] Benjamin Thérien, Xiaolong Huang, Aaron Defazio, Irina Rish, and Eugene Belilovsky. MuLoCo: Muon is a practical inner optimizer for DiLoCo. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 120419–120462. PMLR, 2026. URL https://proceedings.mlr.press/v306/therien26a.html. (cited on page 7)

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. (cited on page 7)

[53] Jia-Nan Wang, Zixun Huang, Kairui Li, Zilin Wang, and Lei Wu. Momentum in large-batch training: Polyak enlarges the critical batch size, Nesterov improves data efficiency. *arXiv preprint arXiv:2609.02728*, 2026. (cited on page 40)

[54] Jinbo Wang, Binghui Li, Zhanpeng Zhou, Mingze Wang, Yuxuan Sun, Jiaqi Zhang, Xunliang Cai, and Lei Wu. Fast catch-up, late switching: Optimal batch size scheduling via functional scaling laws. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PXWgzUkVwo. (cited on page 17)

[55] 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 20)

[56] 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*, pages 23549–23588. PMLR, 2022. (cited on page 17)

[57] Lei Wu, Mingze Wang, and Weijie Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693, 2022. (cited on page 20)

[58] Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signSGD in high-dimensions: quantifying preconditioning and noise-compression effects. In *Forty-second International Conference on Machine Learning*, 2025. (cited on pages 2 and 17)

[59] Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam Exploits ℓ_∞-Geometry of Loss Landscape via Coordinate-wise Adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.08198. (cited on page 45)

[60] 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. *arXiv preprint arXiv:2511.13421*, 2025. (cited on page 17)

[61] Dechen Zhang, Xuan Tang, Yingyu Liang, and Difan Zou. Scaling laws for precision in high-dimensional linear regression. *arXiv preprint arXiv:2602.19241*, 2026. URL https://arxiv.org/abs/2602.19241. (cited on page 17)

[62] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024. (cited on page 1)

Appendix

Table of contents

A Additional related work **17**

B Background on SDEs **17**

B.1 The generator for expectation estimates 17

B.2 Time-change formula for SDEs 19

C The SDE modeling for SGD and SignSGD **20**

C.1 The structure of gradient noise 20

C.2 The SDE modeling for SGD 21

C.3 The SDE modeling for SignSGD 22

C.4 Properties of the smoothed sign function 23

C.5 An extension to non-Gaussian gradient noise 24

D FSL of SignSGD (Proof of Theorem 3.1) **25**

D.1 A technical lemma: power-law spectral sums 25

D.2 Time rescaling and normalized modal dynamics 26

D.3 The generator and a global lower bound 27

D.4 A Foster–Lyapunov upper bound 28

D.5 Control of the burn-in process 30

D.6 Completion of the proof 31

D.7 Discussion: Initialization and relaxation across spectral modes 32

E Data scaling analysis **33**

E.1 Power decay schedules 33

E.2 Data scaling for SignSGD (Proof of Theorem 4.1) 34

E.3 Data scaling of SGD 35

E.4 Optimality of power decay schedules in data scaling 36

F Batch-size scaling analysis **37**

G Heuristic extension to RMSprop and Adam **38**

G.1 Second moments and coordinate curvatures 38

G.2 The effective RMSprop dynamics 38

G.3 The FSL of RMSprop 39

G.4 The FSL of Adam 40

H Experimental details and additional results **40**

H.1 Fitting and prediction using FSL 40

H.2 PLK regression experiments 42

H.3 Language model experiments 46

A Additional related work

Theoretical analysis of SignSGD. Existing studies have primarily focused on convergence guarantees of SignSGD [4, 10, 19, 20, 50]. Another line of work derives continuous-time stochastic differential equations (SDEs) for modeling SignSGD dynamics [11, 36, 58]. We build on this continuous-time approach, and use it to derive a trajectory-level functional scaling law for SignSGD.

Theoretical studies of scaling laws. A growing literature seeks theoretical explanations for scaling laws [1, 2, 5, 6, 14, 15, 18, 24, 27, 28, 29, 30, 31, 32, 37, 39, 40, 42, 45, 54, 56, 60, 61], but predominantly analyzes vanilla SGD. In contrast, we study SignSGD, which is closer to optimizers used in large-scale practice, and show that it exhibits qualitatively different scaling behavior. Another related work is Kunstner and Bach [24], which studies full-batch sign gradient descent (signGD) in a linear bigram model and identifies scaling properties of deterministic sign dynamics. We instead focus on the stochastic regime, where gradient noise plays a central role and interacts nontrivially with sign normalization to shape the scaling behavior.

Comparison with Xiao et al. [58]. Xiao et al. [58] derive a high-dimensional SDE and a deterministic risk ODE for SignSGD, quantifying diagonal preconditioning and noise reshaping. Their formulation allows time-dependent learning rates, and they also discuss a minibatch extension. Their main high-dimensional approximation theorem assumes covariance eigenvalues bounded above and away from zero uniformly in dimension. We instead study a diagonal power-law spectrum with eigenvalues tending to zero, focusing on the label-noise-dominated SDE model. In this setting, we obtain an explicit FSL with power-law signal-learning and noise-forgetting profiles under general LR schedules. This yields optimal data-scaling rates over all admissible schedules and comparisons of critical batch sizes.

Comparison with Kim et al. [22]. The most closely related work is Kim et al. [22], which also provides a scaling-law analysis of SignSGD in PLK regression. Our work differs in three main aspects. First, we focus on the label-noise-dominated regime and establish an explicit FSL for general LR schedules in our diagonal SDE model, whereas their general trajectory formulation is implicit and depends on the unknown risk. Second, we characterize the batch-size dependence through the intrinsic clock $t = \int_0^z \frac{\eta(u)}{\sigma_B}\,\mathrm{d}u$, whereas their theory is restricted to $B = 1$. Third, we derive the optimal data-scaling rate of SignSGD over all admissible LR schedules and obtain a complete phase diagram comparing the rates of SignSGD and SGD under optimally tuned LR schedules. In contrast, their comparison is restricted to specific schedules.

B Background on SDEs

This appendix reviews two standard tools for analyzing Itô stochastic differential equations (SDEs): the generator for expectation estimates and a time-change formula. These tools are used in the SDE modeling of Appendix C and in the proof in Appendix D.

B.1 The generator for expectation estimates

We briefly recall how the generator of an Itô SDE can be used to derive expectation and moment estimates. For the details of this method, one can refer to Sec.7 of Øksendal [41]. We now state a more general version (time-inhomogeneous case) and derive the corresponding results here.

Consider the following SDE

$$\mathrm{d}\mathbf{Y}_t = \mathbf{b}(t, \mathbf{Y}_t)\,\mathrm{d}t + \mathbf{G}(t, \mathbf{Y}_t)\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{Y}_t \in \mathbb{R}^n, \tag{14}$$

where $\mathbf{W}_t$ is a d-dimensional Brownian motion, $\mathbf{b}(t, \mathbf{y}) \in \mathbb{R}^n$, and $\mathbf{G}(t, \mathbf{y}) \in \mathbb{R}^{n\times d}$.

We first recall the time-homogeneous case, in which $\mathbf{b}$ and $\mathbf{G}$ do not depend explicitly on t. Let $\mathbf{Y}_t^{\mathbf{y}}$ denote the solution initialized at $\mathbf{Y}_0^{\mathbf{y}} = \mathbf{y}$. The infinitesimal generator is defined by

$$(\mathcal{L}f)(\mathbf{y}) \coloneqq \lim_{h\downarrow 0} \frac{\mathbb{E}[f(\mathbf{Y}_h^{\mathbf{y}})] - f(\mathbf{y})}{h}, \tag{15}$$

whenever the limit exists. For $f \in C^2(\mathbb{R}^n)$, Itô's formula gives the explicit representation

$$(\mathcal{L}f)(\mathbf{y}) = \mathbf{b}(\mathbf{y})^\top \nabla f(\mathbf{y}) + \frac{1}{2}\,\mathrm{tr}\big(\mathbf{G}(\mathbf{y})\mathbf{G}(\mathbf{y})^\top \nabla^2 f(\mathbf{y})\big)\,. \tag{16}$$

For the time-inhomogeneous SDE (14), we define the instantaneous generator at time t by

$$(\mathcal{L}_t f)(\mathbf{y}) \coloneqq \mathbf{b}(t,\mathbf{y})^\top \nabla f(\mathbf{y}) + \frac{1}{2}\operatorname{tr}\big(\mathbf{G}(t,\mathbf{y})\mathbf{G}(t,\mathbf{y})^\top \nabla^2 f(\mathbf{y})\big)\,. \tag{17}$$

Under the standard Markov regularity conditions, this operator agrees with the infinitesimal conditional expectation

$$(\mathcal{L}_t f)(\mathbf{y}) = \lim_{h\downarrow 0} \frac{\mathbb{E}[f(\mathbf{Y}_{t+h}) \mid \mathbf{Y}_t = \mathbf{y}] - f(\mathbf{y})}{h}, \tag{18}$$

whenever the limit exists. Thus, $\mathcal{L}_t f$ describes the instantaneous rate of change of $f(\mathbf{Y}_t)$ after the stochastic fluctuation has been averaged out.

Applying Itô's formula to (14) gives

$$f(\mathbf{Y}_t) = f(\mathbf{Y}_0) + \int_0^t (\mathcal{L}_s f)(\mathbf{Y}_s)\,\mathrm{d}s + \int_0^t \nabla f(\mathbf{Y}_s)^\top \mathbf{G}(s,\mathbf{Y}_s)\,\mathrm{d}\mathbf{W}_s. \tag{19}$$

Under suitable integrability conditions, the stochastic integral has zero expectation. Therefore, the generator $\mathcal{L}_t f$ determines the evolution of expectation $\mathbb{E}[f(\mathbf{Y}_t)]$.

The following lemma states Dynkin's formula and the upper and lower expectation bounds obtained from generator inequalities. Its proof combines Itô's formula with an integrating factor and also applies to time-inhomogeneous coefficients.

Lemma B.1 (Dynkin's formula and expectation bounds)**.** *Let $\mathbf{Y}_t$ satisfy (14), and let $V \in C^2(\mathbb{R}^n;[0,\infty))$. Assume that $\sup_{t\leqslant T} V(\mathbf{Y}_t)$ and $\int_0^T |(\mathcal{L}_t V)(\mathbf{Y}_t)|\,\mathrm{d}t$ are integrable. Then*

$$\mathbb{E}[V(\mathbf{Y}_t)] = \mathbb{E}[V(\mathbf{Y}_0)] + \int_0^t \mathbb{E}[(\mathcal{L}_s V)(\mathbf{Y}_s)]\,\mathrm{d}s. \tag{20}$$

Suppose, in addition, that there exist $c_V > 0$ and a nonnegative function $r : [0,T]\times\mathbb{R}^n \to [0,\infty)$ such that

$$(\mathcal{L}_t V)(\mathbf{y}) \leqslant -c_V V(\mathbf{y}) + r(t,\mathbf{y}) \tag{21}$$

for every $t \in [0,T]$ and $\mathbf{y}\in\mathbb{R}^n$. Then

$$\mathbb{E}[V(\mathbf{Y}_t)] \leqslant e^{-c_V t}\mathbb{E}[V(\mathbf{Y}_0)] + \int_0^t e^{-c_V(t-s)}\mathbb{E}[r(s,\mathbf{Y}_s)]\,\mathrm{d}s. \tag{22}$$

If instead the generator satisfies

$$(\mathcal{L}_t V)(\mathbf{y}) \geqslant -c_V V(\mathbf{y}) + r(t,\mathbf{y}) \tag{23}$$

for every $t \in [0,T]$ and $\mathbf{y}\in\mathbb{R}^n$, then

$$\mathbb{E}[V(\mathbf{Y}_t)] \geqslant e^{-c_V t}\mathbb{E}[V(\mathbf{Y}_0)] + \int_0^t e^{-c_V(t-s)}\mathbb{E}[r(s,\mathbf{Y}_s)]\,\mathrm{d}s. \tag{24}$$

Proof. For $R>0$, define the stopping time

$$\rho_R \coloneqq \inf\left\{ t \geqslant 0 : \int_0^t \left\|\nabla V(\mathbf{Y}_s)^\top \mathbf{G}(s,\mathbf{Y}_s)\right\|^2 \mathrm{d}s \geqslant R \right\}.$$

Applying Itô's formula to $V(\mathbf{Y}_{t\wedge\rho_R})$ gives

$$\begin{aligned} V(\mathbf{Y}_{t\wedge\rho_R}) = V(\mathbf{Y}_0) &+ \int_0^{t\wedge\rho_R} (\mathcal{L}_s V)(\mathbf{Y}_s)\,\mathrm{d}s \\ &+ \int_0^{t\wedge\rho_R} \nabla V(\mathbf{Y}_s)^\top \mathbf{G}(s,\mathbf{Y}_s)\,\mathrm{d}\mathbf{W}_s. \end{aligned}$$

The stopped stochastic integral is a martingale and therefore has zero expectation. Hence,

$$\mathbb{E}[V(\mathbf{Y}_{t\wedge\rho_R})] = \mathbb{E}[V(\mathbf{Y}_0)] + \mathbb{E}\left[\int_0^{t\wedge\rho_R} (\mathcal{L}_s V)(\mathbf{Y}_s)\,\mathrm{d}s\right].$$

Letting $R \to \infty$ under the stated integrability assumptions and applying Fubini's theorem gives (20). If we further assume (21), differentiating (20) gives, for almost every t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[V(\mathbf{Y}_t)] \leqslant -c_V\mathbb{E}[V(\mathbf{Y}_t)] + \mathbb{E}[r(t,\mathbf{Y}_t)].$$

Multiplying by $e^{c_V t}$ and integrating from 0 to t proves (22). Under (23), the same argument with both inequalities reversed proves (24). □

To obtain a moment bound, we choose a test function V, establish a pointwise inequality for $\mathcal{L}_t V$, and apply the corresponding expectation bound in Lemma B.1. This strategy is used in Appendix D. Specifically, Corollary D.3 combines the quadratic drift estimate in Lemma D.2 with the lower bound (24). Corollary D.6 combines the Lyapunov drift estimate in Lemma D.5 with the upper bound (22).

B.2 Time-change formula for SDEs

We next recall how an SDE transforms under a deterministic time change. This follows from the standard time-change formula for Itô integrals; see Sec.8.5 of Øksendal [41]. To make the presentation self-contained, we state and prove the version needed here.

Lemma B.2 (Deterministic time change for SDEs)**.** *Let $\Gamma : [0, Z] \to [0, T]$ be a continuously differentiable bijection satisfying $\Gamma(0) = 0$ and $\Gamma'(z) > 0$ for all $z \in [0, Z]$. Suppose that $(\widetilde{\mathbf{Y}}_z)_{z\in[0,Z]}$ satisfies*

$$\mathrm{d}\widetilde{\mathbf{Y}}_z = \mathbf{a}(z, \widetilde{\mathbf{Y}}_z)\ \mathrm{d}z + \mathbf{C}(z, \widetilde{\mathbf{Y}}_z)\ \mathrm{d}\mathbf{W}_z, \tag{25}$$

where $\mathbf{W}_z$ is a d-dimensional Brownian motion. Define

$$\overline{\mathbf{Y}}_t \coloneqq \widetilde{\mathbf{Y}}_{\Gamma^{-1}(t)}, \qquad t \in [0, T].$$

Then there exists a d-dimensional Brownian motion $\overline{\mathbf{W}}_t$ such that

$$\mathrm{d}\overline{\mathbf{Y}}_t = (\Gamma^{-1})'(t)\mathbf{a}(\Gamma^{-1}(t), \overline{\mathbf{Y}}_t)\ \mathrm{d}t + \sqrt{(\Gamma^{-1})'(t)}\mathbf{C}(\Gamma^{-1}(t), \overline{\mathbf{Y}}_t)\ \mathrm{d}\overline{\mathbf{W}}_t. \tag{26}$$

Proof. Write (25) in integral form. For $t \in [0, T]$,

$$\overline{\mathbf{Y}}_t = \widetilde{\mathbf{Y}}_0 + \int_0^{\Gamma^{-1}(t)} \mathbf{a}(u, \widetilde{\mathbf{Y}}_u)\ \mathrm{d}u + \int_0^{\Gamma^{-1}(t)} \mathbf{C}(u, \widetilde{\mathbf{Y}}_u)\ \mathrm{d}\mathbf{W}_u. \tag{27}$$

For the drift term, using the change of variables $u = \Gamma^{-1}(s)$, we have

$$\int_0^{\Gamma^{-1}(t)} \mathbf{a}(u, \widetilde{\mathbf{Y}}_u)\ \mathrm{d}u = \int_0^t (\Gamma^{-1})'(s)\mathbf{a}(\Gamma^{-1}(s), \overline{\mathbf{Y}}_s)\ \mathrm{d}s. \tag{28}$$

To make the change of the stochastic integral explicit, define

$$\overline{\mathbf{W}}_t \coloneqq \int_0^{\Gamma^{-1}(t)} \sqrt{\Gamma'(u)}\ \mathrm{d}\mathbf{W}_u.$$

We claim that $\overline{\mathbf{W}}$ is a standard Brownian motion in the new time variable t. Indeed, for $0 \leqslant s < t$,

$$\overline{\mathbf{W}}_t - \overline{\mathbf{W}}_s = \int_{\Gamma^{-1}(s)}^{\Gamma^{-1}(t)} \sqrt{\Gamma'(u)}\ \mathrm{d}\mathbf{W}_u.$$

This increment is Gaussian with mean zero and covariance

$$\int_{\Gamma^{-1}(s)}^{\Gamma^{-1}(t)} \Gamma'(u)\ \mathrm{d}u\, \mathbf{I}_d = \big(\Gamma(\Gamma^{-1}(t)) - \Gamma(\Gamma^{-1}(s))\big)\, \mathbf{I}_d = (t-s)\mathbf{I}_d.$$

Moreover, increments over disjoint intervals in t are independent because they are stochastic integrals over disjoint intervals in the original Brownian time u. Thus,

$$\overline{\mathbf{W}}_t - \overline{\mathbf{W}}_s \sim \mathcal{N}(0, (t-s)\mathbf{I}_d)$$

and has independent increments relative to the time-changed filtration. Therefore, $\overline{\mathbf{W}}$ is a standard d-dimensional Brownian motion in the new time variable.

Using $(\Gamma^{-1})'(t) = \frac{1}{\Gamma'(\Gamma^{-1}(t))}$, the deterministic time-change formula for Itô integrals gives

$$\int_0^{\Gamma^{-1}(t)} \mathbf{C}(u, \widetilde{\mathbf{Y}}_u)\, \mathrm{d}\mathbf{W}_u = \int_0^t \sqrt{(\Gamma^{-1})'(s)}\mathbf{C}(\Gamma^{-1}(s), \overline{\mathbf{Y}}_s)\, \mathrm{d}\overline{\mathbf{W}}_s. \tag{29}$$

Combining (27), (28), and (29), we obtain

$$\overline{\mathbf{Y}}_t = \overline{\mathbf{Y}}_0 + \int_0^t (\Gamma^{-1})'(s)\mathbf{a}(\Gamma^{-1}(s), \overline{\mathbf{Y}}_s)\, \mathrm{d}s + \int_0^t \sqrt{(\Gamma^{-1})'(s)}\mathbf{C}(\Gamma^{-1}(s), \overline{\mathbf{Y}}_s)\, \mathrm{d}\overline{\mathbf{W}}_s.$$

This is precisely the integral form of (26). □

C The SDE modeling for SGD and SignSGD

Throughout this appendix, z denotes physical time measured in training steps, and $\eta(z)$ is a continuous learning rate schedule satisfying $\eta(k) = \eta_k$.

C.1 The structure of gradient noise

We first characterize the covariance structure of the gradient noise. The Hessian-aligned structure of SGD noise has been observed and analyzed in prior works on SGD noise geometry [55, 57]. For completeness, we provide a self-contained derivation in our kernel regression model.

We use the batch noise scale $\sigma_B = 1/\sqrt{B}$ from Section 2. Let $\mathcal{B}$ be a fresh mini-batch of B i.i.d. samples and define its gradient by

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{1}{B} \sum_{(\mathbf{x},y)\in\mathcal{B}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, y).$$

Define the gradient noise by $\boldsymbol{\xi}(\boldsymbol{\theta}) := \mathbf{g}(\boldsymbol{\theta}) - \nabla\mathcal{R}(\boldsymbol{\theta})$. Then $\mathbb{E}_{\mathcal{B}}[\boldsymbol{\xi}(\boldsymbol{\theta})] = \mathbf{0}$. At step k, evaluating $\mathbf{g}(\boldsymbol{\theta})$ and $\boldsymbol{\xi}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_k$ using the fresh batch $\mathcal{B}_k$ gives $\mathbf{g}_k$ and $\boldsymbol{\xi}_k$, respectively. The following theorem shows that the covariance of this noise is aligned with the Hessian.

Theorem C.1 (Gradient-noise covariance)**.** *Consider the kernel regression model in Section 2. The mini-batch and single-sample gradient-noise covariances satisfy*

$$\mathbb{E}_{\mathcal{B}}[\boldsymbol{\xi}(\boldsymbol{\theta})\boldsymbol{\xi}(\boldsymbol{\theta})^\top] = \sigma_B^2 \boldsymbol{\Sigma}(\boldsymbol{\theta}), \qquad \boldsymbol{\Sigma}(\boldsymbol{\theta}) \eqsim \big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2\big)\mathbf{H}.$$

Here and below, for two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \eqsim \mathbf{B}$ if their quadratic forms are equivalent up to universal constants, i.e., there exist constants $c_1, c_2 > 0$, independent of $\mathbf{v}$, such that $c_1\mathbf{v}^\top\mathbf{B}\mathbf{v} \leqslant \mathbf{v}^\top\mathbf{A}\mathbf{v} \leqslant c_2\mathbf{v}^\top\mathbf{B}\mathbf{v}$.

Therefore, for any unit direction $\mathbf{v}$,

$$\mathbb{E}_{\mathcal{B}}[|\boldsymbol{\xi}(\boldsymbol{\theta})^\top\mathbf{v}|^2] = \sigma_B^2\mathbf{v}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} \eqsim \sigma_B^2\big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2\big)\mathbf{v}^\top\mathbf{H}\mathbf{v}.$$

Since $\mathbf{v}^\top\mathbf{H}\mathbf{v}$ is the curvature along $\mathbf{v}$, the gradient noise variance in any direction is proportional to the curvature in that direction (up to constants). In the label-noise-dominated regime $\mathcal{E}(\boldsymbol{\theta}) \lesssim \sigma^2$, this simplifies to

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) \eqsim \mathbf{H}, \qquad \mathbb{E}_{\mathcal{B}}[|\boldsymbol{\xi}(\boldsymbol{\theta})^\top\mathbf{v}|^2] \eqsim \sigma_B^2\mathbf{v}^\top\mathbf{H}\mathbf{v}.$$

This is due to our assumption $\sigma^2 \eqsim 1$. Thus, in coordinate j, the mini-batch noise has standard deviation of order $\sigma_B\sqrt{\lambda_j}$. This coordinate-wise noise scale motivates the normalized noise model used below for the SignSGD scaling analysis.

Proof of Theorem C.1. For a mini-batch of size B, write $\boldsymbol{\xi}(\boldsymbol{\theta}) = \frac{1}{B}\sum_{i=1}^B \boldsymbol{\zeta}_i(\boldsymbol{\theta})$, where each $\boldsymbol{\zeta}_i(\boldsymbol{\theta})$ is the gradient noise over a single data point. Then

$$\mathbb{E}[\boldsymbol{\xi}(\boldsymbol{\theta})\boldsymbol{\xi}(\boldsymbol{\theta})^\top] = \frac{1}{B^2}\sum_{i,j=1}^B \mathbb{E}[\boldsymbol{\zeta}_i(\boldsymbol{\theta})\boldsymbol{\zeta}_j(\boldsymbol{\theta})^\top] = \frac{1}{B^2}\sum_{i=1}^B \mathbb{E}[\boldsymbol{\zeta}_i(\boldsymbol{\theta})\boldsymbol{\zeta}_i(\boldsymbol{\theta})^\top] = \frac{1}{B}\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma_B^2\boldsymbol{\Sigma}(\boldsymbol{\theta}).$$

It remains to compute $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Consider a single data point $(\mathbf{x}, y)$ and let $\mathbf{g}_{\mathbf{x},y}(\boldsymbol{\theta})$ be its gradient. Let $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*, \boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$. Since $y = \langle \boldsymbol{\phi}, \boldsymbol{\theta}^* \rangle + \epsilon$, we have

$$\mathbf{g}_{\mathbf{x},y}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left(\frac{1}{2} (f(\mathbf{x}; \boldsymbol{\theta}) - y)^2 \right) = \big(\langle \boldsymbol{\phi}, \boldsymbol{\theta} \rangle - y \big) \boldsymbol{\phi} = \big(\langle \boldsymbol{\phi}, \boldsymbol{\delta} \rangle - \epsilon \big) \boldsymbol{\phi}.$$

The noise ϵ and feature $\boldsymbol{\phi}$ are independent and satisfy $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top] = \mathbf{H}$. Taking expectations over $(\mathbf{x}, y)$, we obtain

$$\nabla \mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y}[\mathbf{g}_{\mathbf{x},y}(\boldsymbol{\theta})] = \mathbf{H}\boldsymbol{\delta}.$$

Thus, the single-sample gradient noise is

$$\boldsymbol{\zeta}(\boldsymbol{\theta}) = \mathbf{g}_{\mathbf{x},y}(\boldsymbol{\theta}) - \nabla \mathcal{R}(\boldsymbol{\theta}) = (\boldsymbol{\phi}\boldsymbol{\phi}^\top - \mathbf{H})\boldsymbol{\delta} - \epsilon\boldsymbol{\phi}.$$

We next compute its covariance:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y}[\boldsymbol{\zeta}(\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\theta})^\top] = \mathbb{E}\left[(\boldsymbol{\phi}\boldsymbol{\phi}^\top - \mathbf{H})\boldsymbol{\delta}\boldsymbol{\delta}^\top (\boldsymbol{\phi}\boldsymbol{\phi}^\top - \mathbf{H}) \right] + \sigma^2 \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top],$$

where the cross terms vanish because ϵ is independent of $\boldsymbol{\phi}$ and $\mathbb{E}[\epsilon] = 0$. The second term equals $\sigma^2\mathbf{H}$. For the first term, we have

$$\mathbb{E}\left[(\boldsymbol{\phi}\boldsymbol{\phi}^\top - \mathbf{H})\boldsymbol{\delta}\boldsymbol{\delta}^\top (\boldsymbol{\phi}\boldsymbol{\phi}^\top - \mathbf{H}) \right] = \mathbb{E}\left[(\boldsymbol{\delta}^\top \boldsymbol{\phi})^2 \boldsymbol{\phi}\boldsymbol{\phi}^\top \right] - \mathbf{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\mathbf{H} = (\boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta})\mathbf{H} + \mathbf{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\mathbf{H},$$

where we use the Gaussian fourth-moment identity $\mathbb{E}\left[(\boldsymbol{\delta}^\top \boldsymbol{\phi})^2 \boldsymbol{\phi}\boldsymbol{\phi}^\top \right] = (\boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta})\mathbf{H} + 2\mathbf{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\mathbf{H}$. Thus, we obtain the noise covariance of single-sample gradient

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \big(\sigma^2 + \boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} \big)\mathbf{H} + \mathbf{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\mathbf{H}.$$

We claim that, uniformly over all $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{v}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} \eqsim \big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2 \big) \mathbf{v}^\top \mathbf{H}\mathbf{v}.$$

Indeed, using the covariance formula above,

$$\mathbf{v}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} = \big(\sigma^2 + \boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} \big) \mathbf{v}^\top\mathbf{H}\mathbf{v} + (\boldsymbol{\delta}^\top\mathbf{H}\mathbf{v})^2.$$

The lower bound follows immediately by dropping the nonnegative second term: $\mathbf{v}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} \geqslant \big(\sigma^2 + \boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} \big) \mathbf{v}^\top\mathbf{H}\mathbf{v}$. For the upper bound, the Cauchy–Schwarz inequality gives $(\boldsymbol{\delta}^\top\mathbf{H}\mathbf{v})^2 \leqslant (\boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta})(\mathbf{v}^\top\mathbf{H}\mathbf{v})$. Therefore,

$$\mathbf{v}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} \leqslant \big(\sigma^2 + 2\boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} \big) \mathbf{v}^\top\mathbf{H}\mathbf{v}.$$

Using $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta}$ and combining this with the lower bound, we obtain

$$\big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2 \big) \mathbf{v}^\top\mathbf{H}\mathbf{v} \leqslant \mathbf{v}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{v} \leqslant 4\big(\mathcal{E}(\boldsymbol{\theta}) + \sigma^2 \big) \mathbf{v}^\top\mathbf{H}\mathbf{v}.$$

This proves

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) \eqsim (\mathcal{E}(\boldsymbol{\theta}) + \sigma^2)\mathbf{H}.$$

□

Remark C.2 (Beyond Gaussian features). Gaussianity is not needed for curvature alignment in the label-noise-dominated regime. Suppose the hypercontractivity condition $\mathbb{E}[(\mathbf{v}^\top\boldsymbol{\phi})^4] \leqslant C(\mathbf{v}^\top\mathbf{H}\mathbf{v})^2$ holds for every $\mathbf{v} \in \mathbb{R}^d$, with a constant C independent of $\mathbf{v}$ and d. A similar argument gives

$$\sigma_B^2\sigma^2\mathbf{H} \lesssim \sigma_B^2 \boldsymbol{\Sigma}(\boldsymbol{\theta}) \lesssim \sigma_B^2 \big(\sigma^2 + \mathcal{E}(\boldsymbol{\theta}) \big) \mathbf{H}.$$

This leads to the same result $\sigma_B^2 \boldsymbol{\Sigma}(\boldsymbol{\theta}) \eqsim \sigma_B^2\mathbf{H}$ in the label-noise-dominated regime $\mathcal{E}(\boldsymbol{\theta}) \lesssim \sigma^2 \eqsim 1$.

C.2 The SDE modeling for SGD

We briefly review the SDE modeling for SGD in Li et al. [27]. The SGD update is

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{g}_k. \tag{30}$$

We decompose the mini-batch gradient as $\mathbf{g}_k = \nabla \mathcal{R}(\boldsymbol{\theta}_k) + \boldsymbol{\xi}_k$. By Theorem C.1, its conditional moments satisfy

$$\mathbb{E}[\boldsymbol{\xi}_k \mid \boldsymbol{\theta}_k] = \mathbf{0}, \qquad \mathbb{E}[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top \mid \boldsymbol{\theta}_k] = \sigma_B^2 \boldsymbol{\Sigma}(\boldsymbol{\theta}_k).$$

Here $\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)$ denotes the covariance of a single-sample stochastic gradient.

Therefore, one SGD step can be written as $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta_k \nabla \mathcal{R}(\boldsymbol{\theta}_k) - \eta_k \boldsymbol{\xi}_k$, consisting of a drift term $-\eta_k \nabla \mathcal{R}(\boldsymbol{\theta}_k)$ and a fluctuation term with covariance $\eta_k^2 \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)/B$. This motivates the physical-time SDE:

$$\mathrm{d}\widetilde{\boldsymbol{\theta}}_z = -\eta(z) \nabla \mathcal{R}(\widetilde{\boldsymbol{\theta}}_z)\, \mathrm{d}z + \frac{\eta(z)}{\sqrt{B}} \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}_z)^{1/2}\, \mathrm{d}\mathbf{W}_z, \tag{31}$$

where $\eta(z)$ is the continuous-time LRS satisfying $\eta(k) = \eta_k$ for $k \in \mathbb{N}$, and z denotes the physical time, which is a continuous analogue of the training step. To account for the influence of LRS, we define the intrinsic time $t = \Gamma_{\mathrm{sgd}}(z) \coloneqq \int_0^z \eta(u)\, \mathrm{d}u$, which can be viewed as the continuous-time analogue of $\sum_{j=0}^{k-1} \eta_j$. Let $\boldsymbol{\vartheta}_t = \widetilde{\boldsymbol{\theta}}_{\Gamma_{\mathrm{sgd}}^{-1}(t)}$. Since $\frac{\mathrm{d}}{\mathrm{d}t}\Gamma_{\mathrm{sgd}}^{-1}(t) = \frac{1}{\eta(\Gamma_{\mathrm{sgd}}^{-1}(t))}$, applying the time-change formula in Lemma B.2 with $\mathbf{a}(z, \boldsymbol{\theta}) = -\eta(z)\nabla \mathcal{R}(\boldsymbol{\theta}), \mathbf{C}(z, \boldsymbol{\theta}) = \frac{\eta(z)}{\sqrt{B}} \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}$, yields the intrinsic-time SDE

$$\mathrm{d}\boldsymbol{\vartheta}_t = -\nabla \mathcal{R}(\boldsymbol{\vartheta}_t)\, \mathrm{d}t + \sqrt{\frac{\varphi_{\mathrm{sgd}}(t)}{B}}\, \boldsymbol{\Sigma}(\boldsymbol{\vartheta}_t)^{1/2}\, \mathrm{d}\mathbf{W}_t, \tag{32}$$

where $\varphi_{\mathrm{sgd}}(t) \coloneqq \eta(\Gamma_{\mathrm{sgd}}^{-1}(t))$ is the learning rate schedule in intrinsic time. By the time change, we decouple the effect of the learning rate from the drift term.

Here $t = t_{\mathrm{sgd}}(z) = \Gamma_{\mathrm{sgd}}(z)$, and the noise injection intensity is $n_{\mathrm{sgd}}(t) = \varphi_{\mathrm{sgd}}(t)/B$.

C.3 The SDE modeling for SignSGD

We now derive the SDE modeling for SignSGD. The SignSGD update is

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \operatorname{sgn}(\mathbf{g}_k). \tag{33}$$

As in Section 2, write $\mathbf{g}_k = \boldsymbol{\mu}_k + \boldsymbol{\xi}_k$, where $\boldsymbol{\mu}_k = \nabla \mathcal{R}(\boldsymbol{\theta}_k)$, and use the fixed-covariance Gaussian model $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \sigma_B^2 \boldsymbol{\Sigma})$. Recall the scaled smoothed sign operator for batch size B:

$$\mathbf{m}_B(\mathbf{x}) \coloneqq \sigma_B \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \big[\operatorname{sgn}(\mathbf{x} + \sigma_B \boldsymbol{\xi})\big].$$

Then

$$\mathbb{E}[\operatorname{sgn}(\mathbf{g}_k) \mid \boldsymbol{\theta}_k] = \mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B.$$

Define the centered sign noise and its covariance by

$$\mathbf{n}_k \coloneqq \operatorname{sgn}(\mathbf{g}_k) - \mathbf{m}_B(\boldsymbol{\mu}_k)/\sigma_B, \qquad \mathbf{Q}_B(\boldsymbol{\theta}_k) \coloneqq \mathbb{E}[\mathbf{n}_k \mathbf{n}_k^\top \mid \boldsymbol{\theta}_k].$$

Thus the update reads

$$\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\frac{\eta_k}{\sigma_B}\, \mathbf{m}_B(\nabla \mathcal{R}(\boldsymbol{\theta}_k)) - \eta_k \mathbf{n}_k.$$

Matching its conditional mean and covariance gives the physical-time model

$$\mathrm{d}\widetilde{\boldsymbol{\theta}}_z = -\frac{\eta(z)}{\sigma_B}\, \mathbf{m}_B(\nabla \mathcal{R}(\widetilde{\boldsymbol{\theta}}_z))\, \mathrm{d}z + \eta(z) \mathbf{Q}_B(\widetilde{\boldsymbol{\theta}}_z)^{1/2}\, \mathrm{d}\mathbf{W}_z. \tag{34}$$

Define the SignSGD intrinsic clock, its intrinsic-time learning rate, and its noise injection intensity by

$$t = \Gamma_B(z) \coloneqq \int_0^z \frac{\eta(u)}{\sigma_B}\, \mathrm{d}u,$$
$$\varphi(t) = \varphi_{\mathrm{sgn}}(t) \coloneqq \eta(\Gamma_B^{-1}(t)), \qquad n_B(t) \coloneqq \sigma_B \varphi(t).$$

Set $\boldsymbol{\vartheta}_t = \widetilde{\boldsymbol{\theta}}_{\Gamma_B^{-1}(t)}$. Applying Lemma B.2 with

$$\mathbf{a}(z, \boldsymbol{\theta}) = -\frac{\eta(z)}{\sigma_B}\, \mathbf{m}_B(\nabla \mathcal{R}(\boldsymbol{\theta})), \qquad \mathbf{C}(z, \boldsymbol{\theta}) = \eta(z) \mathbf{Q}_B(\boldsymbol{\theta})^{1/2},$$

and $(\Gamma_B^{-1})'(t) = \sigma_B/\varphi(t)$ gives

$$\mathrm{d}\boldsymbol{\vartheta}_t = -\mathbf{m}_B(\nabla \mathcal{R}(\boldsymbol{\vartheta}_t))\, \mathrm{d}t + \sqrt{n_B(t)}\, \mathbf{Q}_B(\boldsymbol{\vartheta}_t)^{1/2}\, \mathrm{d}\mathbf{W}_t. \tag{35}$$

This is the same SDE as (2). Its instantaneous diffusion covariance is $n_B(t)\mathbf{Q}_B(\boldsymbol{\vartheta}_t)$, and $\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)]$ denotes the expected excess risk. At zeros of the learning rate the dynamics are understood by continuity. On any interval where $\eta = 0$, both the parameter process and its clock are constant.

Under Assumption 2.5, the marginal standard deviations satisfy $a_j := \sqrt{\Sigma_{jj}} \eqsim \sqrt{\lambda_j}$. We use these marginals to derive the coordinate equations for Appendix D.

Lemma C.3 (Smoothed sign under Gaussian noise)**.** *Let $Z \sim \mathcal{N}(0,1)$, let $\tau > 0$, and define the smoothed sign function $m(u) := 2\Phi(u) - 1$. Then, for any $\mu \in \mathbb{R}$,*

$$\mathbb{E}[\operatorname{sgn}(\mu + \tau Z)] = m\left(\frac{\mu}{\tau}\right), \qquad \operatorname{Var}(\operatorname{sgn}(\mu + \tau Z)) = 1 - m\left(\frac{\mu}{\tau}\right)^2.$$

Proof. Since $\operatorname{sgn}(\mu + \tau Z) = 1$ on the event $Z > -\mu/\tau$ and equals -1 on its complement, we have

$$\mathbb{E}[\operatorname{sgn}(\mu + \tau Z)] = \mathbb{P}\left(Z > -\frac{\mu}{\tau}\right) - \mathbb{P}\left(Z \leqslant -\frac{\mu}{\tau}\right) = 2\Phi\left(\frac{\mu}{\tau}\right) - 1 = m\left(\frac{\mu}{\tau}\right).$$

Moreover, $\operatorname{sgn}(\mu + \tau Z)^2 = 1$ almost surely, so

$$\operatorname{Var}(\operatorname{sgn}(\mu + \tau Z)) = \mathbb{E}[\operatorname{sgn}(\mu + \tau Z)^2] - \mathbb{E}[\operatorname{sgn}(\mu + \tau Z)]^2 = 1 - m\left(\frac{\mu}{\tau}\right)^2.$$

□

Applying Lemma C.3 coordinate-wise to the drift and sign-noise covariance in (35) yields the following explicit dynamics.

Theorem C.4 (Coordinate-wise dynamics)**.** *Suppose $\mathbf{H} = \operatorname{diag}(\lambda_j)$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_B^2 \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \eqsim \mathbf{H}$. Set $k_j = \lambda_j/(\sigma_B a_j)$. For each coordinate j, we have*

$$[\mathbf{m}_B(\nabla\mathcal{R}(\boldsymbol{\theta}))]_j = \sigma_B m\big(k_j(\theta_j - \theta_j^*)\big), \qquad [\mathbf{Q}_B(\boldsymbol{\theta})]_{jj} = 1 - m\big(k_j(\theta_j - \theta_j^*)\big)^2.$$

Accordingly, the j-th coordinate of (35) satisfies

$$\mathrm{d}X_j(t) = -\sigma_B m(k_j X_j(t))\,\mathrm{d}t + \sqrt{n_B(t)\left[1 - m(k_j X_j(t))^2\right]}\,\mathrm{d}\widetilde{W}_{j,t}, \tag{36}$$

where $X_j(t) = (\boldsymbol{\vartheta}_t)_j - \theta_j^*$ and $X_j(0) = -\theta_j^*$ under the initialization $\boldsymbol{\vartheta}_0 = \mathbf{0}$.

Proof. The j-th gradient coordinate has mean $\lambda_j(\theta_j - \theta_j^*)$ and noise standard deviation $\sigma_B a_j$. Lemma C.3 therefore gives its mean sign $m(k_j(\theta_j - \theta_j^*))$ and variance $1 - m(k_j(\theta_j - \theta_j^*))^2$. Multiplying the mean by σ_B gives the scaled drift. The j-th coordinate martingale in (35) has instantaneous variance $n_B(t)[\mathbf{Q}_B(\boldsymbol{\vartheta}_t)]_{jj}$, so it admits the scalar Brownian representation in (36). The Brownian motions for different coordinates may be correlated. □

C.4 Properties of the smoothed sign function

We collect the properties of the smoothed sign function $m(z) = 2\Phi(z) - 1$ introduced in Appendix C.3. These estimates will be used in Appendix D to obtain lower and upper bounds for the modal second moments. Throughout the following analysis, we write

$$c := m'(0) = \sqrt{2/\pi}.$$

Lemma C.5 (Properties of the smoothed sign function)**.** *The smoothed sign function m has the following properties.*

1. *The function m is smooth, odd, and strictly increasing, with $|m(z)| < 1$ for any $z \in \mathbb{R}$.*
2. *The following bounds hold globally:*
$$0 \leqslant zm(z) \leqslant cz^2, \qquad m(z)^2 \leqslant c^2 z^2, \qquad z \in \mathbb{R}. \tag{37}$$
3. *Near the origin, we have*
$$\begin{aligned} m(z) &= cz + O(z^3), \quad z \to 0, \\ zm(z) &\geqslant \frac{c}{2} z^2, \qquad\quad |z| \leqslant 1. \end{aligned} \tag{38}$$

4. In the tails,

$$m(z)\,\mathrm{sgn}(z) \to 1, \qquad 1 - m(z)^2 \to 0, \qquad |z| \to \infty. \tag{39}$$

Proof. The definition of m gives

$$m(z) = c\int_0^z e^{-u^2/2}\,\mathrm{d}u, \qquad m'(z) = ce^{-z^2/2}.$$

Thus m is smooth, odd, and strictly increasing. Since $0 < \Phi(z) < 1$ for any finite z, we also have $|m(z)| < 1$. The integral representation gives $|m(z)| \leqslant c|z|$, and $m(z)$ has the same sign as z. These observations prove (37).

Expanding the integrand at zero gives $m(z) = cz + O(z^3)$. For $|z| \leqslant 1$, the same integral representation yields

$$zm(z) = c|z|\int_0^{|z|} e^{-u^2/2}\,\mathrm{d}u \geqslant ce^{-1/2}z^2 \geqslant \frac{c}{2}z^2,$$

which proves (38). Finally, $\Phi(z) \to 1$ as $z \to \infty$ and $\Phi(z) \to 0$ as $z \to -\infty$, giving (39). □

C.5 An extension to non-Gaussian gradient noise

This extension is supplementary and may be skipped when reading the main proof in Appendix D, which uses the Gaussian SDE from Appendix C.3. We show that the original mini-batch noise also yields the relevant smoothed-sign properties for integer $B \geqslant 2$.

Mini-batch gradient and density at zero. Fix $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and a coordinate j. For independent samples from the Gaussian regression model,

$$g_{B,j} = \frac{1}{B}\sum_{i=1}^{B}\phi_{ij}\left(\boldsymbol{\phi}_i^\top\boldsymbol{\delta} - \epsilon_i\right) = \lambda_j\delta_j + \xi_j.$$

Set

$$\sigma^2_{\mathrm{res},j} = \sigma^2 + \sum_{\ell\neq j}\lambda_\ell\delta_\ell^2, \qquad R_B = \frac{1}{B\lambda_j}\sum_{i=1}^{B}\phi_{ij}^2 \sim \frac{\chi^2_B}{B}.$$

The residual $\sum_{\ell\neq j}\phi_{i\ell}\delta_\ell - \epsilon_i$ is $\mathcal{N}(0, \sigma^2_{\mathrm{res},j})$ and independent of ϕ_{ij}. Conditioning on the coordinate features gives

$$g_{B,j} \overset{\mathrm{d}}{=} \lambda_j\delta_j R_B + \sigma_{\mathrm{res},j}\sqrt{\frac{\lambda_j}{B}}\sqrt{R_B}\,Z, \qquad Z \perp R_B.$$

Here $Z \sim \mathcal{N}(0,1)$. At zero coordinate error, $\delta_j = 0$, the noise is a symmetric Gaussian scale mixture with $\widehat{\sigma}_j = \sigma_{\mathrm{res},j}\sqrt{\lambda_j/B}$. Its density f_j satisfies

$$f_j(0) = \frac{\mathbb{E}[R_B^{-1/2}]}{\sqrt{2\pi}\,\widehat{\sigma}_j} = \frac{\sqrt{B}}{2\sqrt{\pi}\,\widehat{\sigma}_j}\frac{\Gamma((B-1)/2)}{\Gamma(B/2)} \eqsim \frac{1}{\widehat{\sigma}_j}.$$

For $B \geqslant 3$, $\mathbb{E}[R_B^{-1}] = B/(B-2) \leqslant 3$ gives $1 \leqslant \mathbb{E}[R_B^{-1/2}] \leqslant \sqrt{3}$. For $B = 2$, the mixture has the Laplace density

$$f_j(x) = \frac{1}{\sqrt{2}\,\widehat{\sigma}_j}\exp\left(-\frac{\sqrt{2}\,|x|}{\widehat{\sigma}_j}\right).$$

Thus the density is continuous, finite, and strictly positive at zero, with $f_j(0) \eqsim \widehat{\sigma}_j^{-1}$ uniformly over $B \geqslant 2$.

Smoothed sign and its role in the proof. To account for the dependence of the noise on $\boldsymbol{\delta}$, compute the mean sign directly. With $m(u) = 2\Phi(u) - 1$, conditioning gives

$$\mathbb{E}[\mathrm{sgn}(g_{B,j}) \mid \boldsymbol{\delta}] = \widetilde{m}_B\left(\frac{\sqrt{B\lambda_j}\,\delta_j}{\sigma_{\mathrm{res},j}}\right), \qquad \widetilde{m}_B(u) := \mathbb{E}[m(u\sqrt{R_B})]. \tag{40}$$

Integrating against the χ^2_B density yields

$$\widetilde{m}'_B(u) = \frac{2}{\sqrt{\pi B}}\frac{\Gamma((B+1)/2)}{\Gamma(B/2)}\left(1+\frac{u^2}{B}\right)^{-(B+1)/2}.$$

Hence $\widetilde{m}'_B(0) \eqsim 1$, with a uniform upper bound on the derivative and a positive lower bound on $|u| \leqslant 1$. For $B \geqslant 3$, $\widetilde{m}_B(u) = \mathbb{E}[\mathrm{sgn}(u + Z/\sqrt{R_B})]$ and $\mathbb{E}[Z^2/R_B] \leqslant 3$, so Chebyshev's inequality gives $1 - \widetilde{m}_B(u)^2 \leqslant 12/u^2$ for $|u| \geqslant 1$. For $B = 2$, direct integration gives

$$\widetilde{m}_2(u) = \frac{u}{\sqrt{u^2+2}}, \qquad 1 - \widetilde{m}_2(u)^2 = \frac{2}{u^2+2},$$

which yields the same tail bound. These estimates provide the local linear response, global linear upper bound, and restoring drift used in Appendix D, uniformly over $B \geqslant 2$.

If $\sigma^2 \eqsim 1$ and $\boldsymbol{\delta}^\top \mathbf{H}\boldsymbol{\delta} \lesssim 1$, then $\sigma_{\mathrm{res},j} \eqsim 1$ uniformly in j. Consequently, the response in (40) has the same coordinate-wise scale $\sqrt{B\lambda_j}$ as the Gaussian model of Appendix C.3, with comparable local response and restoring drift bounds uniformly over $B \geqslant 2$. These estimates support the Gaussian modeling choice. Extending the full FSL proof to the original mini-batch noise would additionally require controlling the state-dependent residual scales along the trajectory.

D FSL of SignSGD (Proof of Theorem 3.1)

Throughout this appendix, $t = \Gamma_B(z) = \sqrt{B}\int_0^z \eta(u)\,\mathrm{d}u$ denotes SignSGD intrinsic time. We work under the assumptions and the learning-rate and training-time conditions of Theorem 3.1, using the Gaussian model with $\boldsymbol{\Sigma} \eqsim \mathbf{H}$ specified in Appendix C.3.

We use $\varphi(t) = \varphi_{\mathrm{sgn}}(t) = \eta(\Gamma_B^{-1}(t))$ and $n_B(t) = \varphi(t)/\sqrt{B}$, consistently with Section 2. The scalar $m(u) = 2\Phi(u) - 1$ is the unscaled Gaussian sign response; the scaled vector drift in the main text is $\mathbf{m}_B$.

We prove the functional scaling law by obtaining upper and lower bounds for every modal second moment and then summing over the spectrum. The lower bound follows directly from the exact second-moment equation. For the upper bound, a Foster–Lyapunov estimate controls the dynamics, while the training-time requirement ensures burn-in control for modes with large normalized initial values. Throughout the proof, constants may depend on q, s, the fixed bound C_0 in $\sqrt{B}\,\eta_{\mathrm{peak}} \leqslant C_0$, and the comparison constants in Assumptions 2.3 and 2.5, but not on B, t or the particular learning rate schedule φ.

D.1 A technical lemma: power-law spectral sums

The modal estimates below reduce the signal learning and noise accumulation terms to sums of exponentials over the spectrum. The following lemma converts both sums into power laws and will be used in Appendix D.6 to complete the proof of Theorem 3.1.

Lemma D.1 (Power-law spectral sums)**.** *Let $r > 1$ and $\theta > 0$, and suppose $b_j \eqsim j^{-r}$ and $\rho_j \eqsim j^{-\theta}$. There is a constant $C \geqslant 1$ such that, for every $Q \geqslant 0$,*

$$C^{-1}(1+Q)^{-(r-1)/\theta} \leqslant \sum_{j=1}^{\infty} b_j e^{-Q\rho_j} \leqslant C(1+Q)^{-(r-1)/\theta}. \tag{41}$$

Proof. It suffices to consider $b_j = j^{-r}$ and $\rho_j = j^{-\theta}$. First consider $0 \leqslant Q \leqslant 1$. Then the $j = 1$ term gives a positive constant lower bound, while

$$\sum_{j=1}^{\infty} j^{-r} e^{-Qj^{-\theta}} \leqslant \sum_{j=1}^{\infty} j^{-r} < \infty.$$

Also $(1+Q)^{-(r-1)/\theta} \eqsim 1$ on this interval.

Now consider $Q > 1$. Let $N = Q^{1/\theta}$. For the upper bound, split the sum at N. The tail satisfies

$$\sum_{j>N} j^{-r} e^{-Qj^{-\theta}} \leqslant \sum_{j>N} j^{-r} \leqslant CN^{1-r}.$$

For the part $j \leqslant N$, divide the indices into dyadic blocks on which j is comparable to $2^{-k}N$. The contribution of the k-th block is bounded by

$$CN^{1-r} 2^{k(r-1)} e^{-2^{k\theta}}.$$

The sum over $k \geqslant 0$ is finite, so the contribution of $j \leqslant N$ is also bounded by $CN^{1-r} = CQ^{-(r-1)/\theta}$.

For the lower bound, consider the integer interval $\lceil N \rceil \leqslant j \leqslant \lfloor 3N \rfloor$, which contains at least N indices. On this interval, $e^{-Qj^{-\theta}} \geqslant e^{-1}$ and $j^{-r} \geqslant (3N)^{-r}$. Therefore,

$$\sum_{j=1}^{\infty} j^{-r} e^{-Qj^{-\theta}} \geqslant e^{-1} N (3N)^{-r} \gtrsim Q^{-(r-1)/\theta}.$$

Since $Q \geqslant 1$, this is equivalent to $(1+Q)^{-(r-1)/\theta}$. □

D.2 Time rescaling and normalized modal dynamics

Let $a_j = \sqrt{\Sigma_{jj}}$ be the single-sample noise standard deviation, as in Section 6. With $k_j = \sqrt{B}\,\lambda_j / a_j$, Theorem C.4 gives

$$\mathrm{d}X_j(t) = -\frac{1}{\sqrt{B}} m(k_j X_j(t))\,\mathrm{d}t + \sqrt{n_B(t)\left[1 - m(k_j X_j(t))^2\right]}\,\mathrm{d}\widetilde{W}_{j,t}, \tag{42}$$

where $X_j(t) = (\boldsymbol{\vartheta}_t)_j - \theta^*_j$, $X_j(0) = -\theta^*_j$, and $m(z) = 2\Phi(z) - 1$ is the smoothed sign function.

The initial signal energy in mode j is

$$w_j \coloneqq \lambda_j X_j(0)^2 = \lambda_j |\theta^*_j|^2.$$

By Assumption 2.3, we have $\lambda_j \eqsim j^{-q}$ and $w_j \eqsim j^{-1-qs}$. Tonelli's theorem gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] = \frac{1}{2} \sum_{j=1}^{\infty} \lambda_j \mathbb{E}[X_j(t)^2]. \tag{43}$$

The Brownian motions in different coordinates may be correlated. Since $\mathbf{H}$ is diagonal, (43) requires only the marginal second moments, and the scalar drift and diffusion coefficients in (42) depend only on $X_j(t)$.

At intrinsic time t, the elapsed time for mode j in rescaled units is

$$\widehat{t}_j = \frac{k_j t}{\sqrt{B}} = \frac{\lambda_j}{a_j}\, t.$$

Writing τ for the time variable in these units, define

$$Z_j(\tau) \coloneqq k_j\, X_j\left(\frac{\sqrt{B}\,\tau}{k_j}\right). \tag{44}$$

Then $Z_j(0)^2 = B(\lambda_j / a_j^2) w_j$. Applying Lemma B.2 with the time change $\tau = k_j t / \sqrt{B}$, and then using the rescaling (44), we obtain

$$\mathrm{d}Z_j(\tau) = -m(Z_j(\tau))\,\mathrm{d}\tau + \varepsilon_j(\tau)\sqrt{1 - m(Z_j(\tau))^2}\,\mathrm{d}\widetilde{W}_{j,\tau}, \tag{45}$$

where $\widetilde{W}_{j,\tau}$ is a standard Brownian motion in the rescaled time τ, and

$$\varepsilon_j(\tau)^2 := k_j\,\varphi\left(\frac{\sqrt{B}\,\tau}{k_j}\right) = \sqrt{B}\,k_j\,n_B\left(\frac{\sqrt{B}\,\tau}{k_j}\right). \tag{46}$$

Choose fixed constants $\lambda_+, w_+, h_-, h_+ > 0$ such that

$$\lambda_j \leqslant \lambda_+ j^{-q}, \qquad w_j \leqslant w_+ j^{-1-qs}, \qquad h_-\lambda_j \leqslant a_j^2 \leqslant h_+\lambda_j.$$

By the learning-rate condition in Theorem 3.1, fix $C_0 > 0$ such that $\sqrt{B}\,\|\varphi\|_\infty = \sqrt{B}\,\eta_{\text{peak}} \leqslant C_0$. This gives the uniform normalized noise bound

$$\varepsilon_j(\tau)^2 \leqslant k_j\|\varphi\|_\infty \leqslant C_0\sqrt{\lambda_+/h_-} =: \varepsilon_{\max}^2. \tag{47}$$

Here we used $k_j/\sqrt{B} = \lambda_j/a_j \leqslant \sqrt{\lambda_+/h_-}$.

Returning to the original risk weights in (43) gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j}{k_j^2}\mathbb{E}[Z_j(\widehat{t}_j)^2], \qquad \frac{\lambda_j}{k_j^2} = \frac{a_j^2}{B\lambda_j}. \tag{48}$$

D.3 The generator and a global lower bound

For $f \in C^2(\mathbb{R})$, we can write the instantaneous generator of the SDE (45) as

$$(\mathcal{L}_{j,\tau}f)(z) = -m(z)f'(z) + \frac{\varepsilon_j(\tau)^2}{2}\left(1 - m(z)^2\right)f''(z). \tag{49}$$

Note that both j and τ enter the instantaneous generator only through the scalar $\varepsilon_j(\tau)$. We therefore introduce the following operator, which provides a common representation of all the instantaneous generators:

$$(L_\varepsilon f)(z) := -m(z)f'(z) + \frac{\varepsilon^2}{2}\left(1 - m(z)^2\right)f''(z). \tag{50}$$

We have

$$\mathcal{L}_{j,\tau} = L_{\varepsilon_j(\tau)}.$$

After the time rescaling, the drift $-m(z)$ is identical for every mode, while all dependence on the mode j and on the learning-rate schedule is encoded in the single scalar $\varepsilon_j(\tau)$. It therefore suffices to establish drift estimates for a single generator L_ε, rather than treating each mode separately.

We use the properties of m collected in Appendix C.4, with $c = m'(0)$ as defined there. The local expansion (38) identifies c as the strength of the local linear drift, which enters both the lower and upper moment bounds. This local behavior motivates the quadratic test function $V(z) = z^2$. Since the drift saturates in the tails and the process can leave the neighborhood of the origin, the global upper bound uses a Lyapunov function that is quadratic near the origin and exponential outside a bounded interval.

Lemma D.2 (Quadratic drift lower bound)**.** *For $V(z) = z^2$, set*

$$\omega := 2c + c^2\varepsilon_{\max}^2. \tag{51}$$

Then, for any $z \in \mathbb{R}$ and $0 \leqslant \varepsilon^2 \leqslant \varepsilon_{\max}^2$,

$$L_\varepsilon V(z) \geqslant -\omega V(z) + \varepsilon^2. \tag{52}$$

Proof. For the test function $V(z) = z^2$, the global bounds (37) in Lemma C.5 give

$$\begin{aligned} L_\varepsilon V(z) &= -2zm(z) + \varepsilon^2\left(1 - m(z)^2\right) \\ &\geqslant -\left(2c + c^2\varepsilon^2\right)z^2 + \varepsilon^2 \\ &\geqslant -\omega V(z) + \varepsilon^2. \end{aligned}$$

□

Applying the lower-bound conclusion of Lemma B.1 to this drift estimate gives the following second-moment bound.

Corollary D.3 (Global second-moment lower bound)**.** *Let Z solve*

$$\mathrm{d}Z(\tau) = -m(Z(\tau))\ \mathrm{d}\tau + \varepsilon(\tau)\sqrt{1 - m(Z(\tau))^2}\ \mathrm{d}W_\tau, \qquad Z(0) = z_0,$$

where $0 \leqslant \varepsilon(\tau)^2 \leqslant \varepsilon_{\max}^2$. With ω as in (51), for any $\tau \geqslant 0$,

$$\mathbb{E}[Z(\tau)^2] \geqslant z_0^2 e^{-\omega\tau} + \int_0^\tau e^{-\omega(\tau-r)}\varepsilon(r)^2\ \mathrm{d}r. \tag{53}$$

Proof. Apply (24) in Lemma B.1 with $V(z) = z^2$, $c_V = \omega$, and $r(\tau, z) = \varepsilon(\tau)^2$, using Lemma D.2. □

D.4 A Foster–Lyapunov upper bound

We first construct a Lyapunov function, then use it to prove the second-moment upper bound.

Lemma D.4 (Auxiliary Lyapunov function)**.** *For any $R \geqslant 1$, there exists a nonnegative even function $\Psi_R \in C^2(\mathbb{R})$ with the following forms in the three regions:*

1. *For $|z| \leqslant \frac{1}{2}$, the function is quadratic:*
$$\Psi_R(z) = z^2. \tag{54}$$
2. *For $\frac{1}{2} < |z| < R$, the function smoothly interpolates between the quadratic and exponential pieces and satisfies*
$$z^2 \leqslant \Psi_R(z) \leqslant e^{|z|}.$$
3. *For $|z| \geqslant R$, the function is exponential:*
$$\Psi_R(z) = e^{|z|}.$$

In particular, the following bounds hold globally:

$$z^2 \leqslant \Psi_R(z) \leqslant 4z^2 e^{|z|}, \qquad z \in \mathbb{R}. \tag{55}$$

Proof. For $0 \leqslant y \leqslant 1$, let

$$h(y) = 10y^3 - 15y^4 + 6y^5.$$

Its first two derivatives are

$$h'(y) = 30y^2(1-y)^2, \qquad h''(y) = 60y(1-y)(1-2y).$$

In particular, we have $h(0) = 0, h(1) = 1$, and $h'(0) = h'(1) = h''(0) = h''(1) = 0$. Since $h'(y) \geqslant 0$, we also have $0 \leqslant h(y) \leqslant 1$ on $[0,1]$.

For $r \geqslant 0$, define

$$\psi_R(r) = \begin{cases} r^2, & 0 \leqslant r \leqslant \frac{1}{2}, \\ (1 - h(y))r^2 + h(y)e^r, \quad y = \dfrac{r - \frac{1}{2}}{R - \frac{1}{2}}, & \frac{1}{2} \leqslant r \leqslant R, \\ e^r, & r \geqslant R, \end{cases} \qquad \Psi_R(z) \coloneqq \psi_R(|z|).$$

The definition gives the required quadratic and exponential forms. At $r = \frac{1}{2}$ and $r = R$, the endpoint conditions on h, h', h'' make the values and the first two derivatives of the adjacent pieces agree. Thus ψ_R is C^2 across both boundaries. Since $\Psi_R(z) = z^2$ near zero, the even extension $\Psi_R(z) = \psi_R(|z|)$ belongs to $C^2(\mathbb{R})$. It is nonnegative because each piece is nonnegative.

On the connecting interval, $\psi_R(r)$ is a convex combination of r^2 and e^r. Since $e^r \geqslant r^2$ for $r \geqslant 0$, we have

$$r^2 \leqslant \psi_R(r) \leqslant e^r \leqslant e^R.$$

Finally, the construction gives $\Psi_R(z) \geqslant z^2$ for any $z \in \mathbb{R}$. For $|z| \geqslant \frac{1}{2}$, the connecting and tail regions satisfy

$$\Psi_R(z) \leqslant e^{|z|} \leqslant 4z^2 e^{|z|};$$

for $|z| \leqslant \frac{1}{2}$, $\Psi_R(z) = z^2 \leqslant 4z^2 e^{|z|}$. This proves (55). □

Lemma D.5 (Foster–Lyapunov drift estimate)**.** *Let $\varepsilon_{\max}^2$ be the fixed noise bound in (47). There exist a Lyapunov function $\mathcal{V}$ of the form constructed in Lemma D.4 and constants $c_L > 0$, $C \geqslant 1$ such that*

$$L_\varepsilon \mathcal{V}(z) \leqslant -c_L \mathcal{V}(z) + C\varepsilon^2$$

for any $z \in \mathbb{R}$ and $0 \leqslant \varepsilon^2 \leqslant \varepsilon_{\max}^2$.

Proof. By the tail limits (39) in Lemma C.5, we may choose $R \geqslant 1$ sufficiently large for the bound $\varepsilon_{\max}^2$ in (47) so that

$$m(z)\operatorname{sgn}(z) \geqslant \frac{3}{4}, \qquad \frac{\varepsilon_{\max}^2}{2}\left(1 - m(z)^2\right) \leqslant \frac{1}{4}, \qquad |z| \geqslant R.$$

Set $\mathcal{V} = \Psi_R$, using the function constructed in Lemma D.4. We estimate its generator in the local, tail, and connecting regions.

For $|z| \leqslant \frac{1}{2}$, we have $\mathcal{V}(z) = z^2$, so

$$L_\varepsilon z^2 = -2zm(z) + \varepsilon^2(1 - m(z)^2).$$

Using the local lower bound (38) in Lemma C.5 and $1 - m(z)^2 \leqslant 1$, we obtain

$$L_\varepsilon z^2 \leqslant -cz^2 + \varepsilon^2.$$

Thus the desired drift inequality holds in the local region.

On the tail region $|z| \geqslant R$, we have $\mathcal{V}(z) = e^{|z|}$. Hence

$$\mathcal{V}'(z) = \operatorname{sgn}(z)\mathcal{V}(z), \qquad \mathcal{V}''(z) = \mathcal{V}(z).$$

Therefore, for any $\varepsilon^2 \leqslant \varepsilon_{\max}^2$,

$$L_\varepsilon \mathcal{V}(z) = \mathcal{V}(z)\left[-m(z)\operatorname{sgn}(z) + \frac{\varepsilon^2}{2}(1 - m(z)^2)\right] \leqslant \mathcal{V}(z)\left[-\frac{3}{4} + \frac{1}{4}\right] = -\frac{1}{2}\mathcal{V}(z).$$

On the connecting region $\frac{1}{2} \leqslant |z| \leqslant R$, write $r = |z|$ and $y = (r - \frac{1}{2})/(R - \frac{1}{2})$. Differentiating the interpolation in Lemma D.4 gives

$$\mathcal{V}'(z)\operatorname{sgn}(z) = (1 - h(y))2r + h(y)e^r + \frac{h'(y)}{R - \frac{1}{2}}(e^r - r^2) \geqslant 1,$$

since $2r \geqslant 1$, $e^r \geqslant 1$, and $h'(y) \geqslant 0$. Smoothness on this compact region gives $|\mathcal{V}''(z)| \leqslant b_R$ for some finite $b_R > 0$. By the oddness and strict monotonicity in Lemma C.5, we have $m(z)\operatorname{sgn}(z) \geqslant m_* := m(\frac{1}{2}) > 0$ on this region, so

$$\begin{aligned} L_\varepsilon \mathcal{V}(z) &= -m(z)\mathcal{V}'(z) + \frac{\varepsilon^2}{2}(1 - m(z)^2)\mathcal{V}''(z) \\ &\leqslant -m_* + \frac{b_R}{2}\varepsilon^2. \end{aligned}$$

Choose

$$c_L := \min\left\{c, \frac{1}{2}, \frac{m_*}{2e^R}\right\} > 0, \qquad C := \max\left\{1, \frac{b_R}{2}\right\}.$$

Because $\mathcal{V}(z) \leqslant e^R$ on the connecting region, we have

$$-m_* \leqslant -c_L \mathcal{V}(z).$$

Together with the local and tail estimates, these choices give

$$L_\varepsilon \mathcal{V}(z) \leqslant -c_L \mathcal{V}(z) + C\varepsilon^2$$

for any $z \in \mathbb{R}$ and $0 \leqslant \varepsilon^2 \leqslant \varepsilon_{\max}^2$. □

Corollary D.6 (Second-moment upper bound)**.** *Let $\mathcal{V}$, c_L, and C be as in Lemma D.5. If Z solves*

$$\mathrm{d}Z(\tau) = -m(Z(\tau))\,\mathrm{d}\tau + \varepsilon(\tau)\sqrt{1 - m(Z(\tau))^2}\,\mathrm{d}W_\tau,$$

with deterministic initial value $Z(0)$ and $\sup_{\tau \geqslant 0} \varepsilon(\tau)^2 \leqslant \varepsilon_{\max}^2$, then

$$\mathbb{E}[Z(\tau)^2] \leqslant Ce^{-c_L\tau}\mathcal{V}(Z(0)) + C\int_0^\tau e^{-c_L(\tau - r)}\varepsilon(r)^2\,\mathrm{d}r.$$

Proof. The bounded drift and diffusion in the scalar SDE give exponential moments of every order, uniformly on each finite time interval. This ensures integrability in Dynkin's formula for $\mathcal{V}$ and its derivatives. Using the drift estimate in Lemma D.5, the upper-bound conclusion (22) of Lemma B.1 with $V = \mathcal{V}$, $c_V = c_L$, and $r(\tau, z) = C\varepsilon(\tau)^2$ gives

$$\mathbb{E}[\mathcal{V}(Z(\tau))] \leqslant e^{-c_L\tau}\mathcal{V}(Z(0)) + C\int_0^\tau e^{-c_L(\tau-r)}\varepsilon(r)^2\,\mathrm{d}r.$$

The bound $z^2 \leqslant \mathcal{V}(z)$ in (55) and $C \geqslant 1$ give the claimed second-moment upper bound. □

D.5 Control of the burn-in process

The following argument controls the burn-in phase for modes with large normalized initial values. Recall that $|Z_j(0)| = k_j|\theta_j^*|$. A large $|Z_j(0)|$ places the mode in the saturated region at initialization. Therefore, we require the intrinsic time t to be sufficiently large relative to $\sqrt{B}\,|\theta_j^*|$ to go through this phase.

We compare the signal learning terms in Corollary D.3 and Corollary D.6, which involve $Z_j(0)^2$ and $\mathcal{V}(Z_j(0))$, respectively. For modes with $|Z_j(0)| \leqslant \frac{1}{2}$, the identity $\mathcal{V}(Z_j(0)) = Z_j(0)^2$ gives a matching initialization in both bounds. For modes with large $|Z_j(0)|$, the factor $\mathcal{V}(Z_j(0))$ cannot be bounded by $Z_j(0)^2$ up to a uniform constant, so we need to control the effect of burn-in. Each modal signal learning term decays exponentially in time, allowing a sufficiently large t to suppress this extra factor. For these modes, the global envelope (55) gives

$$\mathcal{V}(Z_j(0)) \leqslant 4Z_j(0)^2 e^{|Z_j(0)|}.$$

The next lemma uses (57) to absorb $e^{|Z_j(0)|}$ into the exponential decay. To separate the modes according to their initial values, set $\zeta := 1 + qs$ and define

$$J_B := \max\left\{1, \left\lceil (4w_+B/h_-)^{1/\zeta} \right\rceil\right\}. \tag{56}$$

The cutoff satisfies $J_B \eqsim B^{1/\zeta} = B^{1/(1+qs)}$. Its purpose is to separate two cases in the signal learning bound. For $j \geqslant J_B$, the bounds on w_j and λ_j/a_j^2 give

$$Z_j(0)^2 = B(\lambda_j/a_j^2)w_j \leqslant \tfrac{1}{4}, \qquad \mathcal{V}(Z_j(0)) = Z_j(0)^2.$$

These modes already have a matching initialization in both bounds. For $j < J_B$, the initial value may grow with B. The conditions in (57) ensure that t exceeds the burn-in scale for these modes by a sufficient constant factor, allowing the extra factor to be absorbed into the exponential decay in t.

Lemma D.7 (Burn-in control for small indices)**.** *Suppose*

$$t \geqslant C_1\sqrt{B}, \qquad t \geqslant C_3 B^{1/(2s+2/q)}, \tag{57}$$

where $C_1, C_3 > 0$ are sufficiently large constants depending only on the fixed model parameters. Then, for any integer $1 \leqslant j < J_B$,

$$e^{-c_L k_j t/\sqrt{B}}\mathcal{V}(Z_j(0)) \leqslant 4Z_j(0)^2 e^{-c_L k_j t/(2\sqrt{B})}. \tag{58}$$

Proof. Fix an integer $1 \leqslant j < J_B$. Applying the global envelope (55) to the signal learning term gives

$$e^{-c_L k_j t/\sqrt{B}}\mathcal{V}(Z_j(0)) \leqslant 4Z_j(0)^2 e^{|Z_j(0)| - c_L k_j t/\sqrt{B}}.$$

Our goal is to bound this expression by $4Z_j(0)^2 e^{-c_L k_j t/(2\sqrt{B})}$. It therefore suffices to show

$$|Z_j(0)| \leqslant \frac{c_L k_j t}{2\sqrt{B}}.$$

Since $|Z_j(0)| = k_j|\theta_j^*|$ and $k_j > 0$, this condition reduces to

$$t \geqslant \frac{2\sqrt{B}\,|\theta_j^*|}{c_L}. \tag{59}$$

The target (59) compares the observation time with the burn-in scale $\sqrt{B}\,|\theta_j^*|$. We next estimate this scale for $j < J_B$. The power-law assumptions give

$$|\theta_j^*| = \sqrt{\frac{w_j}{\lambda_j}} \lesssim j^{(q-\zeta)/2}.$$

The sign of $q-\zeta$ determines whether this bound grows with j. If $\zeta \geqslant q$, it is bounded by a constant C_θ independent of B. Taking $C_1 \geqslant 2C_\theta/c_L$, the first condition in (57) gives

$$t \geqslant C_1\sqrt{B} \geqslant \frac{2C_\theta\sqrt{B}}{c_L} \geqslant \frac{2\sqrt{B}\,|\theta_j^*|}{c_L}.$$

If $\zeta < q$, the bound grows with j, so we use $j < J_B$ and $J_B \eqsim B^{1/\zeta}$ to obtain

$$|\theta_j^*| \leqslant C_\theta B^{(q-\zeta)/(2\zeta)}.$$

Thus we need t to dominate $\sqrt{B}$ times this polynomial in B. This is precisely what the second condition in (57) ensures: choosing $C_3 \geqslant 2C_\theta/c_L$ gives

$$t \geqslant C_3 B^{q/(2\zeta)} \geqslant \frac{2C_\theta\sqrt{B}}{c_L} B^{(q-\zeta)/(2\zeta)} \geqslant \frac{2\sqrt{B}\,|\theta_j^*|}{c_L}.$$

Thus, (59) holds in both cases, which proves the claim. □

Combining this control of the burn-in process with the normalized second-moment estimates gives the following bounds, with signal learning terms that depend quadratically on $Z_j(0)$ on both sides.

Proposition D.8 (Uniform two-sided modal bounds)**.** *Under the condition (57), there is a constant $C \geqslant 1$, independent of B, t and the learning rate schedule, such that, for any $j \geqslant 1$ and $\widehat{t}_j = k_j t/\sqrt{B}$,*

$$\mathbb{E}[Z_j(\widehat{t}_j)^2] \geqslant Z_j(0)^2 e^{-\omega\widehat{t}_j} + \int_0^{\widehat{t}_j} e^{-\omega(\widehat{t}_j - r)} \varepsilon_j(r)^2 \,\mathrm{d}r, \tag{60}$$

and

$$\mathbb{E}[Z_j(\widehat{t}_j)^2] \leqslant C\left[Z_j(0)^2 e^{-c_L\widehat{t}_j/2} + \int_0^{\widehat{t}_j} e^{-c_L(\widehat{t}_j - r)/2} \varepsilon_j(r)^2 \,\mathrm{d}r\right]. \tag{61}$$

Here ω and c_L are as in Lemma D.2 and Lemma D.5, respectively.

Proof. The lower bound follows directly from Corollary D.3. For the upper bound, apply Corollary D.6 to Z_j. When $j < J_B$, Lemma D.7 bounds the signal learning term by $Z_j(0)^2 e^{-c_L\widehat{t}_j/2}$ up to a constant factor. When $j \geqslant J_B$, the identity $\mathcal{V}(Z_j(0)) = Z_j(0)^2$ in (54) gives the same bound. Since $\varepsilon_j(r)^2 \geqslant 0$, replacing $e^{-c_L(\widehat{t}_j - r)}$ by $e^{-c_L(\widehat{t}_j - r)/2}$ bounds the noise integral from above. These estimates give (61). □

D.6 Completion of the proof

Recall from (48) that

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] = \frac{1}{2}\sum_{j=1}^{\infty} \frac{\lambda_j}{k_j^2}\mathbb{E}[Z_j(\widehat{t}_j)^2].$$

We next prove the upper bound for the expected excess risk. The lower bound follows by the same argument. The key step is to sum the exponentially decaying modal terms and apply Lemma D.1 to obtain the power laws. Choose C_1, C_3 as in Lemma D.7. The training-time condition in Theorem 3.1 then implies (57), so Proposition D.8 applies.

For the signal term, the exact risk weight gives $(\lambda_j/k_j^2)Z_j(0)^2 = w_j$. For the noise term, substitute (46) and change variables $r = k_j u/\sqrt{B}$. The factors k_j cancel, giving

$$\frac{\lambda_j}{k_j^2}\int_0^{\widehat{t}_j} e^{-c_L(\widehat{t}_j - r)/2}\varepsilon_j(r)^2 \,\mathrm{d}r = \lambda_j \int_0^t e^{-c_L k_j(t-u)/(2\sqrt{B})} n_B(u) \,\mathrm{d}u.$$

Substituting these identities into (61) and summing over j, we obtain

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] \lesssim \sum_{j=1}^{\infty} w_j e^{-c_L k_j t/(2\sqrt{B})} + \int_0^t \left(\sum_{j=1}^{\infty} \lambda_j e^{-c_L k_j (t-u)/(2\sqrt{B})} \right) n_B(u)\, \mathrm{d}u. \tag{62}$$

Since $w_j \asymp j^{-1-qs}$ and $k_j/\sqrt{B} = \lambda_j/a_j \asymp \sqrt{\lambda_j} \asymp j^{-q/2}$, Lemma D.1 with weights w_j, rates $k_j/\sqrt{B}$, and $Q = c_L t/2$ gives

$$\sum_{j=1}^{\infty} w_j e^{-c_L k_j t/(2\sqrt{B})} \asymp (1 + c_L t/2)^{-2s} \asymp (1+t)^{-2s}. \tag{63}$$

Here in the last step we used the fact that c_L is a fixed positive constant. For the noise accumulation term, apply the same lemma with $b_j = \lambda_j$, the same rates $k_j/\sqrt{B}$, $r = q$, $\theta = q/2$, and $Q = c_L(t-u)/2$. This gives

$$\sum_{j=1}^{\infty} \lambda_j e^{-c_L k_j (t-u)/(2\sqrt{B})} \asymp (1 + t - u)^{-(2-2/q)}, \qquad 0 \leqslant u \leqslant t. \tag{64}$$

The noise accumulation estimate also holds at $u = t$, where $Q = 0$. Substituting (63) and (64) into (62) gives the required upper bound.

For the lower bound, apply the same steps to (60), taking $Q = \omega t$ and $Q = \omega(t-u)$ in Lemma D.1. Since ω is also a fixed positive constant, the same signal and noise weights w_j and λ_j give the same power laws. Together, the two bounds prove

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_t)] \asymp (1+t)^{-2s} + \int_0^t (1 + t - r)^{-(2-2/q)} n_B(r)\, \mathrm{d}r.$$

Since $n_B(r) = \sigma_B \varphi(r)$, this establishes the FSL (3) with the bounded-time profiles

$$S(t) \asymp (1+t)^{-2s}, \qquad K(t) \asymp (1+t)^{-(2-2/q)}.$$

These profiles have the large-time decay rates stated in Theorem 3.1, while the shift by 1 keeps the forgetting kernel finite at zero lag. All comparison constants are independent of B, t and the particular schedule φ. This completes the proof of Theorem 3.1.

D.7 Discussion: Initialization and relaxation across spectral modes

The initial cutoff in (56) and the characteristic frontier from Section 6.3 satisfy

$$J_B \asymp B^{1/(1+qs)}, \qquad j_t \asymp t^{2/q}.$$

Here J_B separates modes by their initial normalized error, with $|Z_j(0)| \leqslant 1/2$ for $j \geqslant J_B$. The frontier j_t compares the observation time with the characteristic relaxation time:

$$\tau_j \asymp \frac{\sqrt{B}}{k_j} \asymp j^{q/2}, \qquad \tau_{j_t} \asymp t.$$

Thus, J_B is fixed at a given batch size, while j_t moves to larger indices as training proceeds.

The bound $t \gtrsim B^{1/(2s+2/q)}$ in (57) gives $J_B \lesssim j_t$. Burn-in control also uses $t \gtrsim \sqrt{B}$, which is the stronger bound when $s + 1/q > 1$.

For the ordering $J_B < j_t$ shown in Figure 5, the three regions have the following meanings.

- **Burn-in-time dominated ($j < J_B$).** These modes may start with a large normalized error. When $|Z_j(0)| \gg 1$, the saturated drift gives an initial burn-in scale of order $\sqrt{B}\,|\theta_j^*|$, which can exceed τ_j. Lemma D.7 controls the effect of this initial transient under the training-time requirement.
- **Effectively linear modes ($J_B \leqslant j < j_t$).** These modes start with $|Z_j(0)| \leqslant 1/2$ and have $\tau_j \lesssim t$. Their initial signal decays substantially when $j \ll j_t$, as reflected in the exponential terms of the modal bounds.

- **Unlearned tail ($j > j_t$).** These modes have $\tau_j \gtrsim t$, so modes with $j \gg j_t$ retain most of their initial signal. Summing over this tail gives a signal contribution of order $j_t^{-qs} \eqsim t^{-2s}$.

These regions describe initial conditions and relaxation time scales. The "unlearned" label refers to the initial signal, while update noise contributes through the convolution term in the FSL.

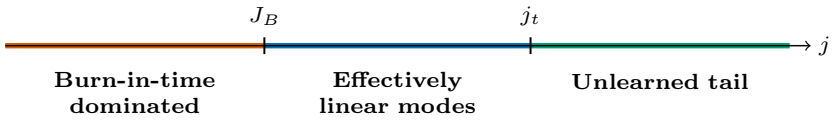


Figure 5: **Regime separation for different modes when $J_B < j_t$.** The left region identifies modes that may require initial burn-in, controlled by Lemma D.7. The middle region contains modes with small initial normalized errors and relaxation times of order t or shorter. The right region contains slower modes whose initial signal remains partly unlearned. The boundaries indicate characteristic scales.

E Data scaling analysis

In this appendix, we analyze the data scaling behavior of SignSGD and SGD. Throughout, we fix the batch size at $B = 1$, so the number of iterations equals the data budget, $N = D/B = D$, where D denotes the total number of samples.

At $B = 1$, the two intrinsic clocks coincide: $\Gamma(z) \coloneqq \Gamma_1(z) = \Gamma_{\mathrm{sgd}}(z) = \int_0^z \eta(u)\,\mathrm{d}u$. We use Γ in the main-text data-scaling analysis. Similarly, $\varphi = \varphi_{\mathrm{sgn}} = \varphi_{\mathrm{sgd}}$ and $n_1(t) = \varphi(t)$. The final SignSGD risk is $\mathcal{E}_D^{\mathrm{sgn}} = \mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_{\Gamma(D)})]$.

E.1 Power decay schedules

The goal of the data scaling analysis is to characterize the best rate achievable by SignSGD or SGD through an appropriate choice of the learning rate schedule. To this end, we introduce the power decay schedule [30] and derive the rate it achieves in the following analysis. In Appendix E.4 we show that this schedule indeed gives the optimal rate. For $k \in \{0, 1, \ldots, N-1\}$, the schedule is defined as

$$\eta_k = \eta_{\mathrm{peak}} \left(1 - \frac{k}{N}\right)^p. \tag{65}$$

It is determined by two hyperparameters: the peak learning rate η_{peak} and the decay rate $p > 0$. In the continuous modeling, the power decay schedule takes the form

$$\eta(z) = \eta_{\mathrm{peak}} \left(1 - \frac{z}{N}\right)^p,$$

where $z \in [0, N]$ is the continuous training step. The intrinsic time is given by

$$t = \Gamma(z) = \int_0^z \eta_{\mathrm{peak}} \left(1 - \frac{u}{N}\right)^p \mathrm{d}u = \frac{\eta_{\mathrm{peak}} N}{p+1} \left[1 - \left(1 - \frac{z}{N}\right)^{p+1}\right].$$

Thus, the total intrinsic time of the entire training process is

$$T_D = \Gamma(D) = \frac{\eta_{\mathrm{peak}} D}{p+1} \eqsim \eta_{\mathrm{peak}} D.$$

The learning rate schedule can be expressed in intrinsic time as

$$\varphi(t) = \eta(\Gamma^{-1}(t)) = \eta_{\mathrm{peak}} \left(1 - \frac{t}{T_D}\right)^{\frac{p}{p+1}}. \tag{66}$$

Intuition behind the power decay schedule. We now provide an intuition for the choice of power decay. In our FSL, the loss is determined by two competing terms: the signal term, which prefers a large total intrinsic time, and the noise term, which prefers sufficiently fast decay near the end of training so that the accumulated noise can be forgotten. Since both SGD and SignSGD have power-law forgetting kernels (see Section 3), a power decay tail is the natural schedule shape for matching the kernel.

E.2 Data scaling for SignSGD (Proof of Theorem 4.1)

We first prove the data scaling result for SignSGD. Let $\mathcal{E}_D^{\mathrm{sgn}} := \mathbb{E}[\mathcal{E}(\vartheta_{T_D})]$ denote the final-step excess risk of SignSGD where $T_D = \Gamma(D)$ is the total intrinsic time. The SignSGD FSL from Theorem 3.1 gives [1]

$$\mathcal{E}_D^{\mathrm{sgn}} \eqsim T_D^{-2s} + \int_0^{T_D} (1 + T_D - t)^{-(2-2/q)} \varphi(t)\, \mathrm{d}t. \tag{67}$$

The following lemma simplifies the FSL under the power decay schedule.

Lemma E.1 (SignSGD FSL under power decay schedules)**.** *Under the assumptions of Theorem 3.1, let $B = 1$ and use the power-decay schedule (66). Let $\alpha_{\mathrm{sgn}} = \min\{\frac{q}{2}, p+1\}$, where p is the power decay exponent. Then the final excess risk satisfies*

$$\mathcal{E}_D^{\mathrm{sgn}} \eqsim (\eta_{\mathrm{peak}} D)^{-2s} + \eta_{\mathrm{peak}} (\eta_{\mathrm{peak}} D)^{-(1-\frac{1}{\alpha_{\mathrm{sgn}}})} \tag{68}$$

up to a logarithmic factor when $p + 1 = \frac{q}{2}$.

Proof. For the power decay schedule, the total intrinsic time satisfies $T_D \eqsim \eta_{\mathrm{peak}} D$. By the SignSGD FSL, it remains to estimate the noise term $\int_0^{T_D} (1 + T_D - t)^{-(2-2/q)} \varphi(t)\, \mathrm{d}t$. Write $u = T_D - t$. Then, under the power decay schedule (66),

$$\varphi(T_D - u) = \eta_{\mathrm{peak}} \left(\frac{u}{T_D}\right)^{\frac{p}{p+1}}.$$

Hence the noise term is

$$\eta_{\mathrm{peak}} T_D^{-\frac{p}{p+1}} \int_0^{T_D} (1+u)^{-(2-2/q)} u^{\frac{p}{p+1}}\, \mathrm{d}u.$$

Let $a = \frac{p}{p+1}$. Since $(1+u)^{-(2-2/q)} u^a \eqsim u^{a-(2-2/q)}$ for large u, we have

$$\int_0^{T_D} (1+u)^{-(2-2/q)} u^a\, \mathrm{d}u \eqsim \begin{cases} 1, & a - (2 - 2/q) < -1, \\ \log T_D, & a - (2-2/q) = -1, \\ T_D^{a-(2-2/q)+1}, & a - (2-2/q) > -1. \end{cases}$$

The boundary condition $a - (2 - 2/q) = -1$ is equivalent to $p + 1 = \frac{q}{2}$. Therefore, the noise term satisfies

$$\int_0^{T_D} (1 + T_D - t)^{-(2-2/q)} \varphi(t)\, \mathrm{d}t \eqsim \eta_{\mathrm{peak}} T_D^{-\left(1-\frac{1}{\alpha_{\mathrm{sgn}}}\right)}$$

up to a logarithmic factor at the boundary, where $\alpha_{\mathrm{sgn}} = \min\{\frac{q}{2}, p+1\}$. Hence we have

$$\mathcal{E}_D^{\mathrm{sgn}} \eqsim (\eta_{\mathrm{peak}} D)^{-2s} + \eta_{\mathrm{peak}} (\eta_{\mathrm{peak}} D)^{-\left(1-\frac{1}{\alpha_{\mathrm{sgn}}}\right)}.$$

□

Optimizing (68) over the peak learning rate gives the data-scaling result. The following theorem is a restatement of Theorem 4.1.

Theorem E.2 (Data scaling for SignSGD)**.** *Under the assumptions of Theorem 3.1, let $B = 1$ and use the power-decay schedule (66). Then the following rates hold.*

- *If $s \geqslant \frac{1}{2} - \frac{1}{q}$, then choosing $p > \frac{q}{2} - 1, \eta_{\mathrm{peak}} \eqsim D^{-\frac{2s-1+2/q}{2s+2/q}}$ yields*

$$\mathcal{E}_D^{\mathrm{sgn}} \eqsim D^{-\frac{sq}{sq+1}}.$$

[1] Here and throughout the data-scaling analysis, we work in the regime $T_D \gg 1$. Accordingly, for the signal learning term we have $(1+T_D)^{-2s} \eqsim T_D^{-2s}$ for SignSGD and $(1+T_D)^{-s} \eqsim T_D^{-s}$ for SGD, and we use the latter form to simplify the derivations.

- *If* $0 < s < \frac{1}{2} - \frac{1}{q}$, *then choosing* $p > \frac{2s}{1-2s}, \eta_{\text{peak}} \eqsim 1$ *yields*

$$\mathcal{E}_D^{\text{sgn}} \eqsim D^{-2s}.$$

Proof. First consider $s \geqslant \frac{1}{2} - \frac{1}{q}$. Choosing $p > \frac{q}{2} - 1$ gives $\alpha_{\text{sgn}} = \frac{q}{2}$. Then, applying Lemma E.1, we have

$$\mathcal{E}_D^{\text{sgn}} \eqsim (\eta_{\text{peak}} D)^{-2s} + \eta_{\text{peak}} (\eta_{\text{peak}} D)^{-(1-2/q)}.$$

Balancing the two terms yields

$$\eta_{\text{peak}} \eqsim D^{-\frac{2s-1+2/q}{2s+2/q}}.$$

Substituting this into the loss formula gives

$$\mathcal{E}_D^{\text{sgn}} \eqsim D^{-\frac{sq}{sq+1}}.$$

If $0 < s < \frac{1}{2} - \frac{1}{q}$, the choice of η_{peak} above grows with D, which violates the stability condition $\eta_{\text{peak}} \lesssim 1$. We therefore take $\eta_{\text{peak}} \eqsim 1$. Choosing $p > \frac{2s}{1-2s}$ gives $\alpha_{\text{sgn}} = \min\{\frac{q}{2}, p+1\} > \frac{1}{1-2s}$. The noise term is $o(D^{-2s})$, including the logarithmic factor when $p + 1 = q/2$. Hence

$$\mathcal{E}_D^{\text{sgn}} \eqsim D^{-2s}.$$

□

E.3 Data scaling of SGD

We now review the data scaling result for SGD. The SGD FSL yields the following achievable rates for power decay [30]. Proposition E.5 supplies the matching lower bounds for all admissible schedules.

For SGD, let $\boldsymbol{\vartheta}_t$ denote the process in (32) and $\mathcal{E}_D^{\text{sgd}} := \mathbb{E}[\mathcal{E}(\boldsymbol{\vartheta}_{T_D})]$ its final expected excess risk at $T_D = \Gamma(D)$. The SGD FSL from Li et al. [27] gives

$$\mathcal{E}_D^{\text{sgd}} \eqsim T_D^{-s} + \int_0^{T_D} (1 + T_D - t)^{-(2-1/q)} \varphi(t) \, \mathrm{d}t. \tag{69}$$

The following lemma summarizes the SGD FSL under power decay schedules.

Lemma E.3 (SGD FSL under power decay schedules; Li et al. [30])**.** *Consider the SGD FSL* (69) *with the power decay schedule* (66)*. Let* $\alpha_{\text{sgd}} = \min\{q, p+1\}$*, where* p *is the power decay exponent. Then the final excess risk satisfies*

$$\mathcal{E}_D^{\text{sgd}} \eqsim (\eta_{\text{peak}} D)^{-s} + \eta_{\text{peak}} (\eta_{\text{peak}} D)^{-\left(1 - \frac{1}{\alpha_{\text{sgd}}}\right)} \tag{70}$$

up to a logarithmic factor when $p + 1 = q$*.*

Proof. The proof is the same as that of Lemma E.1, with the SignSGD signal exponent $2s$ replaced by the SGD signal exponent s, and the effective capacity exponent $q/2$ replaced by q. Therefore, with $\alpha_{\text{sgd}} = \min\{q, p+1\}$, the claimed result follows from the same calculation. □

Optimizing (70) over the peak learning rate gives the data-scaling result:

Theorem E.4 (Data scaling for SGD; Li et al. [30])**.** *Consider the intrinsic time SDE for SGD* (32) *with the power decay schedule* (66)*. Then we have:*

- *If* $s \geqslant 1 - \frac{1}{q}$, *then choosing* $p > q - 1, \eta_{\text{peak}} \eqsim D^{-\frac{s-1+1/q}{s+1/q}}$ *yields*

$$\mathcal{E}_D^{\text{sgd}} \eqsim D^{-\frac{sq}{sq+1}}.$$

- *If* $0 < s < 1 - \frac{1}{q}$, *then choosing* $p > \frac{s}{1-s}, \eta_{\text{peak}} \eqsim 1$ *yields*

$$\mathcal{E}_D^{\text{sgd}} \eqsim D^{-s}.$$

Proof. First consider $s \geqslant 1 - \frac{1}{q}$. Choosing $p > q - 1$ gives $\alpha_{\mathrm{sgd}} = q$, then applying Lemma E.3, we have

$$\mathcal{E}_D^{\mathrm{sgd}} \eqsim (\eta_{\mathrm{peak}} D)^{-s} + \eta_{\mathrm{peak}} (\eta_{\mathrm{peak}} D)^{-(1-1/q)}.$$

Balancing the signal and noise terms gives $\eta_{\mathrm{peak}} \eqsim D^{-\frac{s-1+1/q}{s+1/q}}$ and the minimax rate $D^{-sq/(sq+1)}$.

Now consider $0 < s < 1 - \frac{1}{q}$, the choice of η_{peak} above grows with D, which violates the stability condition $\eta_{\mathrm{peak}} \lesssim 1$. We therefore take $\eta_{\mathrm{peak}} \eqsim 1$. Choosing $p > \frac{s}{1-s}$ gives $\alpha_{\mathrm{sgd}} = \min\{q, p+1\} > \frac{1}{1-s}$. The noise term is $o(D^{-s})$, including the logarithmic factor when $p + 1 = q$. Hence

$$\mathcal{E}_D^{\mathrm{sgd}} \eqsim D^{-s}.$$

□

E.4 Optimality of power decay schedules in data scaling

We compare all admissible learning rate schedules at the same sample budget D. A lower bound on accumulated noise limits the benefit of increasing the total intrinsic time, while the peak learning rate bounds the time available for learning the signal.

Proposition E.5 (Optimal rates over bounded learning rate schedules)**.** *Consider the SignSGD and SGD models of Appendix E.2 and Appendix E.3 with $B = 1$. Let $\mathcal{A}_D$ consist of all nonnegative schedules on $[0, D]$ bounded by a fixed admissible peak $C_0 > 0$, independent of D. Then*

$$\inf_{\eta \in \mathcal{A}_D} \mathcal{E}_D^{\mathrm{sgn}} \eqsim D^{-\min\{2s,\, sq/(sq+1)\}}, \qquad \inf_{\eta \in \mathcal{A}_D} \mathcal{E}_D^{\mathrm{sgd}} \eqsim D^{-\min\{s,\, sq/(sq+1)\}}.$$

Proof. Fix $\eta \in \mathcal{A}_D$ and write $T = \Gamma(D)$. For SignSGD, summing the global moment lower bound in Corollary D.3 by Lemma D.1, as in Appendix D.6, bounds the risk below by the sum of the signal and noise terms for every $T \geqslant 0$. For SGD the corresponding lower bound follows from (69) when $T \geqslant 1$. For $T < 1$, the first coordinate of the mean SGD error retains the factor $e^{-\lambda_1 T}$ from (32), so its risk is bounded below by a positive constant. The SignSGD signal lower bound gives the same conclusion. It suffices to consider $T \geqslant 1$.

For a forgetting kernel $(1+u)^{-\beta}$, write the noise contribution in training steps. Cauchy–Schwarz on $[0, D]$ gives

$$\begin{aligned}\int_0^D (1 + T - \Gamma(z))^{-\beta} \eta(z)^2 \, \mathrm{d}z &\geqslant \frac{1}{D} \left[\int_0^D (1 + T - \Gamma(z))^{-\beta/2} \eta(z) \, \mathrm{d}z \right]^2 \\ &= \frac{1}{D} \left[\int_0^T (1+u)^{-\beta/2} \, \mathrm{d}u \right]^2.\end{aligned}$$

The equality uses $\Gamma'(z) = \eta(z)$ almost everywhere and also holds on schedules with zero learning rates, including entire intervals. With $\beta = 2 - 2/q$ for SignSGD and $\beta = 2 - 1/q$ for SGD, adding the signal lower bounds yields

$$\mathcal{E}_D^{\mathrm{sgn}} \gtrsim T^{-2s} + \frac{T^{2/q}}{D}, \qquad \mathcal{E}_D^{\mathrm{sgd}} \gtrsim T^{-s} + \frac{T^{1/q}}{D}.$$

The two terms balance at $T = D^{1/(2s+2/q)}$ for SignSGD and at $T = D^{1/(s+1/q)}$ for SGD, giving a lower bound of order $D^{-sq/(sq+1)}$ in both cases. The cap gives $T \leqslant C_0 D$, so the signal terms also bound the respective risks below by constant multiples of D^{-2s} and D^{-s}. Taking the larger of the two lower bounds for each optimizer gives the asserted exponents. The power decay schedules in Theorem E.2 and Theorem E.4, with the peak prefactors chosen to satisfy the cap, attain these rates. □

When SignSGD attains $D^{-sq/(sq+1)}$, its exponent agrees with the statistical minimax benchmark discussed in Section 4.1. The proposition establishes optimality over schedules within the continuous models.

F Batch-size scaling analysis

We prove Theorem 4.4 by comparing the intrinsic time needed to preserve the optimal small-batch risk with the time available at batch size B. This gives one argument for all task regimes in the theorem. Write $\alpha = s + 1/q$ and $N = D/B$.

Proof. **The small-batch target.** Let $T = \Gamma_{\mathrm{sgd}}(N)$ and $U = \Gamma_B(N)$ denote the final intrinsic times of SGD and SignSGD. The data-scaling analysis in Appendix E identifies the optimal $B = 1$ time scales

$$T_* = \min\{D, D^{1/\alpha}\}, \qquad U_* = \min\{D, D^{1/(2\alpha)}\}.$$

By Proposition E.5, the corresponding optimal risks satisfy

$$\mathcal{E}_\star^{\mathrm{sgd}}(D, 1) \eqsim T_*^{-s}, \qquad \mathcal{E}_\star^{\mathrm{sgn}}(D, 1) \eqsim U_*^{-2s}.$$

Thus the question is how large B can be while still reaching these time scales.

The time available at batch size B. Since there are only $N = D/B$ updates, the peak learning rate bounds $\eta_{\mathrm{peak}}^{\mathrm{sgd}} \lesssim 1$ and $\sqrt{B}\,\eta_{\mathrm{peak}}^{\mathrm{sgn}} \lesssim 1$ imply

$$T \lesssim \frac{D}{B}, \qquad U \lesssim \frac{D}{B}.$$

The signal lower bounds for SGD and Corollary D.3 for SignSGD hold for every schedule, including before burn-in. Consequently, for $1 \leqslant B \leqslant D$,

$$\mathcal{E}_\star^{\mathrm{sgd}}(D, B) \gtrsim \left(\frac{B}{D}\right)^s, \qquad \mathcal{E}_\star^{\mathrm{sgn}}(D, B) \gtrsim \left(\frac{B}{D}\right)^{2s}.$$

Comparing with the small-batch risks above, preserving a fixed risk ratio requires, respectively,

$$B \lesssim \frac{D}{T_*} = \max\{1, D^{1-1/\alpha}\}, \qquad B \lesssim \frac{D}{U_*} = \max\{1, D^{1-1/(2\alpha)}\}.$$

These are the claimed CBS upper bounds.

Reaching the target times. Use the power-decay schedule with fixed $p > q - 1$ and the peak learning rates stated after Theorem 4.4. Within the respective batch-size ranges above, they give $T \eqsim T_*$ and $U \eqsim U_*$. For SignSGD, we also need the training-time conditions

$$U \gtrsim \sqrt{B}, \qquad U \gtrsim B^{1/(2\alpha)}.$$

If $\alpha \leqslant 1/2$, the claimed batch range is bounded and $U \eqsim D$, so both conditions hold for large D. If $1/2 < \alpha < 3/2$, substituting $B \lesssim D^{1-1/(2\alpha)}$ and $U \eqsim D^{1/(2\alpha)}$ verifies both conditions. The power-decay estimates from Appendix E therefore give

$$\mathcal{E}^{\mathrm{sgd}}(D, B; \eta) \lesssim T_*^{-s} + \frac{T_*^{1/q}}{D} \eqsim T_*^{-s},$$
$$\mathcal{E}^{\mathrm{sgn}}(D, B; \eta) \lesssim U_*^{-2s} + \frac{U_*^{2/q}}{D} \eqsim U_*^{-2s},$$

where $\mathcal{E}^{\mathrm{sgd}}(D, B; \eta)$ and $\mathcal{E}^{\mathrm{sgn}}(D, B; \eta)$ denote the final risks under these schedules. Hence both methods retain their respective small-batch risk throughout the claimed ranges. If the claimed CBS is of constant order, the matching lower bound already follows from $B_{\mathrm{crit}}(D) \geqslant 1$.

The fixed comparison factor. Choose common constants $0 < c_- \leqslant C_+$ such that each optimizer's $B = 1$ optimal risk is at least c_- times its rate above, and the constructed schedules have risk at most C_+ times that rate. These constants do not depend on D or B. For each optimizer, the constructed schedule therefore satisfies

$$\frac{\mathcal{E}(D, B; \eta)}{\mathcal{E}_\star(D, 1)} \leqslant \frac{C_+}{c_-}$$

throughout its claimed batch-size range. Fix the comparison factor $1+\delta$ in Definition 4.3 sufficiently large to absorb this ratio, independently of D and B. The constructed schedules then give the CBS lower bounds for this fixed comparison factor. Together with the upper bounds, this proves (6). □

The super-easy regime. Along the balanced SignSGD schedule, $U \eqsim D^{1/(2\alpha)}$. The burn-in requirement $U \gtrsim \sqrt{B}$ gives $B \lesssim D^{1/\alpha}$, while the peak learning rate cap gives $B \lesssim D^{1-1/(2\alpha)}$. At $s = 3/2 - 1/q$ both bounds have order $D^{2/3}$, and the constants determine which one is reached first. For $s > 3/2 - 1/q$, the burn-in condition restricts this schedule at a smaller batch scale. Beyond that scale the two-sided FSL for the balanced schedule no longer gives the optimal-risk comparison in Theorem 4.4.

G Heuristic extension to RMSprop and Adam

This appendix develops a heuristic extension of the SignSGD analysis to RMSprop and Adam. We work in the diagonal model and approximate the second-moment normalizer by the coordinate-wise gradient-noise variance after burn-in. For Adam, we additionally require a separation between its momentum memory and the coordinate relaxation time. These assumptions yield effective dynamics with the same FSL exponents and batch-size dependence as SignSGD. These results are supported by our experiments involving Adam and RMSprop in Appendix H.2.

Ignoring the ϵ in the denominator and bias correction, the Adam update is

$$\mathbf{m}_{k+1} = \beta_1 \mathbf{m}_k + (1-\beta_1)\mathbf{g}_k,\ \mathbf{v}_{k+1} = \beta_2 \mathbf{v}_k + (1-\beta_2)\mathbf{g}_k^2, \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{m}_{k+1}/\sqrt{\mathbf{v}_{k+1}},$$

where all operations are coordinate-wise, with $\mathbf{m}_0 = \mathbf{v}_0 = \mathbf{0}$. RMSprop corresponds to the case $\beta_1 = 0$. To expose the main mechanism, we first focus on RMSprop and then discuss Adam as a momentum-filtered variant.

G.1 Second moments and coordinate curvatures

Consider one coordinate with curvature λ and write the stochastic gradient as $g_k = \mu_k + \xi_k$ with μ_k being the full-batch gradient. By Assumption 2.5, in the label-noise-dominated regime, $\mathbb{E}[\xi_k] = 0$ and $\mathbb{E}[\xi_k^2] \eqsim \sigma_B^2 \lambda$, where $\sigma_B = 1/\sqrt{B}$ is the batch noise scale from Section 2. For the following heuristic scaling calculation, we suppress fixed noise–curvature comparison constants and use $\sigma_B^2\lambda$ as a representative variance. In the noise-dominated regime: $\mu_k^2 \ll \mathbb{E}[\xi_k^2]$, we have $v_{k+1} = (1-\beta_2)\sum_{j=0}^{k} \beta_2^{k-j} g_j^2 \approx (1-\beta_2)\sum_{j=0}^{k} \beta_2^{k-j}\xi_j^2$. Approximating the noise across steps by independent draws with fixed variance, for $k(1-\beta_2) \gg 1$:

$$\mathbb{E}[v_{k+1}] \approx (1-\beta_2^{k+1})\sigma_B^2\lambda \approx \sigma_B^2\lambda,$$

$$\mathrm{Var}[v_{k+1}] \approx \frac{(1-\beta_2)(1-\beta_2^{2(k+1)})}{1+\beta_2}\,\mathrm{Var}[\xi^2] \eqsim (1-\beta_2)\sigma_B^4\lambda^2,$$

where we use $\mathbb{E}[\xi^4] \lesssim (\mathbb{E}[\xi^2])^2$ for Gaussian noise. Hence, $v_{k+1} \eqsim \sigma_B^2\lambda\left[1 + O_{\mathbb{P}}\left(\sqrt{1-\beta_2}\right)\right]$. When β_2 is close to 1, the EMA self-averages and for $k(1-\beta_2) \gg 1$,

$$v_{k+1} \eqsim \sigma_B^2\lambda.$$

Thus, in the noise-dominated regime, RMSprop and Adam estimate a coordinate-wise gradient-noise variance proportional to curvature.

G.2 The effective RMSprop dynamics

Let $u_k = \theta_k - \theta^*$ be the deviation along the coordinate with curvature λ. Using the representative normalization $\sqrt{v_{k+1}} \approx \sigma_B\sqrt{\lambda}$, the RMSprop update in the noise-dominated regime becomes

$$u_{k+1} \approx u_k - \eta_k \frac{\lambda u_k + \xi_k}{\sigma_B\sqrt{\lambda}} = u_k - \eta_k\left(\sigma_B^{-1}\sqrt{\lambda}u_k + \zeta_k\right), \tag{71}$$

where $\zeta_k = \xi_k/(\sigma_B\sqrt{\lambda})$ is the normalized noise satisfying $\mathbb{E}[\zeta_k^2] \eqsim 1$. In the continuous model, let $u_j(t)$ be the j-th coordinate deviation at intrinsic time $t = \Gamma_B(z) = \sigma_B^{-1}\int_0^z \eta(v)\,\mathrm{d}v$. We use $\varphi(t) = \varphi_{\mathrm{sgn}}(t) = \eta(\Gamma_B^{-1}(t))$ and $n_B(t) = \varphi(t)/\sqrt{B}$, as in the main text. With the fixed normalized noise-variance factor also suppressed, the representative effective dynamics become

$$\mathrm{d}u_j(t) = -\sqrt{\lambda_j}u_j(t)\,\mathrm{d}t + \sqrt{\frac{\varphi(t)}{\sqrt{B}}}\,\mathrm{d}W_j(t). \tag{72}$$

Therefore, the second moment of the j-th coordinate is given by

$$\mathbb{E}[u_j(t)^2] = u_j(0)^2 e^{-2\sqrt{\lambda_j}t} + \frac{1}{\sqrt{B}} \int_0^t e^{-2\sqrt{\lambda_j}(t-t')} \varphi(t')\, \mathrm{d}t'. \tag{73}$$

Remark G.1 (Characteristic time). The j-th coordinate dynamics has characteristic time $\tau_j \eqsim \lambda_j^{-1/2}$ in intrinsic time. In contrast, the characteristic time of SGD is λ_j^{-1}. This means the coordinate-wise normalization introduces acceleration by compressing the characteristic time from λ^{-1} to $\lambda^{-1/2}$, which is curvature adaptive. Batch size accelerates the intrinsic clock through the factor $\sqrt{B}$ in its definition.

G.3 The FSL of RMSprop

We sum the coordinate-wise dynamics over the infinite power-law spectrum to identify the signal-learning and noise-forgetting profiles.

Multiplying (73) by λ_j and summing over coordinates gives the dynamics of total loss $\frac{1}{2}\mathbb{E}[\mathbf{u}_t^\top \mathbf{H} \mathbf{u}_t]$:

$$\frac{1}{2}\mathbb{E}[\mathbf{u}_t^\top \mathbf{H} \mathbf{u}_t] \eqsim \sum_j \lambda_j u_j(0)^2 e^{-2\sqrt{\lambda_j}t} + \frac{1}{\sqrt{B}} \int_0^t \sum_j \lambda_j e^{-2\sqrt{\lambda_j}(t-t')} \varphi(t')\, \mathrm{d}t'. \tag{74}$$

Hence, the loss dynamics of RMSprop admit the FSL with noise injection intensity $n_B(t) = \varphi(t)/\sqrt{B}$ and the signal learning and forgetting kernel given by

$$S_{\mathrm{rms}}(t) = \sum_j \lambda_j u_j(0)^2 e^{-2\sqrt{\lambda_j}t}, \quad K_{\mathrm{rms}}(t) = \sum_j \lambda_j e^{-2\sqrt{\lambda_j}t}.$$

The origin of power-law scaling. Under the power-law assumption (Assumption 2.3): $\lambda_j \eqsim j^{-q}$ and $w_j := \lambda_j u_j(0)^2 \eqsim j^{-(1+sq)}$, we have

$$S_{\mathrm{rms}}(t) \eqsim (1+t)^{-2s}, \qquad K_{\mathrm{rms}}(t) \eqsim (1+t)^{-(2-2/q)}.$$

We now provide a physical-style derivation of why they follow the above power-law decay. The coordinate with curvature λ_j has characteristic time $\tau_j \eqsim \lambda_j^{-1/2}$. Take the signal-learning term as an example, $S(t) = \sum_j w_j e^{-2\sqrt{\lambda_j}t}$. At time t, the modes with $\tau_j \lesssim t$ have essentially been learned, while the modes with $\tau_j \gtrsim t$ are still largely unlearned. Since $\lambda_j \eqsim j^{-q}$, the moving frontier j_t is determined by $\tau_{j_t} \eqsim t$, yielding $j_t \eqsim t^{2/q}$, as in Section 6.3. For $t \gtrsim 1$,

$$S_{\mathrm{rms}}(t) = \sum_{j \leqslant j_t} w_j \underbrace{e^{-2\sqrt{\lambda_j}t}}_{\approx 0} + \sum_{j > j_t} w_j \underbrace{e^{-2\sqrt{\lambda_j}t}}_{\approx 1} \approx \sum_{j > j_t} w_j \eqsim (j_t)^{-sq} \eqsim t^{-2s}. \tag{75}$$

The power-law scaling of the forgetting kernel follows from the same moving-cutoff argument.

Remark G.2 (Tail dominates scaling law). The above derivation suggests a simple picture of learning in power-law data. The expected excess risk is a sum over spectral modes, and each mode has its own characteristic time τ_j. At time t, the moving frontier j_t is determined by $\tau_{j_t} \eqsim t$. Modes with $\tau_j \lesssim t$ have already been learned and contribute little to the remaining signal error, whereas modes with $\tau_j \gtrsim t$ are still largely unlearned. Thus the signal-learning term is essentially the tail mass of the unlearned modes. Analogously, the power-law decay of $S(t)$ therefore comes from how fast this moving frontier advances through the spectrum. In summary,

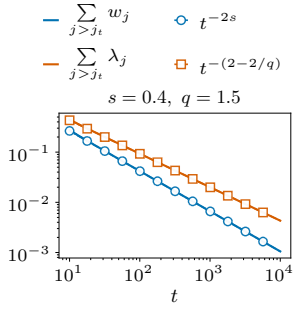


Figure 6: Tail-sum scaling.

$$S(t) \approx \sum_{j \gtrsim j_t} w_j, \qquad K(t) \approx \sum_{j \gtrsim j_t} \lambda_j, \qquad \text{with } \tau_{j_t} \eqsim t. \tag{76}$$

Figure 6 demonstrates that the above moving-tail approximation indeed yields the right scaling for both signal-learning and noise-forgetting terms.

G.4 The FSL of Adam

Adam differs from RMSprop only through the first-moment accumulator in the numerator. For a complementary analysis of how momentum affects data scaling and the critical batch size in one-pass kernel regression, see Wang et al. [53]. Under the same denominator approximation $v_{k+1} \approx \sigma_B^2 \lambda$, the first-moment EMA has an effective memory window $M_1 \asymp (1-\beta_1)^{-1}$ steps. After burn-in, $k \gg M_1$, its expansion is

$$m_{k+1} = (1-\beta_1) \sum_{j=0}^{k} \beta_1^j \mu_{k-j} + (1-\beta_1) \sum_{j=0}^{k} \beta_1^j \xi_{k-j}.$$

Assume that the learning rate varies little across this window and that $\sqrt{B}\,\eta_k M_1 \ll \tau_\lambda$, where $\tau_\lambda \asymp \lambda^{-1/2}$ is the intrinsic relaxation time. Recall from (71) that $\zeta_k = \xi_k/(\sigma_B\sqrt{\lambda})$ is the normalized gradient noise. Its first-moment EMA, denoted by $\widetilde{\zeta}_k$, gives the noise contribution to the normalized Adam update. The drift contribution is then approximately *quasi-static*, giving

$$u_{k+1} \approx u_k - \eta_k \left(\sigma_B^{-1} \sqrt{\lambda} u_k + \widetilde{\zeta}_k \right), \qquad \widetilde{\zeta}_k := (1-\beta_1) \sum_{j=0}^{k} \beta_1^j \zeta_{k-j}. \tag{77}$$

Assume that the normalized noises ζ_k are independent and identically distributed, with mean zero and variance $\sigma_\zeta^2 := \mathrm{Var}[\zeta_k] \asymp 1$. In the stationary approximation after burn-in,

$$\mathrm{Var}[\widetilde{\zeta}_k] \approx (1-\beta_1)^2 \sum_{j=0}^{\infty} \beta_1^{2j} \sigma_\zeta^2 = \frac{1-\beta_1}{1+\beta_1} \sigma_\zeta^2, \quad \mathrm{Cov}(\widetilde{\zeta}_k, \widetilde{\zeta}_{k+h}) \approx \beta_1^h \,\mathrm{Var}[\widetilde{\zeta}_k], \quad h \geqslant 0.$$

The first equation gives the single-step variance reduction. The second equation shows that different steps remain positively correlated because their moving averages share past noise. For a block of $L \gg M_1$ steps, the variance of the accumulated noise therefore satisfies

$$\frac{1}{L} \mathrm{Var}\left[\sum_{r=0}^{L-1} \widetilde{\zeta}_{k+r} \right] \approx \sigma_\zeta^2 \frac{1-\beta_1}{1+\beta_1} \left[1 + 2 \sum_{h=1}^{L-1} \left(1 - \frac{h}{L} \right) \beta_1^h \right] \approx \sigma_\zeta^2 \frac{1-\beta_1}{1+\beta_1} \left(1 + \frac{2\beta_1}{1-\beta_1} \right) = \sigma_\zeta^2.$$

Momentum spreads the noise of each step across its memory window. For $L \gg M_1$, these contributions accumulate within the block, giving the same leading noise variance $L\sigma_\zeta^2$ as the unfiltered noise despite smoother individual updates.

Now choose a block with $M_1 \ll L \ll \tau_\lambda/(\sqrt{B}\,\eta_k)$, and assume that the learning rate remains approximately constant over this block. The intrinsic time elapsed over this block and the accumulated parameter-noise variance are

$$\Delta t \approx \sqrt{B}\, L \eta_k, \qquad \mathrm{Var}\left[\sum_{r=0}^{L-1} \eta_{k+r} \widetilde{\zeta}_{k+r} \right] \approx L \eta_k^2 \sigma_\zeta^2 \approx \frac{\sigma_\zeta^2 \eta_k}{\sqrt{B}} \Delta t.$$

Together with the drift in (77), this yields the effective intrinsic-time SDE

$$\mathrm{d}u(t) = -\sqrt{\lambda} u(t)\, \mathrm{d}t + \sqrt{\frac{\sigma_\zeta^2 \varphi(t)}{\sqrt{B}}}\, \mathrm{d}W_t.$$

This is the same form as the RMSprop model (72), with $\sigma_\zeta^2 \asymp 1$ absorbed into the constants. Therefore, following the same argument as for RMSprop, we obtain the FSL for Adam with noise injection intensity $n(t) \asymp n_B(t) = \varphi(t)/\sqrt{B}$ and

$$S_{\mathrm{adam}}(t) \asymp (1+t)^{-2s}, \qquad K_{\mathrm{adam}}(t) \asymp (1+t)^{-(2-2/q)}. \tag{78}$$

H Experimental details and additional results

H.1 Fitting and prediction using FSL

Although derived from a kernel regression model, the FSL framework is expected to model loss dynamics in more general training scenarios, as the decomposition into "irreducible, signal, and

noise" components carries a broader physical interpretation. Inspired by the functional form of (3), we heuristically propose a parametrized ansatz:

$$L_\Theta(t;\varphi,B) = \underbrace{L_\infty}_{\text{irreducible loss}} + \underbrace{C_S(A_S+t)^{-\alpha}}_{\text{signal-learning term}} + \underbrace{\frac{C_K}{\sqrt{B}}\int_0^t \big(A_K+t-r\big)^{-\nu}\varphi(r)\,\mathrm{d}r}_{\text{noise-forgetting term}}, \tag{79}$$

Here $t = \Gamma_B(z) = \sqrt{B}\int_0^z \eta(u)\,\mathrm{d}u$ denotes intrinsic time, B is the batch size, $\varphi(t) = \eta(\Gamma_B^{-1}(t))$ is the learning rate schedule in intrinsic time, and $\Theta = (L_\infty, C_S, A_S, \alpha, C_K, A_K, \nu)$ denotes fitting parameters. Thus $\varphi = \varphi_{\text{sgn}}$ and $n_B(t) = \varphi(t)/\sqrt{B}$. The noise-forgetting term in (79) is equivalently $C_K\int_0^t (A_K+t-r)^{-\nu} n_B(r)\,\mathrm{d}r$. We use this base form for synthetic experiments and extend its learning-rate dependence for language model training as described in Appendix H.3.

Discrete FSL. To obtain the left-endpoint form used in the main text, define

$$t_k = \sqrt{B}\sum_{j=0}^{k-1}\eta_j, \qquad I_j = [t_j, t_{j+1}),$$
$$\Delta t_j = t_{j+1} - t_j = \sqrt{B}\,\eta_j.$$

On each nonempty interval I_j, take $\varphi(r) = \eta_j$. The left-endpoint rectangle rule gives

$$\begin{aligned}\frac{1}{\sqrt{B}}\int_0^{t_k} K(t_k-r)\varphi(r)\,\mathrm{d}r &\approx \frac{1}{\sqrt{B}}\sum_{j=0}^{k-1} K(t_k-t_j)\varphi(t_j)\Delta t_j \\ &= \sum_{j=0}^{k-1} K(t_k-t_j)\eta_j^2.\end{aligned}$$

Steps with $\eta_j = 0$ contribute zero. With $K(u) = (A_K+u)^{-\nu}$, this yields the parameterized form in (7).

Numerical convention used in the reported fits. We discretize the noise integral on the cumulative-learning-rate grid

$$\tau_k = \sum_{j=0}^{k-1}\eta_j, \qquad \tau_0 = 0, \qquad \Delta\tau_j = \tau_{j+1} - \tau_j = \eta_j.$$

Here k is the number of completed updates, so ℓ_k is the loss at $\boldsymbol{\theta}_k$ and $t_k = \sqrt{B}\,\tau_k$ is the same intrinsic clock as in the main text. The first stored loss corresponds to $k = 1$. Thus the signal term is $C_S(A_S + \sqrt{B}\,\tau_k)^{-\alpha}$.

To discretize the noise integral in (79), set $r = \sqrt{B}\,v$ and write $\psi(v) = \varphi(\sqrt{B}\,v)$ for the learning rate expressed in cumulative-learning-rate time. Then the noise term at t_k is

$$\frac{C_K}{\sqrt{B}}\int_0^{t_k} K(t_k-r)\varphi(r)\,\mathrm{d}r = C_K\int_0^{\tau_k} K\Big(\sqrt{B}(\tau_k-v)\Big)\,\psi(v)\,\mathrm{d}v,$$

where $K(u) = (A_K+u)^{-\nu}$. On each interval $I_j = (\tau_j, \tau_{j+1}]$, take $\psi(v) = \eta_j$. The right-endpoint rectangle rule gives

$$\begin{aligned}\int_0^{\tau_k} K\Big(\sqrt{B}(\tau_k-v)\Big)\,\psi(v)\,\mathrm{d}v &\approx \sum_{j=0}^{k-1} K\Big(\sqrt{B}(\tau_k-\tau_{j+1})\Big)\,\eta_j\Delta\tau_j \\ &= \sum_{j=0}^{k-1} K\Big(\sqrt{B}(\tau_k-\tau_{j+1})\Big)\,\eta_j^2.\end{aligned}$$

```
adam_optimizer = torch.optim.Adam(
    model.parameters(),
    lr=0.05,
    weight_decay=1e-6,
    betas=(0.9, 0.95),
    eps=1e-8,
)
```

```
lbfgs_optimizer = torch.optim.LBFGS(
    model.parameters(),
    lr=0.1,
    max_iter=50,
    tolerance_grad=1e-20,
    tolerance_change=1e-20,
    history_size=100,
    line_search_fn="strong_wolfe",
)
```

Figure 7: Optimizer configurations used in the two-stage FSL fitting procedure. **(Left)** Stage 1: Adam. **(Right)** Stage 2: L-BFGS.

Steps with $\eta_j = 0$ have zero interval length and contribute zero. The resulting predictor is

$$\begin{aligned}\widehat{L}_\Theta(k;\boldsymbol{\eta},B) = L_\infty + C_S(A_S + \sqrt{B}\,\tau_k)^{-\alpha} \\ + C_K \sum_{j=0}^{k-1} \big(A_K + \sqrt{B}(\tau_k - \tau_{j+1})\big)^{-\nu} \eta_j^2.\end{aligned} \tag{80}$$

The discretization step is $\Delta\tau_j = \eta_j$. Batch size enters the signal and forgetting arguments through $\sqrt{B}$. Equation (80) records the right-endpoint convention used for the reported fits and predictions. The main-text expression (7) uses left endpoints. Both discretize the continuous FSL (79), but they need not agree at finite step sizes.

Fitting the discrete FSL. For a loss trajectory with learning rate schedule $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_{N-1})$, batch size B, and losses $(\ell_1, \ldots, \ell_N)$, the parameters Θ are estimated by minimizing the mean squared error (MSE):

$$\min_\Theta \frac{1}{|\mathcal{I}|} \sum_{k\in\mathcal{I}} \big(\widehat{L}_\Theta(k;\boldsymbol{\eta},B) - \ell_k\big)^2, \tag{81}$$

where $\mathcal{I} = \{N_{\text{skip}}+1, \ldots, N\}$ denotes the set of step indices used for fitting (the first N_{skip} steps are excluded). We optimize Θ using a two-stage procedure: 2000 steps of Adam followed by 30 steps of L-BFGS. The optimizer configurations are shown in Figure 7.

In some experiments, we fit multiple loss trajectories simultaneously. The fitting objective is the average MSE:

$$\min_\Theta \frac{1}{M} \sum_{m=1}^{M} \frac{1}{|\mathcal{I}^{(m)}|} \sum_{k\in\mathcal{I}^{(m)}} \big(\widehat{L}_\Theta(k;\boldsymbol{\eta}^{(m)},B^{(m)}) - \ell_k^{(m)}\big)^2,$$

where $m \in \{1, \cdots, M\}$ denotes the index of the loss trajectory. Each trajectory uses its own cumulative-learning-rate grid $\tau_k^{(m)} = \sum_{j=0}^{k-1} \eta_j^{(m)}$. Its intrinsic time is $\sqrt{B^{(m)}}\tau_k^{(m)}$.

After fitting the parameters Θ, we use the discrete FSL (80) to predict loss trajectories for held-out learning rate schedules or batch sizes.

H.2 PLK regression experiments

The loss-curve and data-scaling experiments in Section 5.1 use $(s,q) = (0.3, 4)$, which satisfies

$$\frac{1}{2} - \frac{1}{q} \leqslant s < 1 - \frac{1}{q}.$$

As discussed in Section 4.1, this is the regime where SignSGD attains the statistically optimal data scaling rate, whereas SGD does not.

In all experiments, we set dimension $d = 512$, eigenvalues $\lambda_j = j^{-q}$, and target magnitudes $|\theta_j^*| = j^{-\frac{1}{2}(sq-q+1)}$ for $j = 1, \ldots, d$. We fix the noise variance $\sigma^2 = 1.0$, and initialize at $\boldsymbol{\theta}_0 = \mathbf{0}$. Unless otherwise stated, the default peak learning rate and batch size are $\eta_{\text{peak}} = 0.1$ and $B = 1$,

respectively. We report the population excess risk after k updates,

$$\ell_k = \mathcal{E}(\boldsymbol{\theta}_k) = \frac{1}{2}\sum_{j=1}^{d} \lambda_j (\theta_{k,j} - \theta_j^*)^2.$$

Across the synthetic experiments, we use SGD, SignSGD, RMSprop and Adam; the optimizer subset depends on the experiment. For RMSprop, we set $\beta_2 = 0.9$ and $\epsilon = 10^{-8}$. For Adam, we set $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. For the loss-curve and data-scaling experiments, we report the excess risk averaged over 200 independent runs unless otherwise stated.

H.2.1 Risk dynamics across learning rate schedules

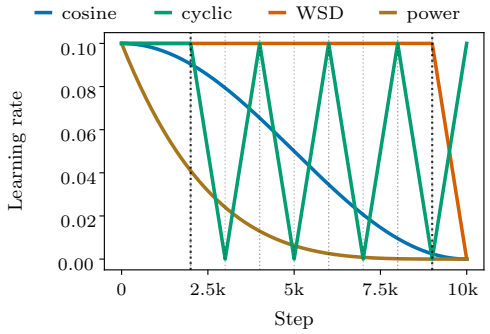


Figure 8: Learning rate schedules in synthetic experiments.

Fitting and prediction for different schedules. We provide experimental details for risk-curve fitting and prediction. For each optimizer, we fit the FSL on cosine [34] and WSD [17] schedules and evaluate its predictive performance on the held-out cyclic schedule [47]. See Figure 8 for a visualization of these schedules. For SignSGD in Figure 1 (left), we set $(s, q, d) = (0.3, 4, 512)$, $B = 1$, and $\sigma^2 = 1$, initialize at zero, and train for 10,000 updates with peak learning rate 0.1. For each schedule, we report the unsmoothed population excess risk averaged over 200 independent runs. We fit six shared parameters in (80) and keep them fixed for the held-out cyclic prediction. We use the fitting procedure of Appendix H.1, consisting of 2000 Adam steps followed by 30 L-BFGS steps, with $L_\infty = 0$ for excess risk and $N_{\text{skip}} = 500$ unless specified otherwise.

For Adam in Figure 9 (left), we set $N_{\text{skip}} = 1000$ and fit piecewise-linear interpolants of the risk curves at zero-based stored indices 1000 through 9999.

Decomposition and intrinsic-time collapse across schedules. Figure 2 uses the same experimental settings and fitted FSL parameters as Figure 1 (left). Figure 2 (a) shows the signal-learning and noise-forgetting terms for the three LR schedules. Figure 2 (b) plots the signal-learning terms against intrinsic time $t_k = \sqrt{B}\,\tau_k$, showing their collapse across schedules.

H.2.2 Data scaling and critical batch size

Data scaling. We provide details for the SGD/SignSGD data-scaling comparison in Figure 1 (middle) and the adaptive-optimizer comparison in Figure 10 (left). Across these comparisons, we run SGD, SignSGD, RMSprop, and Adam using the power decay schedule [30]:

$$\eta_k = \eta_{\text{peak}} \left(1 - \frac{k}{N}\right)^p, \qquad k = 0, 1, \ldots, N-1.$$

Here N is the number of updates and $D = BN$. We choose $p = 4$, which satisfies the power-decay conditions in Theorem 4.1 for SignSGD and in Appendix E.3 for SGD at the experimental parameters. We use the same schedule family for RMSprop and Adam to compare their data-scaling behavior.

All runs use batch size $B = 1$, with η_{peak} tuned at each budget. We use $D \in \{1000, 2000, 4000, 8000, 16000, 32000\}$, evaluate SignSGD on 64 independent runs after tuning,

and compute SGD risk by exact second-moment recursion. The four-optimizer comparison in Figure 10 (left) uses the five budgets from 2000 to 32000, gradient mean $2\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, the same noise covariance, and 200 evaluation runs per setting.

Critical batch size. Figure 1 (right) compares SGD and SignSGD. Figure 10 (right) additionally reports Adam and RMSprop. For these batch-size experiments, we use the common PLK setup with $(s, q) = (1, 4)$. We use the budgets and schedule

$$D \in \{2000, 4000, 8000, 16000, 32000, 64000, 128000\}, \qquad N = \lceil D/B \rceil,$$
$$\eta_k = \eta_{\text{peak}}(1 - k/N)^4, \qquad k = 0, \ldots, N-1.$$

The final batch consumes the remaining samples, so every run uses exactly D samples. For every optimizer and (D, B), we tune η_{peak} within $(0, 1]$. RMSprop uses decay 0.9 and Adam uses $(\beta_1, \beta_2) = (0.9, 0.95)$ with bias correction, both with $\epsilon = 10^{-8}$ and no weight decay. Each risk is normalized by the same optimizer's independently evaluated, optimized $B = 1$ risk at the same D. The CBS is the last tested batch in the passing prefix whose normalized risk is at most 1.2, ending at the first tested failure even if later batches pass again. We refine the batch grid adaptively near this first crossing. SGD risks are computed by an exact second-moment recursion. For the normalized optimizers, learning-rate selection and final risk evaluation use independent random streams, and stratified Monte Carlo includes all 512 coordinates. The dashed curves in the CBS plot in Section 5.1 are ordinary least-squares fits in log–log coordinates over all displayed data budgets.

Batch-size experiment. Figure 1 (right) uses the diagonal Gaussian-gradient model with $D = 32{,}000$, $(s, q, d) = (1, 4, 512)$, unit noise, and zero initialization. At every batch size, we run $N = D/B$ updates with power decay $p = 4$ and retune the peak learning rate up to one. SignSGD risks are averaged over 512 runs independent of tuning, while SGD risks are computed exactly. We take the empirical CBS as the last tested batch before the mean risk exceeds 1.2 times its own $B = 1$ value. This gives 2000 for SignSGD and 16 for SGD. Adding the SignSGD cap $\eta_{\text{peak}} \leqslant B^{-1/2}$ gives 256 versus 16. These finite-grid results illustrate the batch-size advantage. The common-cap experiment does not directly test the theoretical CBS exponent.

H.2.3 Additional results for Adam and RMSprop

Figure 9 shows that the FSL captures risk dynamics for both Adam and RMSprop, which also exhibit data-scaling and batch-size advantages over SGD (Figure 10). In Figure 9 (right), the held-out cyclic prediction has a root mean squared error of 1.90×10^{-3} against the mean excess risk over 200 runs. On the same stored entries (zero-based indices $500, \ldots, 9999$), the fitting errors are 1.60×10^{-3} for cosine and 2.46×10^{-3} for WSD.

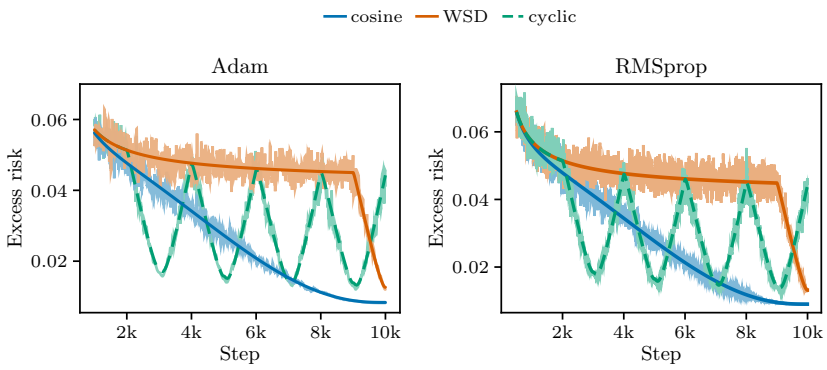


Figure 9: **FSL predictions for Adam and RMSprop.** Risk dynamics for Adam and RMSprop, respectively. Light lines show the mean excess risk over 200 runs. Dark solid lines are FSL fits for cosine and WSD, and dashed lines predict the held-out cyclic schedule.

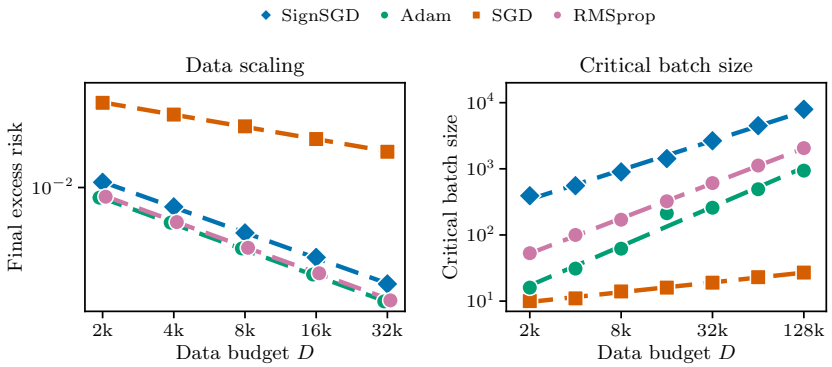


Figure 10: **Data scaling and critical batch size. (Left)** Final excess risk versus data budget for SGD, SignSGD, RMSprop and Adam. **(Right)** Estimated CBS versus data budget for the same optimizers. Dashed lines in both panels denote power-law fits. The CBS uses a 20% risk-increase tolerance relative to each optimizer's $B = 1$ baseline.

H.2.4 Structural-condition experiments

Figure 4 uses quadratic objectives with additive Gaussian gradient noise and the spectral and target-magnitude parameterization above. The target signs are fixed as in the rotation construction below. We set $(s, q, d) = (0.3, 4, 512)$, initialize at zero, and use $B = 1$ and $D = 32{,}000$. All conditions have the same total gradient noise energy. Each optimizer and setting uses a separately tuned peak learning rate in $[10^{-5}, 1]$ and the schedule $\eta_k = \eta_{\text{peak}}(1 - k/N)^4$. Each setting and rotation path is evaluated on 64 held-out stochastic runs. The left panel shows mean risk trajectories, with isotropic-noise controls (iso.) and one fixed rotated geometry (rot.).

For the middle panel, the Hessian remains diagonal and the noise covariance varies as

$$\mathbf{\Sigma}_\gamma = \frac{\operatorname{tr}(\mathbf{H})}{\operatorname{tr}(\mathbf{H}^\gamma)}\mathbf{H}^\gamma, \qquad \gamma \in \{0, 0.1, \ldots, 1\}.$$

The endpoints give isotropic and curvature-aligned noise, respectively. The right panel uses the joint rotations described in the rotation construction below, preserving the Hessian spectrum and target energy in each eigenmode. Each positive rotation strength averages three fixed rotation paths, with one shared unrotated control. Both panels report the ratio of the mean final SGD excess risk to the mean final SignSGD excess risk, so values above one favor SignSGD.

Rotation construction. We use the spectral and target parameterization above, with $(s, q, d) = (0.3, 4, 512)$, $\lambda_j = j^{-q}$, and $|\theta_j^*| = j^{-\frac{1}{2}(sq-q+1)}$. The target has independent random signs sampled once and then fixed. Following the matrix-exponential rotation construction of Xie et al. [59], we draw a matrix $\mathbf{M}$ with independent standard Gaussian entries and set

$$\mathbf{A} = \mathbf{M} - \mathbf{M}^\top, \qquad \mathbf{R}_\rho = \exp(\rho\mathbf{A}),$$
$$\mathbf{H}_0 = \operatorname{diag}(\lambda_1, \ldots, \lambda_d), \qquad \mathbf{H}_\rho = \mathbf{R}_\rho^\top \mathbf{H}_0 \mathbf{R}_\rho, \qquad \boldsymbol{\theta}_\rho^* = \mathbf{R}_\rho^\top \boldsymbol{\theta}^*.$$

No normalization is applied to $\mathbf{A}$. We use three independent generator matrices with seeds 21161024, 21161025, and 21161026, each held fixed across all ρ. A pilot experiment sets the six positive rotation strengths $\rho \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We also evaluate the exactly diagonal control $\rho = 0$ using the same implementation and Gaussian seed protocol.

Training and evaluation. For each fixed ρ, we initialize $\boldsymbol{\theta}_0 = \mathbf{0}$ and set $B = \sigma^2 = 1$. The gradient oracle and population excess risk are

$$\mathbf{g}_k = \mathbf{H}_\rho(\boldsymbol{\theta}_k - \boldsymbol{\theta}_\rho^*) + \boldsymbol{\xi}_k, \qquad \boldsymbol{\xi}_k = \mathbf{R}_\rho^\top \mathbf{H}_0^{1/2} \mathbf{z}_k, \qquad \mathbf{z}_k \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I),$$
$$\ell_k = \mathcal{E}(\boldsymbol{\theta}_k) = \tfrac{1}{2}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_\rho^*)^\top \mathbf{H}_\rho(\boldsymbol{\theta}_k - \boldsymbol{\theta}_\rho^*).$$

Thus $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{H}_\rho)$. This additive Gaussian oracle isolates the noise geometry used in the mechanism analysis, rather than simulating the full state-dependent noise of stochastic regression gradients.

H.3 Language model experiments

Training configuration. We train a 0.1B-parameter LLaMA-style Transformer and track the validation loss throughout training. Unless otherwise specified, each run uses 10B tokens with sequence length 2048, batch size 128, peak learning rate 5×10^{-4}, and gradient clipping threshold 1.0.

We train with SignSGD and Adam. For Adam, we set $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. The schedule-comparison runs use decoupled weight decay with $\lambda = 0.1$ for both optimizers [35], giving the update

$$\boldsymbol{\theta}_{k+1} = (1 - \eta_k \lambda)\boldsymbol{\theta}_k - \eta_k \mathbf{u}_k,$$

where $\mathbf{u}_k$ is the optimizer-specific update at step k.

Learning rate schedules. Figure 11 visualizes the learning rate schedules in our experiments: cosine [34], WSD [17], and 8-1-1 [12]. All schedules incorporate a 2000-step linear warmup from 0 to the peak learning rate η_{peak}. Following warmup, both WSD and 8-1-1 maintain a constant learning rate of η_{peak} until step $0.8N$ (where N denotes the total training steps), after which WSD linearly decays to zero by the end of training, while the 8-1-1 schedule adopts a piecewise constant profile: dropping to $\eta_{\text{peak}}/\sqrt{10}$ at step $0.8N$, and then to $\eta_{\text{peak}}/10$ at step $0.9N$.

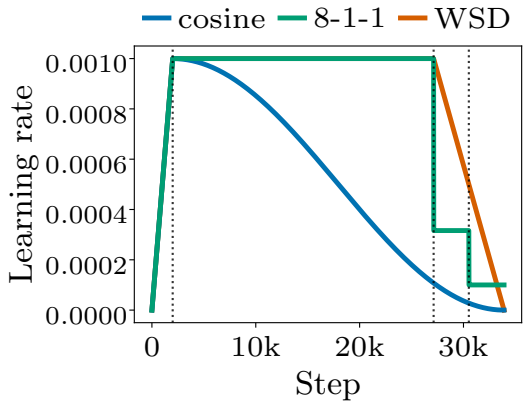


Figure 11: Learning rate schedules for language model training. All schedules use a warmup of 2000 steps and the same peak learning rate.

Details of FSL fitting. We use the fitting procedure described in Appendix H.1 with $N_{\text{skip}} = 2000$.

Loss curve fitting and prediction for different schedules. For Figure 3 (a), we fit the FSL to the cosine and 8-1-1 loss curves and predict the held-out WSD curve. In the fitting objective (81), we use a 100-step moving average of the loss,

$$\ell_k^{\text{smooth}} = \frac{1}{2H} \sum_{j=k-H}^{k+H-1} \widetilde{\ell}_j, \qquad H = 50,$$

where $\widetilde{\ell}_j = \ell_{\min\{N, \max\{1,j\}\}}$ extends the first and last observed losses beyond the endpoints.

Joint fitting across batch sizes under a cosine schedule. We next consider a nonconstant learning rate. Figure 12 (left) reports joint fitting of Adam cosine-schedule runs across all three batch sizes $B \in \{64, 128, 256\}$, with peak learning rate 5×10^{-4}. The FSL closely tracks all three curves, indicating that the same square-root batch-size clock remains effective in this setting. Since all three curves are included in this fit, this experiment tests joint batch-size consistency rather than held-out batch-size extrapolation.

Extrapolation of batch sizes and data size. In Figure 3 (b), we demonstrate that FSL can extrapolate to held-out batch sizes. We now test the simultaneous extrapolation of both batch size and data size. The batch-size comparisons at peak learning rate 5×10^{-4} use the first 80% of the 10B-token runs (about 8B tokens) for $B \in \{64, 128, 256\}$. SignSGD uses a constant rate throughout. The

Adam curves cover the warmup and plateau of the WSD schedule up to the start of decay. We fit a single FSL model using only

- the initial 50% of the $B = 64$ curve,
- the initial 50% of the $B = 128$ curve.

We then evaluate the predictive performance on the remaining parts of the two curves and on the entire held-out $B = 256$ curve. As shown in Figure 12 (right), FSL predicts both the continuation of the fitted batch-size curves and the fully held-out $B = 256$ curve, with small deviations over the longest continuation horizons.

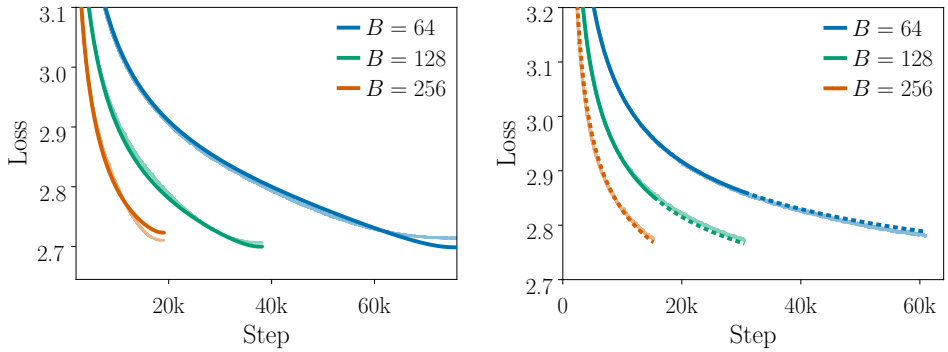


Figure 12: **Joint batch-size fitting and extrapolation for Adam. (Left)** Under a cosine schedule with peak learning rate 5×10^{-4}, the parameters Θ are jointly fitted to all three loss curves for $B \in \{64, 128, 256\}$. **(Right)** Under a constant learning rate $\eta = 5 \times 10^{-4}$, FSL is fitted using the first half of the $B = 64$ and $B = 128$ trajectories, then predicts their remaining parts and the entire held-out $B = 256$ trajectory. Solid portions denote FSL fits, and dashed portions denote predictions on held-out horizons.

Robustness to a larger learning rate. We extend the Adam experiments in Figure 3 (b) and Figure 12 (right) by increasing the learning rate from 5×10^{-4} to 10^{-3}. The model, data, batch sizes, and training horizon are kept unchanged. As shown in Figure 13, FSL continues to capture the loss dynamics under this larger learning rate.

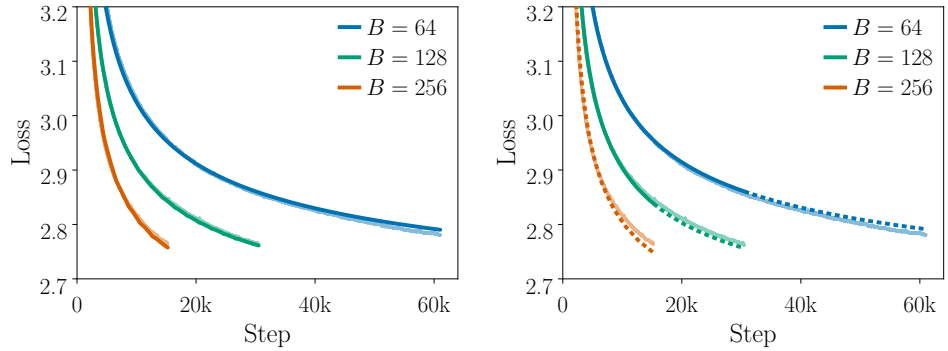


Figure 13: **Fitting and extrapolation under a larger learning rate.** We repeat the Adam experiments in Figure 3 (b) and Figure 12 (right) with the learning rate increased from 5×10^{-4} to 10^{-3}. Solid portions denote FSL fitting results, and dashed portions represent FSL predictions on held-out horizons. **(Left)** Joint fitting across batch sizes. **(Right)** Fitting and extrapolation to held-out batch sizes and training horizons. FSL remains accurate for joint fitting and the held-out batch size, while the longer-horizon continuation predictions show modest deviations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The paper provides the full set of assumptions and a complete (and correct) proof for each theoretical result.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not).

Guidelines:

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

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

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

   Answer: [No]

   Justification: While we provide detailed descriptions of all experimental setups and hyperparameters in Section 5 and Appendix H.2, H.3, we do not currently release code or datasets.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

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

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We use multiple experiments to get mean value to improve statistical significance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: While we describe the experimental setup in full detail, we do not currently report the specific compute resources used.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper does conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our paper is a theoretical work and there is no societal impact.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The code packages and open-source models used in this paper are all properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We describe these in the OpenReview submission. LLMs were used to assist with code writing.

    Guidelines:

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