Phases of MUON: When MUON Eclipses SIGNSGD

Elliot Paquette[*]
McGill University

Noah Marshall
McGill University

Lucas Benigni
Université de Montréal

Guangyuan Wang
McGill University

Atish Agarwala
Google DeepMind

Courtney Paquette
McGill University

Abstract

MUON and related spectral optimizers have recently demonstrated strong empirical performance as scalable stochastic methods, often outperforming ADAM. Yet their behavior remains poorly understood. We analyze stochastic MUON, which orthogonalizes an exponential moving average of minibatch gradients, on a high-dimensional matrix-valued least squares problem, and we compare it with stochastic SIGNSGD, a simple baseline comparable to ADAM. We derive explicit deterministic dynamics that provide a tractable framework for studying learning behavior. We show that MUON has non-linear sensitivity to the effective batch $B_{\text{eff}} = B(1+\beta)/(1-\beta)$. With exact whitening and small B_{eff} MUON fails to resolve smaller eigenmodes, slowing down convergence. Momentum and finite Newton–Shulz iterations can act as a ridge which lessen this effect. This behavior is in contrast with SIGNSGD which has no transition, leading to different optimal learning rates, but with matching convergence rate up to a constant factor in the isotropic setting. For anisotropic data, large batch setting MUON performs a square-root preconditioning with respect to the data covariance spectrum while SIGNSGD provides no preconditioning for generic covariances. We analyzed time to target loss in a power law covariance model and derive a phase diagram in the data exponent α and target exponent ς. MUON is favored for steep covariance spectra with rich targets ($\alpha > 2, \varsigma < \alpha + 1$), SIGNSGD is favored for flat spectra ($\alpha < 1$), and in other regions the winner depends on how the target loss scales with dimension. Our results shed light on why replacing ADAM with MUON is not trivial even from a pure optimization perspective.

1 Introduction

Practical training in machine learning is dominated by stochastic first-order methods with diagonal preconditioning, in particular, robust, scalable variants of ADAM. A long-standing research goal is to develop efficient and robust *non-diagonal* preconditioners which exploit the matrix structure of neural network parameters [30, 46, 64, 72]. Recently the MUON optimizer [34] has gained traction as a simple yet scalable optimizer in this class which can outperform ADAM in many settings [34, 42, 62, 76]. MUON efficiently approximates spectral whitening of matrix-valued gradients via Newton–Schulz iterations. However, deploying MUON can be difficult; the relationship between the training dynamics and batch size, dimension, and data structure remain poorly understood. This leads to a natural question: *under what conditions does* MUON *outperform* ADAM?

To answer this question, we perform an analysis of the learning dynamics of the family of *stochastic spectral optimizers*, particularly MUON with momentum. We focus on a motivated high-dimensional setting where we can derive deterministic equations for the stochastic loss trajectories. This allows for

[*]Email: elliot.paquette@mcgill.ca and Website: https://elliotpaquette.github.io

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

explicit computation of the scaling of optimal hyperparameters with batch size and dimension. This also reveals a rich phase diagram that quantifies the performance of non-diagonal preconditioners and characterizes the joint conditions on the data and optimization setup under which these methods outperform Adam-like methods. Taken together, our results suggest new directions for future empirical and theoretical study towards improving spectral optimizers.

Related Work. The original motivation for MUON [34] stems from steepest descent with respect to the spectral norm [7]. The method is closely related to stochastic spectral descent [13] and orthogonalized gradient methods [69], and can also be interpreted as a form of basis-aware SIGNSGD [8]. Its strong empirical performance [34, 42, 62, 76] has led to a number of recent variants [2, 23, 43, 44, 56, 61]. On the theoretical side, several works analyze MUON and related methods using worst-case complexity arguments, establishing $\mathcal{O}(1/\sqrt{T})$ convergence rates comparable to SGD [15, 35, 63, 65]. These analyses connect MUON to frameworks such as inexact linear minimization oracles [65] and broader algorithmic families like Lion-$\mathcal{K}$ [15], while also characterizing the role of approximate polarization [35]. Complementary regret guarantees are provided in [33].

A growing body of work studies MUON and the exact whitening variant SIGNSVD in simplified settings. Gradient-flow analyses yield explicit loss dynamics for SIGNSVD and enable comparisons with gradient descent [57, 71], while other works highlight nontrivial behavior even on quadratic objectives [29]. High-dimensional analyses remain limited: [73] derives risk dynamics for SIGNSVD in isotropic least-squares settings but does not explore their implications, while [25] studies the dynamics of MUON and SIGNSVD in full-batch isotropic least-squares settings. Beyond least squares, implicit bias has been studied in linear classification [26, 68], and recent work connects MUON to attention via associative memory models, including results on critical batch size scaling [37, 39, 74]. Several papers also attempt to explain the empirical success of spectral methods through single-step analyses [21, 67], though these do not reliably predict full training behaviour [29]. We instead provide explicit stochastic training dynamics over the full course of training. This extends the deterministic-equivalent / homogenization line of analysis developed for SGD and SIGNSGD on scalar-valued least squares [53, 54, 77] to a matrix-valued setting in which spectral preconditioning is meaningful. See Sec. A for an extended discussion of background works.

Main contributions. We introduce a matrix-valued quadratic optimization problem that captures key aspects of neural-network training, in particular the outer-product structure of fully connected layer gradients (Sec. 1.1). This setting enables the study of a general family of *stochastic spectral optimizers*. The algorithm we analyze is stochastic MUON with exponential-moving-average momentum; its case without momentum is stochastic SIGNSVD, the exact spectral whitening of the gradient, which goes back to stochastic spectral descent [13]. The framework also covers diagonal methods such as stochastic SIGNSGD [8], a simple baseline comparable to ADAM (Sec. 1.2).

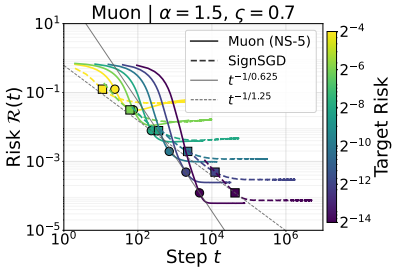


Figure 1: **Random-matrix theory predicts MUON and SIGNSGD loss curves.** Power-law data ($\alpha = 1.5$, $\varsigma = 0.7$, $N = 1024$, $B = 2048$) in the mixed phase C of Fig. 7, where MUON wins at this dimension once ϵ is small. Solid: mean of 16 runs of MUON with 5 Newton–Schulz steps; gray: predicted exponents (Sec. 5).

We derive deterministic learning dynamics for these algorithms via a random matrix theory analysis. Specifically, we show that the stochastic risk is tracked by a deterministic function $\mathcal{R}(t)$, characterized through drift and volatility kernels; this is a theorem on a finite horizon in the proportional limit (Sec. 2, Thm. 2.1). This framework enables us to predict risk trajectories without running the stochastic algorithm and to explicitly relate performance to batch size B, matrix dimension $(N \times N)$, and optimizer hyperparameters (e.g., non-constant learning rate η), thereby providing a precise characterization of scaling behavior (see Fig. 3). ***The dynamics cover a broad class of spectral optimizers as well as the fully anisotropic setting.*** We further leverage this framework to uncover a rich set of phenomena:

- **Critical batch effect for MUON (Sec. 3 & Sec. 4).** A minibatch gradient resolves only part of the parameter eigenspace, and MUON acts as a *square-root preconditioner* only on the resolved eigenspaces. Momentum enlarges the effective batch $B_{\text{eff}} = B(1+\beta)/(1-\beta)$, and as B_{eff} increases, more eigenspaces are resolved, leading to a progressive improvement in convergence. A *critical transition* occurs at $B_{\text{eff}} \asymp N$, where all eigenspaces are resolved and MUON preconditions the full spectrum.

Fixed-point equation	Random-matrix meaning
$\sigma = \dfrac{1}{B}\sum_i \dfrac{\mu_i}{\tilde{s}_1\,\mu_i - z}$	output-side resolvent trace $\frac{1}{B}\operatorname{Tr}\big(R(z)\Sigma_{\text{out}}\big)$
$\tilde{\sigma} = \dfrac{1}{B}\sum_j \dfrac{\lambda_j}{s_1\,\lambda_j - z}$	input-side resolvent trace $\frac{1}{B}\operatorname{Tr}\big(\tilde{R}(z)\Sigma_{\text{in}}\big)$
$\tilde{s}_1\,\sigma = s_1\,\tilde{\sigma} = q\,F_\beta(q),\ q = -z\,\sigma\,\tilde{\sigma}$	Gaussian closure over the moving-average weights
$F_\beta(q) = \sum_{j\geq 0} w_j^2 J(w_j^2 q), \quad L_\beta(q) = \sum_{j\geq 0} w_j J(w_j^2 q)$	
$J(x) = \mathbb{E}_{V\sim\mathcal{N}(0,1)}\big[V^2/(1+xV^2)\big], \quad w_j = (1-\beta)\beta^j$	

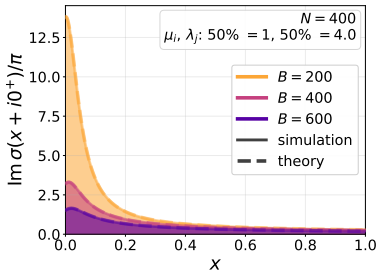


Figure 2: Self-consistent fixed-point system for the deterministic equivalents of the resolvents $R(z) = (M_t M_t^\top/(2\mathcal{R}) - z\,\mathrm{Id}_{N_{\text{out}}})^{-1}$ and $\tilde{R}(z) = (M_t^\top M_t/(2\mathcal{R}) - z\,\mathrm{Id}_{N_{\text{in}}})^{-1}$ of the moving-average buffer, for $z < 0$ (definition I.1 and theorem I.1); at $\beta = 0$, $F_0 = L_0 = J$. *Left:* the governing equations. *Right:* simulation validation at $\beta = 0$ of $\operatorname{Im}\,\sigma(x + i0^+)/\pi$ against the fixed-point solution for a doubly anisotropic spectrum. 50 random data matrices drawn at each batch size B, $N = 400$, imaginary part of z is 10^{-3}.

- **With isotropic data, MUON has a larger learning rate and the same convergence rate up to a constant (Sec. 3, Figs. 5, 6).** The optimal learning rate of MUON exceeds SIGNSGD's by a factor of order $\sqrt{N}$ for $B_{\text{eff}} \gg N$, recovering and refining the empirical $\sqrt{N}$ MOONSHOT rule of [42], and of order $\sqrt{B}$ for $B_{\text{eff}} \ll N$ ($N/\sqrt{B}$ at $\beta = 0$). For $B_{\text{eff}} \gg N$, momentum is a rescaling of the learning rate and for $B_{\text{eff}} \ll N$, it sets the bias of MUON to $1/(2\sqrt{\pi})$. However at the optimal learning rate there is no speedup from MUON on isotropic data.
- **Newton–Shulz is a useful regularizer.** We quantitatively connect finite Newton–Schulz steps to a ridge-regularized whitening, which can be advantageous at small batch: with momentum and $B_{\text{eff}} < N$, a few Newton–Schulz steps outperform exact whitening (Fig. 6, App. C).
- **Anisotropic power-law data reveals a phase diagram (Sec. 4 & Sec. 5).** A power-law covariance with data exponent α and target exponent ς gives a solvable setting in which we can tell which problems favor MUON and which favor SIGNSGD: comparing the time each method needs to reach a target loss divides the (α, ς)-plane into MUON-favored and SIGNSGD-favored phases (Figs. 1, 7). Spectral preconditioning is not uniformly advantageous, even at large batch sizes.

1.1 Model setup

We study matrix-valued linear regression with $N_{\text{out}} \times N_{\text{in}}$ parameters W, a fixed teacher $W^\star$ of the same shape, single-sample loss $\frac{1}{2}\langle x_{\text{out}} \otimes x_{\text{in}},\, W^\star - W\rangle^2$, population risk

$$\mathscr{R}(W) \overset{\text{def}}{=} \tfrac{1}{2}\,\mathbb{E}\big[\langle x_{\text{out}} \otimes x_{\text{in}},\, W - W^\star\rangle^2\big]\,, \tag{1}$$

and no label noise; we write $\mathscr{R}(t) \overset{\text{def}}{=} \mathscr{R}(W_t)$. This is a minimal toy model that retains the gradient structure spectral preconditioners are designed to exploit. The single sample gradient $(x_{\text{out}} \otimes x_{\text{in}})\,\langle x_{\text{out}} \otimes x_{\text{in}},\, W - W^\star\rangle$ has exactly the rank-one outer-product structure of a single linear layer's gradient under backprop, with x_{out} playing the role of the upstream error and x_{in} the activation (see Sec. SA for longer discussion). The data $x_{\text{in}}, x_{\text{out}}$ are independent Gaussian with covariances $\Sigma_{\text{in}}, \Sigma_{\text{out}}$; as we will see, different choices in the data covariance will dramatically change the dynamics. The proportional-regime assumptions (and aspect ratios) controlling $N_{\text{in}}, N_{\text{out}}, B$ and $\|\Sigma_{\text{in}}\|_{\text{op}}, \|\Sigma_{\text{out}}\|_{\text{op}}$ are formally defined in Assumption 2 and Assumption 3.

1.2 Stochastic spectral algorithms.

The minibatch gradient G_{t+1} of (1) is almost surely a rank-$\min\{N_{\text{in}}, N_{\text{out}}, B\}$ matrix where B is the size of the minibatch. *Stochastic spectral optimizers* process the singular values of a moving average of minibatch gradients through a function $\varphi : [0, \infty) \to \mathbb{R}$:

Algorithm MUON$_\varphi$. Fix a momentum $\beta \in [0,1)$, a learning rate $\eta > 0$ and $M_{-1} = 0$. With $\Delta_t \overset{\text{def}}{=} W_t - W^\star$, for $t \geq 0$,

$$G_{t+1} \overset{\text{def}}{=} \frac{1}{B}\sum_{i=1}^{B}(x^i_{\text{out}} \otimes x^i_{\text{in}})\,\langle x^i_{\text{out}} \otimes x^i_{\text{in}},\, \Delta_t\rangle, \tag{2}$$

$$M_t \;=\; \beta M_{t-1} + (1-\beta)\,G_{t+1}, \qquad W_{t+1} \;=\; W_t - \eta\,\varphi(M_t M_t^\top)\,M_t. \tag{3}$$

MUON is $\varphi(z) = z^{-1/2}$, the polar factor $\operatorname{polar}(M) = (MM^\top)^{-1/2}M = UV^\top$ of an SVD $M = U\Sigma V^\top$:

$$W_{t+1} \;=\; W_t - \eta \operatorname{polar}(M_t). \tag{4}$$

SIGNSVD is MUON at $\beta = 0$. The **regularized polar map** is $\varphi(z) = (z + 2\mathscr{R}(t)\varrho)^{-1/2}$, with a normalized ridge $\varrho > 0$.

SIGNSVD, the exact spectral whitening of the gradient, goes back to stochastic spectral descent [13]. In practice the polar factor is computed by K Newton–Schulz iterations on the Frobenius-normalized buffer [2, 34]. The regularized polar map is the mathematical simplification for which the dynamical theorem is proved (Thm. 2.1; Def. F.1), and a qualitative model of finite Newton–Schulz: K iterations map a small singular value σ to about $a^K\sigma$, with $a \approx 3.44$ for the standard coefficients, an effective ridge of order a^{-2K} (about $4 \cdot 10^{-6}$ at $K = 5$); Thm. K.5 and K.6 describe what a small ridge does. Sec. D compares the momentum theory with simulations, and we validate predictions on MUON (NS-5) numerically (e.g. Fig. 1). Note that recent work has introduced MUON with *hybrid Newton-Schulz* which results in the near exact SIGNSVD whitening of singular values [23].[2]

SIGNSGD. We compare with stochastic SIGNSGD [8], a simple baseline comparable to ADAM, which acts *entrywise* (not spectrally) on G_{t+1}:

$$W_{t+1} = W_t - \eta\,\widehat{G}_{t+1}, \qquad (\widehat{G}_{t+1})_{ij} \overset{\text{def}}{=} \operatorname{sign}\big((G_{t+1})_{ij}\big). \tag{5}$$

SIGNSGD coincides with ADAM at $\beta_1 = \beta_2 = 0$, so comparing it against SIGNSVD captures, to leading order, the structural difference between ADAM and MUON. We write *s-SGD* for SIGNSGD, *s-SVD* for SIGNSVD, and SPEC for the general spectral family throughout.

2 Learning curves and deterministic dynamics

We derive a deterministic function $\mathcal{R}(t)$ that tracks the stochastic risk $\mathscr{R}(W_t)$ in the proportional limit (Thm. H.1). Working in a projected basis, we extend prior deterministic-equivalent / homogenization analyses for SGD and SIGNSGD [53, 77] to SPEC. MUON and SIGNSGD both admit closed deterministic risk recursions of the same form. They differ only through their drift $\mathfrak{d}_{ij}$ and volatility $\mathfrak{v}_{ij}$ kernels.

Let (μ_i, u_i) and (λ_j, v_j) be the eigenpairs of Σ_{out} and Σ_{in}. Define the *projected risk* $\mathscr{Q}_{ij}(t) \overset{\text{def}}{=} (u_i^\top \Delta_t v_j)^2$, so that $\mathscr{R}(W_t) = \frac{1}{2}\sum_{ij}\mu_i\lambda_j\,\mathscr{Q}_{ij}(t)$. In the proportional limit, the projected risks $\mathscr{Q}_{ij}$ follow, in aggregate, a deterministic trajectory $\mathcal{Q}_{ij}$ (Thm. 2.1 below) governed by the universal recursion

$$\mathcal{Q}_{ij}(t{+}1) \;=\; \mathcal{Q}_{ij}(t) \;-\; \frac{2\eta\,\mathfrak{d}_{ij}(t)}{\sqrt{\mathcal{R}(t)}}\,\mathcal{Q}_{ij}(t) \;+\; \eta^2\,\mathfrak{v}_{ij}(t), \tag{6}$$

with $\mathcal{R}(t) \overset{\text{def}}{=} \frac{1}{2}\sum_{ij}\mu_i\lambda_j\,\mathcal{Q}_{ij}(t)$ the theory prediction for $\mathscr{R}$.

Theorem 2.1 (Deterministic dynamics, informal)**.** *Let $N_{\text{in}}, N_{\text{out}}, B \to \infty$ proportionally, and run the algorithm of Sec. 1.2 with a fixed momentum $\beta \in [0,1)$, a fixed normalized ridge $\varrho > 0$ and step $\eta = h/B$. Then $\sup_{t \leq BT_\star}|\mathscr{R}(W_t) - \mathcal{R}(t)| \to 0$ in probability, where $T_\star > 0$ depends only on h, ϱ, β, the aspect ratios, the covariance bound and the initial risk.*

[2] Early MUON implementations aggressively normalize the gradient, prior to using a Newton-Schulz scheme that prioritizes rapid increase of singular values from 0, at the cost of slower convergence to 1.

See Thm. H.1 for the precise form (hypotheses on the initial state, the spectral-measure topology and the error terms). The rest of the paper analyzes this deterministic equivalent at zero ridge ($\varrho \downarrow 0$) as the proxy for MUON. The system is closed in the high-dimensional limit. The drift kernel $\mathfrak{d}_{ij}$ and the volatility kernel $\mathfrak{v}_{ij}$ depend only on (μ_i, λ_j), the aspect ratios, β and the algorithm. They need no further information from Δ_t. The kernels are computed in Sec. 3 and Sec. 4. For complete details, see Sec. SA to SE and Sec. SJ.

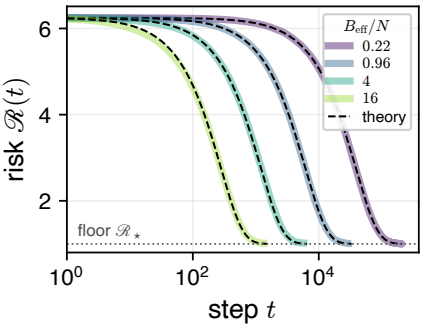


Figure 3: **Theory tracks MUON.** Isotropic data, $N = 256$, $\beta = 0.9$, exact polar factor, constant η, floor $\mathcal{R}_\star = 1$. Thick: simulations (App. B); dashed: the recursion (8). Relative error below 2%.

Computing the kernels: a self-consistent fixed-point system. For the spectral family, $\mathfrak{d}_{ij}$ and $\mathfrak{v}_{ij}$ are spectral integrals against the deterministic equivalent of the resolvent of the Gram matrix of the moving-average buffer. After rescaling spectral variables by the risk ($z \to 2\mathcal{R} \cdot z$), the deterministic equivalent solves the self-consistent fixed-point system in Fig. 2. The system needs only the eigenvalues $\{\mu_i\}, \{\lambda_j\}$, the aspect ratios and the momentum. It belongs to the family of deterministic-equivalent / homogenization fixed-point systems used in high-dimensional learning theory and random-matrix theory [1, 20, 31, 53, 55, 66, 75].

The drift kernel is

$$\mathfrak{d}_{ij} \;=\; \frac{\sqrt{2}}{\pi} \int_0^\infty \frac{z\, L_\beta\big(q(-z)\big)\, \mu_i\, \lambda_j}{\big(z + \tilde{s}_1(-z)\, \mu_i\big)\big(z + s_1(-z)\, \lambda_j\big)}\, \mathrm{d}u, \qquad z = u^2, \tag{7}$$

where $\tilde{s}_1, s_1, q$ solve the system of Fig. 2 at $-z$. At a positive ridge, $z = \varrho + u^2$ (corollary I.1). The volatility kernel $\mathfrak{v}_{ij}$ is the noise the update injects into mode (i, j) per step. It includes the correlation between successive updates that momentum creates. App. SD.1 gives it in closed form. Neither kernel depends on $\mathcal{R}$, because the spectral variables are normalized by the risk. For simplicity, the main text takes $N_{\text{in}} = N_{\text{out}} = N$. The appendices treat the general case.

3 MUON and SIGNSGD in the isotropic setting

The simplest comparison of MUON and SIGNSGD uses isotropic Gaussian data, $x_{\text{in}}, x_{\text{out}} \sim \mathcal{N}(0, \mathrm{Id}_N)$. With identity covariances there is no spectral structure for MUON to exploit. The comparison comes down to how each method handles minibatch noise relative to the dimension N. For MUON, the batch size and the momentum set that noise jointly, through the ***effective batch*** $B_{\text{eff}} \overset{\text{def}}{=} B(1+\beta)/(1-\beta)$. The behaviour of MUON changes qualitatively at $B_{\text{eff}} \asymp N$.

Deterministic risk recursion. With identity covariances, the dynamics of the projected risk $\mathcal{Q}_{ij}(t) = (u_i^\top \Delta_t v_j)^2$ are *eigenmode-independent*. The kernels are constant, $\mathfrak{d}_{ij} = \mathfrak{d}$ and $\mathfrak{v}_{ij} = \mathfrak{v}/N^2$. Summing over (i, j) gives a deterministic *scalar risk recursion* for $\mathcal{R} = \frac{1}{2}\sum_{ij} \mathcal{Q}_{ij}$, of the same form for both algorithms (Muon: Prop. K.2 & Cor. K.1; SIGNSGD: Prop. N.2):

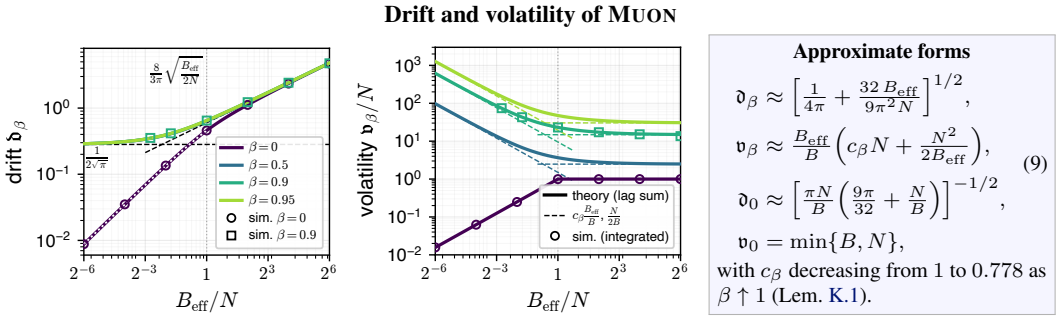
$$\mathcal{R}(t{+}1) = \mathcal{R}(t) - \underbrace{2\eta\, \mathfrak{d}\, \sqrt{\mathcal{R}(t)}}_{\text{drift}} + \underbrace{\tfrac{\eta^2}{2}\, \mathfrak{v}}_{\text{volatility}}. \tag{8}$$

For MUON we write the kernels as a drift $\mathfrak{d} = \mathfrak{d}_\beta$ and a volatility $\mathfrak{v} = \mathfrak{v}_\beta$. They come from the fixed-point system of Fig. 2 at $\mu_i = \lambda_j = 1$ and depend only on B_{eff}/N and β (Fig. 4; Eq. (SF.21), Def. K.1 and Prop. K.3, where they are written C_β and $\mathfrak{v}_{\text{iso},\beta}$). For SIGNSGD the kernels are explicit, $\mathfrak{d} = \mathcal{N}_B/\sqrt{\pi} \approx \sqrt{B/\pi}$ and $\mathfrak{v} = N^2$ (Prop. N.2, N.1). We use $\mathcal{N}_B \approx \sqrt{B}$ throughout. The remaining analyses depend on the kernels only through two ratios: The drift to volatility ratio $\mathfrak{d}/\mathfrak{v}$ sets the learning rate, and the signal-to-noise ratio $\mathfrak{d}^2/\mathfrak{v}$ sets the speed of training.

3.1 Learning rates: floors and schedules

Constant learning rate: floors. Both updates have a fixed size, so at a constant learning rate the risk does not go to zero, but rather balances at the risk floor (Fig. 3):

$$\mathcal{R}_\star^{\text{Muon}} = \left(\frac{\eta\, \mathfrak{v}_\beta}{4\, \mathfrak{d}_\beta}\right)^2 \quad \text{and} \quad \mathcal{R}_\star^{\text{s-SGD}} = \frac{\pi\, \eta^2 N^4}{16\, B}. \tag{10}$$



(a) Drift (b) Volatility (c)

Figure 4: **Isotropic kernels of MUON against the effective batch** $B_{\rm eff} = B(1+\beta)/(1-\beta)$, square case. Colored curves represent the theory (Eq. (SF.21), Def. K.1), black dashed lines are the limits $B_{\rm eff} \gg N$ and $B_{\rm eff} \ll N$, which differ for $\beta = 0$ and $\beta > 0$ (c). Markers are simulations at $N = 512$ with the exact polar factor. Drift and volatility change character at $B_{\rm eff} \asymp N$, vanishing for $\beta = 0$ for $B \ll N$, since the gradient is only rank B even after SIGNSVD. With momentum the buffer has full rank and the drift tends to $1/(2\sqrt{\pi})$ (Thm. K.4). See Sec. B.

At the same η, the MUON floor is far lower. The factor is of order N for $B_{\rm eff} \gg N$, and of order B for $B_{\rm eff} \ll N$ with momentum (N^2/B at $\beta = 0$). To reach the same floor, the MUON learning rate must be larger by the square root of this factor.

Scheduled learning rate. The learning rate that decreases the risk the most at time t is $\eta_\star(t) = 2\mathfrak{d}\sqrt{\mathcal{R}(t)}/\mathfrak{v}$. This turns out to be a schedule that gives the lowest risk at all times (something which need not even exist for anisotropic data). With this learning rate, the recursion (8) gives $\mathcal{R}(t+1) = (1 - 2\mathfrak{d}^2/\mathfrak{v})\,\mathcal{R}(t)$. Thus we can compare the optimal learning rates of MUON and SIGNSGD at a given risk by looking at the ratio:

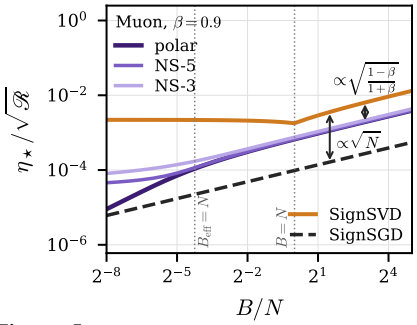


Figure 5: **Optimal learning rates** $\eta_\star/\sqrt{\mathcal{R}} = 2\mathfrak{d}/\mathfrak{v}$, $N = 512$; NS-K: calibrated ridges (App. C). Arrows: the large-batch gaps in (11).

$$\frac{\eta_\star^{\text{Muon}}}{\eta_\star^{\text{s-SGD}}} = \frac{\sqrt{\pi}\,N^2\,\mathfrak{d}_\beta}{\sqrt{B}\,\mathfrak{v}_\beta} \approx \begin{cases} \dfrac{8\sqrt{N}}{3c_\beta\sqrt{2\pi}}\sqrt{\dfrac{1-\beta}{1+\beta}}, & B_{\rm eff} \gg N, \\ \sqrt{B}, & B_{\rm eff} \ll N, \end{cases} \tag{11}$$

At $\beta = 0$ and $B \le N$ the ratio is $N/\sqrt{B}$, since $\mathfrak{v}_0 = \min\{B, N\}$ (Prop. K.1). With momentum, the MUON learning rate bends at $B_{\rm eff} \asymp N$, from $\eta_\star \propto B$ below to $\eta_\star \propto \sqrt{B}$ above (Fig. 5). Finite Newton–Schulz flattens the part below $B_{\rm eff} = N$. With five steps the slope drops from about 1 to about 0.3, close to the picture at $\beta = 0$ (Fig. 5; App. C.1.4). The MUON learning rate is always the larger, as in practice [23, 42]. Momentum narrows the gap at large batch by a factor $\sqrt{(1-\beta)/(1+\beta)}$. The batch limits hold up to a factor $1 + o(1)$, as $N/B \to 0$ for $B_{\rm eff} \gg N$ (Thm. K.2, K.1, K.3, Lem. K.1) and as $N/B \to \infty$ for $B_{\rm eff} \ll N$ (Thm. K.4).

Key Takeaway: The MUON learning rate should exceed that of SIGNSGD by a factor of order $\sqrt{N}$ for $B_{\rm eff} \gg N$ and $\sqrt{B}$ for $B_{\rm eff} \ll N$, whether the learning rate is constant (matched floors) or scheduled. The $\sqrt{N}$ factor is the MOONSHOT scaling rule [42] (see also [23]).

3.2 Rate comparison: the MUON speedup

Under the scheduled learning rates both algorithms contract by $\approx e^{-2\mathfrak{d}^2/\mathfrak{v}}$ per step. We call the ratio of the two exponents the MUON *speedup*:

$$\mathcal{S}_{\text{Muon}} \overset{\text{def}}{=} \frac{2\mathfrak{d}_\beta^2/\mathfrak{v}_\beta}{2B/(\pi N^2)} \approx \begin{cases} \frac{32}{9\pi c_\beta}, & B_{\rm eff} \gg N, \\ \frac{1}{2}, & B_{\rm eff} \ll N,\ \beta > 0, \\ 1, & B_{\rm eff} \ll N,\ \beta = 0. \end{cases} \tag{12}$$

It is the number of SIGNSGD steps that one MUON step is worth. It is also the ratio of the times to a common floor at constant learning rates (Thm. M.2, M.1).

Whitening does not help on isotropic data. The speedup stays between $\frac{1}{2}$ and 1.46 for every β, B and N (Fig. 6). Once each learning rate is tuned, the two algorithms are interchangeable up to an absolute constant. Whitening the minibatch is in fact mildly harmful. Figure 6 shows where. *Large batch.* For $B_{\text{eff}} \gg N$ the minibatch singular values already approach 1, the population limit, and so the speedup approaches $\pi/2$ as B_{eff}/N continues to grow (Rmk. K.1). One Newton–Schulz step, the softest whitening and essentially normalized SGD, attains $\pi/2$ at every batch size (App. C). *Small batch without momentum.* For $B_{\text{eff}} \ll N$ the speedup tends to 1, as at small batch, the whitening of the individual per-example gradients can essentially be done per example. *Near* $B_{\text{eff}} = N$. Without momentum the speedup dips to about 0.7 at $B_{\text{eff}} = N$. There the minibatch gradient is square and its singular values reach down to zero, as in double descent. Whitening amplifies these weakest directions the most. *Small batch with momentum.* For $B_{\text{eff}} \ll N$ the buffer has full rank but many weak directions. Exact whitening amplifies all of them and halves the rate.

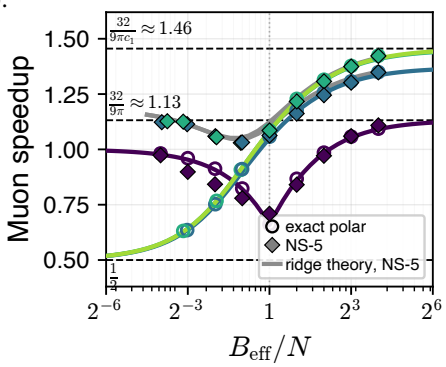


Figure 6: **MUON speedup** (12); above 1, MUON is faster. Colors as in Fig. 4. Curves: zero ridge; markers: simulations at $N = 512$, exact polar (open) and NS-5 (filled); grey: ridge theory at the calibrated NS-5 ridge (App. C).

Newton–Schulz is ridge-regularized whitening. Practical MUON does not exactly whiten. K Newton–Schulz steps map small singular values σ to about $a^K\sigma$. This acts like the regularized polar map with ridge $\varrho \approx 0.31\, a^{-2K}\|M_t\|_{\text{F}}^2$ (App. C). At this calibrated ridge the theory predicts the NS-5 speedups without refitting (Fig. 6, grey; Thm. K.6, $a = 3.4445$). On isotropic data, whitening less is better. NS-5 beats exact whitening for $B_{\text{eff}} < N$, and the speedup decreases with K at every β and B_{eff} (Fig. 14).

Key Takeaway: On isotropic data, tuned MUON and SIGNSGD converge at the same rate up to a factor of at most 2. There is no structure to whiten, and the less MUON whitens small sinbular values (fewer Newton–Schulz steps, larger effective ridge), the faster it converges.

4 Anisotropic output covariance ($\Sigma_{\text{in}} = \text{Id}_N$, $\Sigma_{\text{out}} =$ general)

The previous analysis suggests, not surprisingly, that any real advantage of MUON stems from anisotropy. To explore this further we consider the *half-anisotropic* setting: $\Sigma_{\text{in}} = \text{Id}_N$ and Σ_{out} is a general covariance matrix. This setting captures the effect of non-uniform signal strengths across output modes but has vastly simpler random matrix theory. In particular we focus on the setting:

$$\Sigma_{\text{out}} \;=\; U\ \text{diag}(\mu_1, \ldots, \mu_N)\, U^\top\,, \qquad U \sim \text{Haar}\big(N\big)\,, \tag{13}$$

with U Haar distributed and independent of the data. We note that the distribution of U matters for SIGNSGD but *not* for MUON. Proofs are in the appendix (Cor.I.2, H.1; Sec. SD.9, SG, SJ.4). We begin with the deterministic expression for the risk. Recall the projected risk Ω_{ij} from Sec. 2. In this setting we can partially sum the Ω_{ij} to produce:

$$\mathcal{R}(t) = \tfrac{1}{2}\sum_i^N \mu_i\, \Omega_i, \quad \text{where the } \textit{projected row risks} \text{ satisfy} \quad \Omega_i(t) \overset{\text{def}}{=} \sum_{j=1}^N \Omega_{ij}(t) = \|\Delta_t^\top u_i\|^2.$$

The Ω_i play the role of the Ω_{ij} in Sec. 2 and satisfy a deterministic recursion, which the algorithm follows in the sense of corollary H.1,

$$\Omega_i(t{+}1) = \Omega_i(t) - \frac{2\eta\, \mathfrak{d}_i}{\sqrt{\mathcal{R}(t)}}\Omega_i(t) + \eta^2\, \mathfrak{v}_i. \tag{14}$$

Here the drift kernel $\mathfrak{d}_i$ is the common value of the $\mathfrak{d}_{ij}$, $j = 1, \ldots, N$, which do not depend on j when $\Sigma_{\text{in}} = \text{Id}_N$ (corollary I.2), and the volatility kernel $\mathfrak{v}_i = \sum_{j=1}^N \mathfrak{v}_{ij}$ is their sum (proposition J.1);

both are deterministic functions of the spectrum $\{\mu_i\}_{i=1}^N$. *Both* MUON and SIGNSGD admit risk formulations and projected row risks of the form (14); once more the *only differences* are the specific drift and volatility kernels. SIGNSGD admits the same row recursion on a Haar eigenbasis.[3]

4.1 Large batch: square-root preconditioning

For sufficiently large effective batch $B_{\text{eff}} \gg N$, MUON gives simple drift and volatility coefficients:

Constant volatility. The volatility is the main simplification of the half-anisotropic setting. Without momentum, the update $UV^\top$ is orthogonal, because $B \geq N$. So $\|(UV^\top)^\top u_i\| = 1$, and $\mathfrak{v}_i = 1$ for every mode and every spectrum (Rmk. J.1). With momentum, successive buffers are correlated, and the volatility grows. We use $\mathfrak{v}_i = (1+\beta)/(1-\beta)$, the value proved on the resolved modes at small batch (Sec. 4.2). In the isotropic case the exact value lies between 1 and B_{eff}/B (Lem. K.1). A constant factor in the volatility only rescales the time to reach a target (Rmk. M.2).

Square-root drift. The drift is a square-root law for every spectrum, up to a constant. The constant depends only on how the spectrum is packed around each mode.

Definition 4.1 (Packed mode)**.** *Let $K \geq 1$ and $\delta \in (0,1]$. Mode i is (K,δ)-*packed *if at least δN eigenvalues satisfy $\mu_k \leq K\mu_i$.*

Modes high in the spectrum are packed with δ close to 1. The count of eigenvalues below $K\mu_i$ grows with μ_i, so if the smallest eigenvalue is packed, so is every mode. Packing holds when the spectrum does not decay too fast. Every mode of a power law $\mu_k = k^{-\alpha}$ is $(2, 1-2^{-1/\alpha})$-packed. A geometric spectrum $\mu_k = q^k$ with $q < 1$ decays so fast that its smallest eigenvalues are not packed.

Theorem 4.1 (Large-batch drift; Cor. L.3)**.** *Let $\beta \in [0,1)$ and $\epsilon_{\text{os}} \overset{\text{def}}{=} 3(1+\beta)^2 N/B_{\text{eff}} < 1$. The zero-ridge drift kernel satisfies*

$$(1-\epsilon_{\text{os}})\, c_{K,\delta} \sqrt{\tfrac{B_{\text{eff}}}{2N}}\, \sqrt{\mu_i} \;\leq\; \mathfrak{d}_i \;\leq\; \tfrac{1}{1-\epsilon_{\text{os}}} \sqrt{\tfrac{B_{\text{eff}}}{2N}}\, \sqrt{\mu_i}, \qquad c_{K,\delta} = 0.94\sqrt{\delta/(1+K)}.$$

The upper bound holds for every spectrum and every mode i. The lower bound holds if mode i is (K,δ)-packed.

In expectation and with small learning rate, the MUON update on packed modes is therefore, up to constants (Cor. L.2),

$$W_{t+1} - W_t \approx -\eta \sqrt{\tfrac{B_{\text{eff}}}{2N}}\; \cdot \sqrt{\mathcal{K}} \cdot \tfrac{W_t - W^*}{\sqrt{\mathcal{R}(t)}},$$

where $\mathcal{K}$ is the covariance operator of the problem. We note that gradient descent would instead perform the update $W_{t+1} - W_t = -\eta\mathcal{K}(W_t - W^*)$, so MUON operates like a *square-root preconditioner*. For modes high in the spectrum, the constant tends to 1 (Prop. L.4). When every mode is packed, this preconditioning has an effect comparable to optimal momentum methods, which *accelerate* strongly convex optimization problems by a factor of the square root of the condition number [50, 59]. Not every mode benefits, however. A mode that is not packed, such as an isolated small eigenvalue, can be as slow as under gradient descent, up to a logarithm (Prop. L.5).

If $\mathcal{K}$ is diagonal, SIGNSGD performs the same update as large-batch SIGNSVD, after rescaling the learning rate by $\sqrt{N}$ [77, Eqn. (29)]. In this case SIGNSVD is similar to SIGNSGD, and MUON may be similar to ADAM. But MUON achieves this preconditioning in *any* basis. In the Haar basis above, SIGNSGD loses all of it.

Key Takeaway: For large effective batch, MUON is a square-root preconditioner on every packed eigendirection, in any basis.

4.2 Power-law specialization $\mu_i = i^{-\alpha}$

We now take the *power law* $\mu_i = i^{-\alpha}$ with $\alpha > 0$. It captures key phenomenology of power-law scaling on real data [9, 41, 54, 60].

[3] We derive the SIGNSGD row kernels in a more restricted setting: one step, a Haar eigenbasis, comparable dimensions and a covariance bounded below (Prop. N.3, Thm. N.1). There the per-row volatility carries an extra nonnegative term. Power-law spectra are not bounded below, so for them we use the simpler kernels of Table 1, which omit this term (Rmk. M.1).

Assumption 1 (Power-law covariance.)**.** *Suppose $\mu_i = i^{-\alpha}$ with $\alpha > 0$, and $\Omega_i(0) = i^{-\varsigma}$ such that $\alpha+\varsigma > 1$. The parameters (α, ς) are related to what is known in the literature as source and capacity conditions [12, 14, 24, 58] and the half-space $\alpha + \varsigma > 1$ is precisely where we have summability at initialization.*

Small batch. When $B_{\text{eff}} \ll N$, the batch cannot resolve every direction of the spectrum. MUON whitens only the top modes, up to a *resolution index* $i_\#$. Modes $i \ll i_\#$ keep the square-root drift, $\mathfrak{d}_i \propto \sqrt{\mu_i}$. Modes $i \gg i_\#$ are not whitened. At $\beta = 0$ their drift is $\mathfrak{d}_i \propto \mu_i$. This is the drift of SGD, and of SIGNSGD in the Haar basis. As $N/B \to \infty$, the resolution index is (Prop. L.2, Cor. L.1(iv))

$$i_\# \asymp \begin{cases} B_{\text{eff}}, & \alpha > 1, \\ B_{\text{eff}}/\ln^3(\frac{N}{B}), & \alpha = 1, \\ \left(B_{\text{eff}}/\ln^2(\frac{N}{B})\right)^{1/\alpha}/N^{1/\alpha-1}, & \alpha < 1. \end{cases} \tag{15}$$

Large batch. The theory of Sec. 4.1 applies directly. Every mode of the power law is $(2, 1-2^{-1/\alpha})$-packed (Rmk. L.2). So every mode keeps the square-root law up to a constant that depends only on α. The drift is also sharper than Theorem 4.1. For the modes $i \ll N$, we have $\mathfrak{d}_i \sim \sqrt{\mu_i B_{\text{eff}}/(2N)}$ as $B_{\text{eff}}/N \to \infty$ (Cor. L.4). The rest of this section treats small batch, $B_{\text{eff}} \ll N$.

Drift and volatility kernels. Table 1 gives the kernels on either side of $i_\#$. The resolved modes $i \ll i_\#$ keep the large-batch kernels of Sec. 4.1.

	Drift $\mathfrak{d}_i$	Volatility $\mathfrak{v}_i$
MUON, $B_{\text{eff}} \ll N$		
$i \ll i_\#$	$\sqrt{\mu_i B_{\text{eff}}/(2N)}$	B_{eff}/B
$i \gg i_\#, \beta > 0$	$N^{-\alpha/2}\,\Psi_\alpha(i/N)/\sqrt{2}$	$\frac{N}{B}\,\Gamma_\alpha(i/N)$
$i \gg i_\#, \beta = 0$	$A_\alpha\,\mu_i\sqrt{\kappa B/(2N)}$	$\kappa\mu_i/(1+\kappa\mu_i)$
SIGNSGD	$\mu_i\sqrt{B/(\pi\bar{\mu})}$	$N\big(1+\frac{2}{\pi}\big(\frac{\mu_i}{\bar{\mu}}-1\big)\big)$

$x \downarrow 0$	$\Psi_\alpha(x)$	$\Gamma_\alpha(x)$
$\alpha < 1$	$\log\frac{1}{x}$	$x^\alpha \log^2\frac{1}{x}$
$\alpha = 1$	$\log^{3/2}\frac{1}{x}$	$x\log^3\frac{1}{x}$
$\alpha > 1$	$x^{-(\alpha-1)/2}$	x

$\bar{\mu} = \frac{1}{N}\sum_k \mu_k$; A_α is a constant. κ solves $\sum_i \frac{\kappa\mu_i}{1+\kappa\mu_i} = B$.

Table 1: **Zero-ridge kernels for power-law covariance**, under Assumption 1 and (13), at fixed $\beta \in [0, 1)$. The MUON rows are proved as $N/B \to \infty$ (Cor. L.1; Thm. L.1, L.2; Rmk. L.1). The profiles $\Psi_\alpha, \Gamma_\alpha$ of $x = i/N$ are positive, continuous and independent of β (Def. L.1). Right: their behavior as $x \downarrow 0$, up to constants (Prop. L.1), and the remaining symbols (Prop. L.3).

Key Takeaway: The effective batch B_{eff} is a spectral resolution scale for MUON. It sets how many top eigenmodes are whitened. This number depends sharply on the tail exponent α.

5 Time to ϵ-approximate solution: performance of MUON vs. SIGNSGD

We compare the two algorithms at matched limit loss on power-law data with $B \geq N$. Given a target ϵ, each algorithm runs at the constant learning rate

$$\eta^\star = \frac{2\sqrt{\epsilon}}{\mathcal{S}}, \qquad \mathcal{S} = \sum_{i=1}^{N} \frac{\mu_i\,\mathfrak{v}_i}{2\,\mathfrak{d}_i}, \tag{16}$$

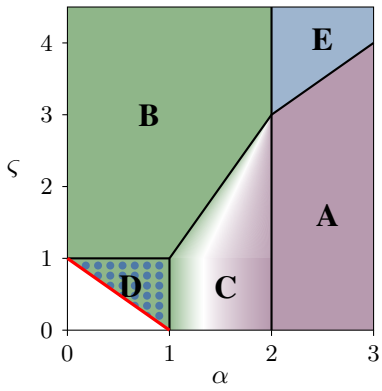
which makes its risk floor $\mathcal{R}(\infty) = \epsilon$, and $t_{4\epsilon} = \inf\{t \geq 0 : \mathcal{R}(t) \leq 4\epsilon\}$ is the first time its risk is within a factor 4 of the floor. A constant learning rate avoids comparing two optimized schedules, which in the anisotropic case differ between the methods. On the other hand, this constant learning rate ensures a fair comparison since the risk floors are the same for both algorithms.

Theorem 5.1 (Speed ratio, informal, see Thm. M.3, Cor. M.2, Thm. M.4 for precise form.)**.** *Under Assumption 1, take $B \geq N$, the large-batch* MUON *kernels of Sec. 4.1, the* SIGNSGD *kernels of Table 1, $\eta = \eta^\star$, and $\alpha \notin \{\frac{1}{2}, 1\}$. For large N and small ϵ, the ratio of times-to-4ϵ $t_{4\epsilon}$ is*

$$\frac{t_{4\epsilon}^{\textit{s-SGD}}}{t_{4\epsilon}^{\textit{Muon}}} \asymp D_N \max\{\epsilon, \epsilon_N\}^{-\kappa}, \qquad D_N \overset{\text{def}}{=} \frac{\pi\sum_{i\leq N} i^{-\alpha}}{\sum_{i\leq N} i^{-\alpha/2}}, \qquad \epsilon_N \overset{\text{def}}{=} N^{-(\alpha+\varsigma-1)}. \tag{17}$$

Here $\kappa \overset{\text{def}}{=} e_{\textit{s-SGD}} - e_{\textit{Muon}} \geq 0$, where

$$e_{\textit{Muon}} = \max\left\{\frac{1}{2}, \frac{\alpha}{2(\alpha+\varsigma-1)}\right\}, \qquad e_{\textit{s-SGD}} = \max\left\{\frac{1}{2}, \frac{\alpha}{\alpha+\varsigma-1}\right\}. \tag{18}$$



Phase	Ratio $t_{4\epsilon}^{\text{s-SGD}}/t_{4\epsilon}^{\text{Muon}}$ at $\epsilon = N^{-\omega}$	
A: $\alpha > 2, \varsigma < \alpha{+}1$	MUON favored (Prop. M.1).	
B: $\alpha < 1 < \varsigma$; $1{<}\alpha{<}2, \varsigma{>}2\alpha{-}1$	SIGNSGD favored (Thm. M.5).	
C: $1{<}\alpha{<}2, \varsigma{<}2\alpha{-}1$	if $\omega < \omega_c$, SIGNSGD favored; if $\omega > \omega_c$, MUON favored (Cor. M.2).	
D: $\alpha < 1, \varsigma < 1$	if $\omega < \alpha{+}\varsigma{-}1$, SIGNSGD favored; if $\omega > \alpha{+}\varsigma{-}1$, algs. perform equally.	(Thm. M.4)
E: $\alpha > 2, \varsigma > \alpha{+}1$	Ratio tends to a constant; both algs. perform equally (Cor. M.1).	

Figure 7: **Phase diagram for MUON vs. SIGNSGD in the (α, ς) plane** when $B \geq N$, matched limit loss, and target $\epsilon = N^{-\omega}$. The analysis requires $\alpha + \varsigma > 1$, above the red line. In C, the shading shows the fraction of accuracies (ϵ) at which SIGNSGD still wins, from green (1) through white ($\frac{1}{2}$) to purple (0), MUON favored. Dots in D indicate the SIGNSGD favored and tie performance. Boundary points are given in Cor. M.3.

Phases: which algorithm wins? To determine which algorithm wins, we analyze the time-to-4ϵ ratio (17), which is a competition between an N factor and an ϵ factor. We set the target accuracy to $\epsilon = N^{-\omega}$, so that the final ratio is a power of N. The sign of the resulting exponent of N then gives the phases of Fig. 7 (Cor. M.3). For "hard" targets with poor conditioning and a lot of loss in small eigenmodes, the exponent is positive and MUON is favored (Phase A). In the rest of the phase diagram, the exponent is either negative (favoring SIGNSGD), or zero (tie). The only subtlety is Phase C ($1 < \alpha < 2, \varsigma < 2\alpha - 1$), where the winner depends on the target accuracy ω. Define

$$\omega_c = \frac{(2-\alpha)(\alpha+\varsigma-1)}{\alpha+1-\varsigma} \quad (\varsigma \geq 1), \qquad \omega_c = \frac{(2-\alpha)(\alpha+\varsigma-1)}{\alpha} \quad (\varsigma < 1).$$

For a given N, the ratio (17) grows as ϵ decreases (Cor. M.2), so SIGNSGD wins at low accuracy, $\epsilon \gtrsim N^{-\omega_c}$, and MUON takes over at high accuracy, $\epsilon \lesssim N^{-\omega_c}$. Fig. 8 illustrates the loss behavior across the different phases.

Key Takeaway: At matched limit loss, the winner depends on both the data, through (α, ς), and the target accuracy, so the advantage of MUON is highly data dependent. MUON wins where whitening pays off, on steep spectra and on hard targets at high accuracy. SIGNSGD tends to win on easy targets or at low accuracy, especially for flatter spectra, where whitening does little (Fig. 7 and Fig. 8).

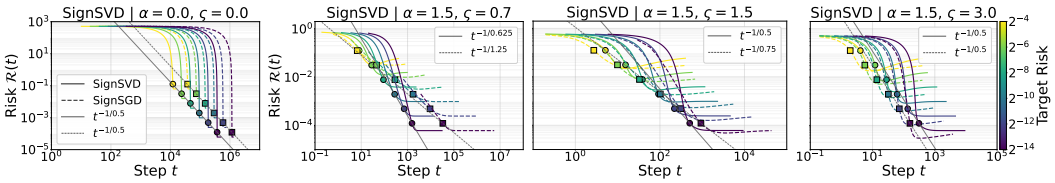


Figure 8: **Deterministic risk trajectories $\mathcal{R}(t)$ and $t_{4\epsilon}$ at $N = 1024$, $B = 2048$**: isotropic data and the power-law points $(\alpha, \varsigma) = (1.5, 0.7), (1.5, 1.5), (1.5, 3.0)$, for $\epsilon \in [2^{-14}, 2^{-4}]$. Circles: $t_{4\epsilon}^{\text{MUON}}$ (the same for every β in the power-law panels); squares: $t_{4\epsilon}^{\text{SIGNSGD}}$; grey lines: the predicted powers of ϵ. At $\varsigma = 0.7$ and $\varsigma = 1.5$ (Phase C) MUON pulls ahead polynomially in $1/\epsilon$; at $\varsigma = 3.0$ (Phase B) both scale as $\epsilon^{-1/2}$ and SIGNSGD wins by the factor $D_N \approx 0.42$.

Conclusions. Our analysis of the high-dimensional dynamics of spectral algorithms in the high-dimensional matrix-valued linear regression suggests that even this relatively simple model can capture key relationships between diagonal and non-diagonal optimization methods. Future work includes studying power-law random features models and deriving explicit compute-optimal neural scaling laws. In particular, our work does not capture the effects of feature learning, or non-linear dynamical effects like the edge-of-stability [16, 17]. The results in the anisotropic case suggest that there is an important interplay between the batch size, data distribution, and the ability to usefully apply non-diagonal preconditioning. This may be helpful to explain the mixed success of deploying MUON in different workloads, and may point to new directions in algorithmic design space.

Acknowledgments and Disclosure of Funding

This research was enabled in part by compute resources, software and technical help provided by McGill math cluster (https://www.mcgill.ca/mathstat/) as well as by support provided by Calcul Québec (https://www.calculquebec.ca/) and the Digital Research Alliance of Canada (alliancecan.ca). No Google computing resources or data were used in this submission.

N. Marshall is supported by a CGS-D from the Natural Science and Engineering Research Council (NSERC). L. Benigni is supported by NSERC RGPIN 2023-03882 & DGECR 2023-00076 and FRQNT Établissement de la Relève Professorale 364387. C. Paquette is a Canadian Institute for Advanced Research (CIFAR) AI chair, Quebec AI Institute (Mila) and a Sloan Research Fellow in Computer Science (2024). C. Paquette is supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada, NSERC CREATE grant Interdisciplinary Math and Artificial Intelligence Program (INTER-MATH-AI), Google x Mila research grant, Fonds de recherche du Quebec Nature et technologies (FRQNT) NOVA Grant (DOI: https://doi.org/10.69777/363444), and CIFAR AI Catalyst Grant. Research by E. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) and a gift from Google Canada. E. Paquette also acknowledges support from AFOSR grant FA9550261B232. Additional revenues related to this work: C. Paquette has 20% part-time employment at Google DeepMind.

References

[1] Ben Adlam and Jeffrey Pennington. Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 11022–11032, 2020.

[2] Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal Matrix Sign Methods and Their Application to the Muon Algorithm. In *International Conference on Learning Representations (ICLR)*, 2026.

[3] Lukas Balles and Philipp Hennig. Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 404–413. PMLR, PMLR, 2018.

[4] Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The Geometry of Sign Gradient Descent. *arXiv preprint arXiv:2002.08056*, 2020.

[5] Gérard Ben Arous, Murat A Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025.

[6] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25349–25362, 2022.

[7] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[8] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569. PMLR, 2018.

[9] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A Dynamical Model of Neural Scaling Laws. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 4345–4382. PMLR, 2024.

[10] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How Feature Learning Can Improve Neural Scaling Laws . *International Conference on Learning Representations (ICLR)*, 2025.

[11] James Bradbury, Roy Frostig, Peter Hawkins, Mathew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018.

[12] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[13] David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Artificial intelligence and statistics*, pages 111–119. PMLR, 2015.

[14] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[15] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research (TMLR)*, 2026.

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

[17] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, pages 1–80, 2021.

[18] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models. *Inf. Inference*, 13(4):Paper No. iaae028, 107, 2024.

[19] Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W Mackenzie, Elliot Paquette, and Courtney Paquette. The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[20] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[21] Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

[22] DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem. `http://github.com/google-deepmind`, 2020.

[23] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[24] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363 – 1399, 2016.

[25] Zhehang Du and Weijie Su. The Newton-Muon Optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

[26] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data. *arXiv preprint arXiv:2502.04664*, 2025.

[27] Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted Momentum Outscales SGD. *arXiv preprint arXiv:2505.16098*, 2025.

[28] Reza Gheissari and Aukosh Jagannath. Universality of high-dimensional scaling limits of stochastic gradient descent. *arXiv preprint arXiv:2512.13634*, 2025.

[29] Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on Muon from Simple Quadratics. *arXiv preprint arXiv:2602.11948*, 2026.

[30] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 2018.

[31] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[32] Aukosh Jagannath, Taj Jones-McCormick, and Varnan Sarangian. High-dimensional limit theorems for SGD: Momentum and Adaptive Step-sizes. *arXiv preprint arXiv:2511.03952*, 2025.

[33] Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive Matrix Online Learning through Smoothing with Guarantees for Nonsmooth Nonconvex Optimization. *arXiv preprint arXiv:2602.08232*, 2026.

[34] Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

[35] Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton-Schulz. In *International Conference on Learning Representations (ICLR)*, pages 1–29, 2026.

[36] Jihwan Kim, Dogyoon Song, and Chulhee Yun. Scaling Laws of SignSGD in Linear Regression: When Does It Outperform SGD? In *International Conference on Learning Representations (ICLR)*, pages 1–89, 2026.

[37] Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D Lee. Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory. *arXiv preprint arXiv:2603.26554*, 2026.

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

[39] 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.

[40] Xuheng Li, Yihe Deng, Jingfeng Wu, Dongruo Zhou, and Quanquan Gu. Risk bounds of accelerated SGD for overparameterized linear regression. In *International Conference on Learning Representations (ICLR)*, pages 1–69, 2024.

[41] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling Laws in Linear Regression: Compute, Parameters, and Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 60556–60606, 2024.

[42] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is Scalable for LLM Training. *arXiv preprint arXiv:2502.16982*, 2025.

[43] Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A Hybrid Adaptive Optimizer for Efficient Training of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2026.

[44] Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: SGD with Normalization and Whitening Enables Stateless LLM training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 41907–41942. PMLR, 2025.

[45] Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To Clip or not to Clip: the Dynamics of SGD with Gradient Clipping in High-dimensions. In *International Conference on Learning Representations (ICLR)*, pages 1–37, 2025.

[46] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015.

[47] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[48] Andrea Montanari and Zihao Wang. Phase transitions for feature learning in neural networks. *arXiv preprint arXiv:2602.01434*, 2026.

[49] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers. In *International Conference on Learning Representations (ICLR)*, 2021.

[50] Yuri Nesterov. *Introductory lectures on convex optimization*. Springer, 2004.

[51] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)*, volume 134, pages 3548–3626, 2021.

[52] Courtney Paquette and Elliot Paquette. Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:9229–9240, 2021.

[53] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: exact dynamics and generalization properties. *Mathematical Programming*, pages 1–90, 2024.

[54] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 Phases of Compute-Optimal Neural Scaling Laws. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[55] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[56] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training Deep Learning Models with Norm-Constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 49069–49104. PMLR, 2025.

[57] Gabriel Peyré. Muon Dynamics as a Spectral Wasserstein Flow. *arXiv preprint arXiv:2604.04891*, 2026.

[58] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[59] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 04, 1964.

[60] Shikai Qiu, Lechao Xiao, Andrew Gordon Wilson, Jeffrey Pennington, and Atish Agarwala. Scaling collapse reveals universal dynamics in compute-optimally trained neural networks. In *International Conference on Machine Learning*, pages 50697–50720. PMLR, 2025.

[61] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion Great Again!(Bridging Theory and Practice of LMO-based Optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025.

[62] Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical Efficiency of Muon for Pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

[63] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the Convergence Analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

[64] Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A Distributed Data-Parallel Pytorch Implementation of the Distributed Shampoo Optimizer for Training Neural Networks At-Scale. *arXiv preprint arXiv:2309.06497*, 2023.

[65] Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the Ideal: Analyzing the Inexact Muon Update. *arXiv preprint arXiv:2510.19933*, 2025.

[66] James B. Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: when infinite overparameterization is optimal and overfitting is obligatory. In *International Conference on Learning Representations (ICLR)*, pages 1–40, 2024.

[67] Weijie Su. Isotropic Curvature Model for Understanding Deep Learning Optimization: Is Gradient Orthogonalization Optimal? *arXiv preprint arXiv:2511.00674*, 2025.

[68] Nikolaos Tsilivis, Eitan Gronich, Julia Kempe, and Gal Vardi. Flavors of Margin: Implicit Bias of Steepest Descent in Homogeneous Neural Networks. In *International Conference on Learning Representations (ICLR)*, pages 1–27, 2025.

[69] Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022.

[70] Aditya Varre and Nicolas Flammarion. Accelerated sgd for non-strongly-convex least squares. In *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*, volume 135 of *Proceedings of Machine Learning Research*, pages 2062–2126, 2022.

[71] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's Spectral Design Benefits Generalization: A Study on Imbalanced Data. In *International Conference on Learning Representations (ICLR)*, pages 1–36, 2026.

[72] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and Stabilizing Shampoo Using Adam. In *International Conference on Learning Representations (ICLR)*, pages 1–22, 2025.

[73] Guangyuan Wang, Elliot Paquette, and Atish Agarwala. High-dimensional isotropic scaling dynamics of Muon and SGD. In *OPT 2025: Optimization for Machine Learning*, 2025.

[74] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon Outperforms Adam in Tail-End Associative Memory Learning. In *International Conference on Learning Representations (ICLR)*, pages 1–38, 2026.

[75] 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 (ICML)*, 2022.

[76] Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic Pretraining Optimizers and Where to Find Them. In *International Conference on Learning Representations (ICLR)*, pages 1–107, 2026.

[77] 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 *International Conference on Learning Representations (ICLR)*, pages 1–48, 2025.

[78] Dmitry Yarotsky and Maksim Velikanov. SGD with Memory: Fundamental Properties and Stochastic Acceleration. *International Conference on Learning Representations (ICLR)*, pages 1–43, 2025.

[79] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Risk Bounds of Multi-Pass SGD for Least Squares in the Interpolation Regime. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 12909–12920, 2022.

Supplementary Material

Broader Impact Statement. The work presented in this paper is foundational research and it is not tied to any particular application. The primary set-up is on a simple well-studied matrix-valued linear regression problem on synthetic data and solved using commonly deployed algorithms, variants of ADAM (SIGNSGD) and MUON (SIGNSVD). We present deterministic dynamics for the training loss for a class spectral algorithms. The results are theoretical and we do not anticipate any direct ethical and societal issues. The results show in a highly synthetic setting the scaling behaviour of MUON/SIGNSVD relative to SIGNSGD as batch size and parameter size grow. This does have some implications to practitioners in designing better compute-optimal scalings for these commonly used algorithms in practice on large language models. While a common positive foreseeable impact of algorithms with better scaling properties is that it would allow more efficient model training and hence lower energy consumption of AI, it is important to acknowledge that, many times across history, cost-lowering technological improvements have led to increase in consumption due to Jevons paradox. Finally, since our main results are theorems regarding scaling laws, we anticipate that one potential negative impact of such a line of research is that fundamental advance in scaling laws will advantage large entities that have more computational resource, thus enhancing centralization of technological power.

LLM Declaration. Large language models were used extensively in this work. The project started from several proofs, some written by hand and some generated semi-autonomously. These were then greatly expanded through an LLM-led exploration, which added the analysis with momentum and a complete proof of the dynamical comparison between the algorithm and its deterministic dynamics. The resulting appendix was audited for style and correctness. Its statements and proofs were then formalized in Lean 4 with Mathlib and checked by machine, using standard results proved in the accompanying foundations library, and the formalization status of every statement is listed in appendix SK of the supplement. Independently, the theorems were verified numerically. In the main text, LLMs were used for some initial drafting, for generating figures, and for connecting the text to the proofs; all text was revised by hand. LLMs were also used to write code for the numerical validation, for example to solve the deterministic equivalent.

Proofs and code. The accompanying website provides interactive access to the statements, proof sketches and complete written proofs: `https://mlexpos.github.io/phases-of-muon/`. The Lean formalization is available at `https://github.com/mlexpos/phases-of-muon`, with its accompanying foundations libraries at `https://github.com/mlexpos/alea-foundations`.

Contents

1 Introduction **1**

1.1 Model setup 3

1.2 Stochastic spectral algorithms. 3

2 Learning curves and deterministic dynamics **4**

3 MUON and SIGNSGD in the isotropic setting **5**

3.1 Learning rates: floors and schedules 5

3.2 Rate comparison: the MUON speedup 6

4 Anisotropic output covariance ($\Sigma_{\text{in}} = \mathrm{Id}_N$, Σ_{out} = general) **7**

4.1 Large batch: square-root preconditioning 8

4.2 Power-law specialization $\mu_i = i^{-\alpha}$ 8

5 Time to ϵ-approximate solution: performance of MUON vs. SIGNSGD **9**

A Additional Related Work **19**

B Numerical simulations for Figures 3, 4 and 6 **21**

C Finite-ridge effects **23**

C.1 Newton–Schulz as a ridge 23

C.1.1 Frozen-error calibration: method 23

C.1.2 Frozen-error calibration: results 24

C.1.3 Dynamical validation 26

C.1.4 Learning rates 27

D Momentum in MUON: theory against an ablation **29**

D.1 What the theory predicts at $B = 2N$ 30

D.2 The noise floor 30

D.3 Time to ϵ 31

E Roadmap of the results and their proofs **34**

F Model, algorithm and standing assumptions **34**

G Proof strategies **36**

G.1 Deterministic equivalent for the EMA Gram resolvent 36

G.2 Drift: the expected regularized update 37

G.3 Volatility and the lag kernel 38

G.4 From the algorithm to the deterministic dynamics 39

G.5 Power-law and large-batch asymptotics 40

G.6 Volterra analysis and the phase map 41

H Dynamics: from the algorithm to the deterministic recursion **42**

I Drift **46**

J Volatility **50**

K The isotropic case **52**

L Power-law spectra and the large-batch regime **65**

M The phase map and time to accuracy **79**

N SignSGD **90**

A Additional Related Work

MUON & SIGNSVD. The original motivation for MUON [34] stems from steepest descent with respect to the spectral norm [7]. The method shares key features with related approaches, including stochastic spectral descent [13] and orthogonalized gradient methods [69], and can also be interpreted as a form of SIGNSGD [8]. Its empirical success [34, 42, 62, 76] has motivated a number of recent variants [43, 44, 56, 61]. Given the growing interest in MUON, the discussion below is necessarily non-exhaustive and highlights only a subset of recent theoretical and empirical developments.

On the theoretical side, several works have analyzed the performance of MUON. For instance, [15, 35, 63, 65] study MUON or closely related variants such as SIGNSVD using classical worst-case complexity arguments for non-convex optimization, showing that these methods achieve $\mathcal{O}(1/\sqrt{T})$ convergence rates comparable to SGD. More specifically, [65] analyzes MUON via an inexact linear minimization oracle, while [15] connects MUON to a broader class of algorithms known as Lion-$\mathcal{K}$. In [35], the authors further characterize the dependence on the number of Newton–Schultz steps and compare MUON with exact polarization. In addition, [33] establishes regret bounds for MUON.

Several works have studied SIGNSVD [71] and MUON [29] in simplified linear settings. In [71], explicit loss dynamics for SIGNSVD are derived in the gradient flow regime (vanishing learning rate and negligible stochasticity), enabling analysis of generalization for both SIGNSVD and standard gradient descent; see also [57] which studies gradient flow SIGNSVD but on more complicated loss functions. Meanwhile, [29] shows that even for simple quadratic objectives, MUON exhibits nontrivial behavior. However, these works do not operate in the high-dimensional regime considered here. In contrast, [73] studies SIGNSVD on a high-dimensional isotropic Gaussian least squares problem and derives explicit risk dynamics, but does not analyze the implications of the dynamics. Building on this line of work, this work moves beyond the isotropic setting and analyze the precise scaling behavior of both MUON and SIGNSVD as well as the explicit dynamics.

Beyond least squares, [26, 68] study the implicit bias of spectral methods in linear multiclass logistic regression with separable data. Additionally, [37, 39] analyze MUON in attention through associative memory models; in particular, [37] shows that the critical batch size scales as $\sqrt{N_{\text{in}}}$ under power-law covariance. To our knowledge, this is the first work to provide explicit training dynamics of spectral methods over the full course of optimization with anisotropic data.

Finally, several works aim to explain the effectiveness of MUON and related spectral optimizers in deep learning, primarily through the lens of a *single update*. For instance, [21] identifies a layerwise condition, known as low stable rank, under which a spectral update yields a larger loss decrease than a Euclidean gradient step in the deterministic MUON setting, and shows that intermediate-layer activations in deep networks often satisfy this property. Similarly, [67] analyzes a single deterministic MUON update under an isotropic assumption on the data and the loss, reducing the loss to a univariate function and showing that MUON is directionally optimal, though not strictly optimal. However, [29] demonstrates on simple quadratic problems that single-step analyses do not reliably predict overall performance. In contrast, our work provides explicit dynamics for *stochastic* spectral optimizers over the full course of training.

Random matrix theory and machine learning. This paper uses random matrix theory to analyze a matrix least squares problem. In particular, our analysis requires analysis of spectral properties of sample covariance matrices (e.g., power-law population covariance), and the use of deterministic equivalents for the resolvents of random matrices. Both these sets of tools have long predated modern machine learning literature, yet have found growing use in the field. See [20] for a modern introduction.

High-dimensional optimization literature for SGD and its variants. Most of the literature on high-dimensional optimization focuses on streaming SGD under certain settings such as least squares beyond isotropic settings [53]. A few works have expanded beyond least squares, see e.g., [6, 18, 47, 48], but again with focus on streaming SGD with many working on power-law covariances [5, 9, 10, 27, 41, 54]. There have been some extensions beyond SGD. Notably, [51, 53, 79] gave exact dynamics for multi-pass SGD on a high-dimensional empirical least squares problem with non-Gaussian covariances. Extensions for deterministic dynamics of momentum algorithms in small batch [40, 52, 70, 78] and large batch setting [38] as well as power-law covariances in [27] where Nesterov acceleration methods with time-dependent momentum parameters were explored on

least squares problems and single/multi-index [28]. Deterministic dynamics for high-dimensional optimization beyond the SGD/SGD with momentum framework have been considered, see [19, 32] for stochastic adaptive learning rates (e.g., AdaGrad-Norm), SIGNSGD [36, 77], and clipped SGD [45]. High-dimensional limits for spectral optimizers have not yet been explored in detail, with the only related result being the workshop paper [73].

SIGNSGD. SIGNSGD has been proposed as a simplified proxy for ADAM, of which it is a special case (setting $\beta_1 = \beta_2 = \varepsilon = 0$) [3, 4]. In [77], the authors study SIGNSGD on a high-dimensional linear regression problem, deriving explicit loss dynamics that precisely quantify the *diagonal preconditioning* effect of SIGNSGD and characterize how it reshapes the distribution of gradient noise.

B Numerical simulations for Figures 3, 4 and 6

This appendix describes the simulations behind Figs. 3, 4 and 6 and the checks of the approximations (9). Nothing in it is used in a proof.

Model and algorithms. All runs use the square isotropic model with $N_{\rm in} = N_{\rm out} = N = 512$ and $\Sigma_{\rm in} = \Sigma_{\rm out} = \mathrm{Id}_N$. At an error $\Delta = W - W^\star$, a batch of size B gives the gradient

$$G = \frac{1}{B}\sum_{b=1}^{B} x^b_{\rm out} \otimes x^b_{\rm in} \left\langle x^b_{\rm out} \otimes x^b_{\rm in}, \Delta \right\rangle, \qquad \mathbb{E}\, G = \Delta, \qquad \mathscr{R} = \tfrac{1}{2}\|\Delta\|_{\rm F}^2.$$

MUON keeps the buffer $M_t = \beta M_{t-1} + (1-\beta)G_{t+1}$ and steps $W_{t+1} = W_t - \eta S_t$. With the exact polar factor, $S_t = UV^\top$ is computed from a singular value decomposition of M_t over the singular values above 10^{-13} times the largest. With five Newton–Schulz steps (NS-5), S_t is the result of five iterations of

$$X \leftarrow aX + \big(bXX^\top + c(XX^\top)^2\big)X$$

from $X_0 = M_t/\|M_t\|_{\rm F}$, with $(a,b,c) = (3.4445, -4.7750, 2.0315)$. SIGNSGD steps $W_{t+1} = W_t - \eta\,\mathrm{sign}(G_{t+1})$, without momentum. All arithmetic is in double precision.

Panels (a) and (b): frozen error. The error is frozen at $\Delta = \mathrm{Id}_N/\sqrt{N}$; by the invariance of the model under $\Delta \mapsto O_1 \Delta O_2$ with orthogonal O_1, O_2, this is equivalent in law to any Δ with the same singular values. The buffer is run to stationarity for $\lceil 37/\log(1/\beta) + N/B \rceil$ steps (none at $\beta = 0$), after which T steps are recorded, with $T = 100$ at $\beta = 0$ and $T = \max(600, 3N/B + 300)$ at $\beta = 0.9$, two seeds each. The drift estimate is the mean of $\langle S_t, \Delta\rangle/(\sqrt{2}\|\Delta\|_{\rm F})$. The volatility estimate is the mean of $\|S_t\|_{\rm F}^2$ plus twice the centred lag sum $\sum_{h\ge 1}\langle S_t - \bar{c}\hat{\Delta},\, S_{t-h} - \bar{c}\hat{\Delta}\rangle$, where $\hat{\Delta} = \Delta/\|\Delta\|_{\rm F}$ and $\bar{c}$ is the mean of $\langle S_t, \hat{\Delta}\rangle$; it is averaged over $t \ge T/2$, so that every lag up to $T/2$ enters with full weight. Both estimates agree with the quadrature to within about 1% for $B_{\rm eff}/N \le 4$. For larger $B_{\rm eff}/N$ the measured volatility lies below the $N \to \infty$ constant by a finite-N correction of relative size about $0.7\, B_{\rm eff}/N^2$: at $\beta = 0.9$ it is 2% at $B_{\rm eff}/N = 16$ and 9% at $B_{\rm eff}/N = 64$ for $N = 512$, and it roughly doubles each time N is halved (18% and 32% at $B_{\rm eff}/N = 64$ for $N = 256$ and 128). Quadrupling T does not change it. The drift is 1.4% low at $B_{\rm eff}/N = 64$.

Panel (c): rates. Each run starts from a Gaussian Δ_0 with $\mathscr{R}(0) = 6.25$ (or 4 at the smallest batches) and uses the constant step $\eta = 4\mathfrak{d}/\mathfrak{v}$ built from the theory constants, which puts the floor at $\mathscr{R}_\star = 1$ for full whitening. The recursion (8), $\mathscr{R}(t+1) = \mathscr{R}(t) - a\sqrt{\mathscr{R}(t)} + b$ with $a = 2\eta\mathfrak{d}$ and $b = \eta^2\mathfrak{v}/2$, is fitted to the seed-averaged risk by least squares in $\log\mathscr{R}$. Near the floor $\sqrt{\mathscr{R}_\star} = b/a$ the recursion contracts by the factor $1 - a^2/(2b) = 1 - 4\mathfrak{d}^2/\mathfrak{v}$ per step, which does not depend on η, and we report

$$1 - \text{rate} = \frac{a^2}{4b} = \frac{2\mathfrak{d}^2}{\mathfrak{v}},$$

the factor of (12). So the measured rate does not depend on the step, even when NS-5 reaches a floor different from the one it was tuned for. Runs last between $7/\kappa$ and $9/\kappa$ steps, $\kappa = 8\mathfrak{d}^2/\mathfrak{v}$.

The SIGNSGD drift is $\mathcal{N}_B/\sqrt{\pi}$ with $\mathcal{N}_B$ as in Sec. 3, and the main text writes $\sqrt{B}$ for $\mathcal{N}_B$. In panel (c) each measured ratio is multiplied by $\mathcal{N}_B^2/B$, which puts SIGNSGD on the $\sqrt{B}$ scale of the curves. We use $\mathcal{N}_1 = 2/\pi$, a Monte Carlo table with $2\cdot 10^6$ samples for $2 \le B \le 64$, and

$$\mathcal{N}_B = \sqrt{B}\Big(1 - \frac{1}{B} + \frac{5}{B^2} + O(B^{-3})\Big)$$

for $B > 64$. The expansion follows from the cumulants 1, 8 and 200 of $(xy)^2$; at $B = 64$ it agrees with the table to $4\cdot 10^{-5}$.

Grid. Panel (c) uses $B_{\rm eff}/N = 2^k$, $k = -4, \dots, 4$, at $\beta \in \{0, 0.5, 0.9\}$, with SIGNSGD, NS-5 and the exact polar factor at each point and two seeds, between 10^2 and $6\cdot 10^5$ steps. Full whitening inverts the oldest direction of the buffer, whose weight is $\beta^{N/B}$. Where this is below 10^{-11} the exact polar factor is not computable in double precision, and that marker is omitted: $\beta = 0.5$ at $B_{\rm eff}/N = 1/16$ ($0.5^{46.5} \approx 10^{-14}$) and $\beta = 0.9$ at $B_{\rm eff}/N = 1/16$ ($0.9^{256} \approx 2\cdot 10^{-12}$). The same rule removes the frozen point at $\beta = 0.9$, $B = 2$ from panels (a) and (b).

Newton–Schulz. The quintic coefficients above do not converge to the polar factor: with many steps the singular values of S_t oscillate in a band near $[0.68, 1.13]$, so $\|S_t\|_{\mathrm{F}}^2 \approx 0.86N$. This is close to a scalar multiple of the polar factor; it changes the floor, but it cancels in $2\mathfrak{d}^2/\mathfrak{v}$. In frozen runs at $\beta = 0.9$, $B = 13$ ($B_{\mathrm{eff}}/N \approx 0.48$), the rate with K steps equals the exact rate within the seed spread for every $K \geq 7$. With $K = 5$ the drift is 16% lower and the volatility 38% lower, so the rate is 14% higher: five steps do not reach the oldest buffer directions, whose singular values are below $a^{-5}\|M_t\|_{\mathrm{F}} \approx 2 \cdot 10^{-3}\|M_t\|_{\mathrm{F}}$, and act as a small positive ridge. The effect grows as B_{eff}/N decreases; at $\beta = 0.5$ the NS-5 rate exceeds the full-whitening rate by factors 1.15, 1.42, 1.76 and 1.96 at $B_{\mathrm{eff}}/N = 1/2, 1/4, 1/8$ and $1/16$, and at $\beta = 0.9$ by 1.37, 1.79 and 1.92 at $B = 7$, 3 and 2 ($B_{\mathrm{eff}}/N \approx 0.26$, 0.11 and 0.07). For $B_{\mathrm{eff}} \geq 2N$ NS-5 and full whitening agree within 3%. Without momentum the update has rank $\min\{B, N\}$ and there is no stale direction to suppress, and the sign of the effect reverses: for $B < N$ the NS-5 rate is 5–8% below the exact one (6%, 8% and 5% at $B/N = 1/8, 1/4$ and $1/2$), and at $B = N$ it is 7% above.

The approximations (9). The forms for $\beta > 0$ interpolate between the two batch limits: additively in C_β^2, because the thin-batch limit $1/(2\sqrt{\pi})$ is a positive constant, and additively in $\mathfrak{v}_{\mathrm{iso},\beta} B/B_{\mathrm{eff}}$. Inserted in the ratio of the factors $1 - \text{rate}$, they give

$$\frac{1 - \text{rate}_{\text{Muon}}}{1 - \text{rate}_{\text{s-SGD}}} \approx \frac{32/(9\pi) + N/(4B_{\mathrm{eff}})}{c_\beta + N/(2B_{\mathrm{eff}})},$$

which tends to $32/(9\pi c_\beta)$ for $B_{\mathrm{eff}} \gg N$ and to $\frac{1}{2}$ for $B_{\mathrm{eff}} \ll N$. Figure 9 compares them with the quadrature of the exact constants, and the table gives the largest relative error over $1/64 \leq B_{\mathrm{eff}}/N \leq 64$; all three approximations overestimate, most near $B_{\mathrm{eff}}/N \approx 0.3$.

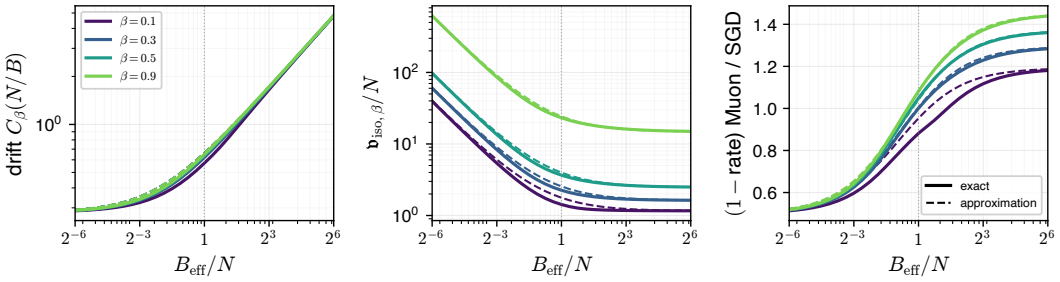


Figure 9: The algebraic forms (9) (dashed) against the exact zero-ridge constants (solid), for $\beta \in \{0.1, 0.3, 0.5, 0.9\}$. **Left:** drift constant. **Centre:** integrated volatility. **Right:** ratio of the factors $1 - \text{rate}$.

β	0.1	0.3	0.5	0.9	0.95
C_β	16%	8.0%	5.7%	4.6%	4.6%
$\mathfrak{v}_{\mathrm{iso},\beta}$	26%	16%	12%	8.8%	8.7%
ratio of $1 - \text{rate}$	8.7%	2.8%	2.1%	2.1%	2.1%

The accuracy degrades as $\beta \downarrow 0$ because the limit $\beta \downarrow 0$ at zero ridge is discontinuous for $N > B$ (proposition SF.12).

C Finite-ridge effects

The theory of the paper is proved for the regularized SIGNSVD map Q_ϱ of definition F.1, and its sharpest statements are made in the limit of vanishing ridge, which is exact whitening. Practical MUON does not whiten exactly: it applies a few Newton–Schulz steps, which amplify small singular values only up to a point. This appendix collects numerical evidence on what such imperfect whitening does, and on when a positive ridge describes it. Nothing in this appendix is used in a proof.

C.1 Newton–Schulz as a ridge

Fig. 6 shows that MUON with five Newton–Schulz steps (NS-5) and momentum behaves like a small positive ridge. This subsection measures, for every number of steps $K \le 7$, which ridge reproduces NS-K in the isotropic model, how accurately, and where the description fails. The calibration is made on a frozen error (Secs. C.1.1 and C.1.2) and then tested, without refitting, on the dynamics (Sec. C.1.3).

C.1.1 Frozen-error calibration: method

Model and updates. All runs use the square isotropic model of Appendix B, with $N = 512$ unless stated otherwise, momentum buffer $M_t = \beta M_{t-1} + (1-\beta)G_t$ and $\beta \in \{0, 0.05, 0.1, 0.25, 0.5, 0.9\}$. Every update we compare is a spectral function of the buffer. If $M_t = \sum_k \sigma_k\, p_k \otimes r_k$ is a singular value decomposition, then the update is $\sum_k \phi(\sigma_k)\, p_k \otimes r_k$ with

$$\phi_{\mathrm{polar}}(\sigma) = 1, \qquad \phi_\varrho(\sigma) = \frac{\sigma}{\sqrt{\sigma^2 + \varrho}}, \qquad \phi_{\text{NS-}K}(\sigma) = p^{\circ K}\Big(\frac{\sigma}{\|M_t\|_{\mathrm{F}}}\Big),$$

where $p(x) = ax + bx^3 + cx^5$ with $(a, b, c) = (3.4445, -4.7750, 2.0315)$ and $p^{\circ K}$ is its K-fold composition. The last formula is exact: K Newton–Schulz iterations from $X_0 = M_t/\|M_t\|_{\mathrm{F}}$ act on each singular value separately.

Units of the ridge. Newton–Schulz acts on $M_t/\|M_t\|_{\mathrm{F}}$, so we parametrize the ridge in the same units,

$$\varrho = \varrho_X\, \|M_t\|_{\mathrm{F}}^2.$$

In the frozen state of Appendix B, which has $2\mathscr{R} = 1$, the normalized and the physical ridge coincide, and the measured $\|M_t\|_{\mathrm{F}}^2$ agrees with $s_2 N^2/B = N^2/B_{\mathrm{eff}}$ to within 5% in every run (median ratio 1.013). A fixed ϱ_X therefore corresponds to the normalized ridge $\varrho_X N^2/B_{\mathrm{eff}}$, which is $\varrho_X N$ times the mean eigenvalue of $M_t M_t^\top$. The heuristic of section 1.2 matches the small-σ slope a^K of $\phi_{\text{NS-}K}$ with the slope $\varrho_X^{-1/2}$ of ϕ_ϱ, which gives $\varrho_X = a^{-2K}$.

The frozen filter bank. The error is frozen at $\Delta = \mathrm{Id}_N/\sqrt{N}$, and the buffer is burnt in and recorded exactly as in Appendix B: $\lceil 37/\log(1/\beta) + N/B \rceil$ burn-in steps (none at $\beta = 0$), then $T = \max(600, 3N/B + 300)$ recorded steps ($T = 400$ at $\beta = 0$). Because the error is frozen, the sequence M_t does not depend on the update rule. One double-precision singular value decomposition per step therefore gives, on the same random numbers, the frozen statistics of the exact polar factor, of NS-K for $K = 1, \dots, 7$, and of exact ridges at the 16 values $\varrho_X \in \{10^{-10}, 10^{-9.33}, \dots, 1\}$. For each update we record the drift $\mathfrak{d}$, the integrated volatility $\mathfrak{v}$ and the same-time volatility $\mathfrak{v}_{\mathrm{same}} = \mathbb{E}\,\|S_t\|_{\mathrm{F}}^2$ with the estimators of Appendix B.

Two scale-free statistics. NS-K does not output singular values equal to one, so its updates differ from any ridge by an overall amplitude. Two statistics do not change when the update is multiplied by a constant:

$$I_1 = \frac{\mathfrak{d}^2}{\mathfrak{v}}, \qquad I_2 = \frac{\mathfrak{v}_{\mathrm{same}}}{\mathfrak{v}}.$$

The first sets the rate at the optimal step, $1 - \mathrm{rate} = 2I_1$, and $\pi\gamma N I_1$ is the ratio plotted in Fig. 6 (the MUON speedup). The second is the share of the volatility that comes from a single step.

Calibration. For each configuration (N, B, β) and each K, the effective ridge $\varrho_X(K)$ is the value at which the exact-ridge bank has the I_1 of NS-K, found by linear interpolation in $\log \varrho_X$. This calibration uses no theory and no finite-N correction, because the bank and NS-K are evaluated on the same buffers. As a second, independent calibration we solve the same equation with the positive-ridge theory of Appendix SF, that is $\mathfrak{d}_{\rm iso}(\varrho)$ and $\mathfrak{v}_{\rm iso}(\varrho)$ at $\varrho = \varrho_X s_2 N^2/B$. When the I_1 of NS-K lies outside the range spanned by the ridges, we report that no ridge matches. The fitted ridge then predicts I_2, which was not used in the fit.

Grid. At $N = 512$: $\beta \in \{0, 0.05, 0.1, 0.25, 0.5, 0.9\}$ and $B_{\rm eff}/N = 2^k$, $k = -4, \dots, 4$, two seeds, 108 runs. For the dependence on N: $N \in \{256, 1024\}$ at $\beta \in \{0.1, 0.9\}$ and $B_{\rm eff}/N \in \{1/8, 1/2, 2\}$, two seeds, 24 runs.

C.1.2 Frozen-error calibration: results

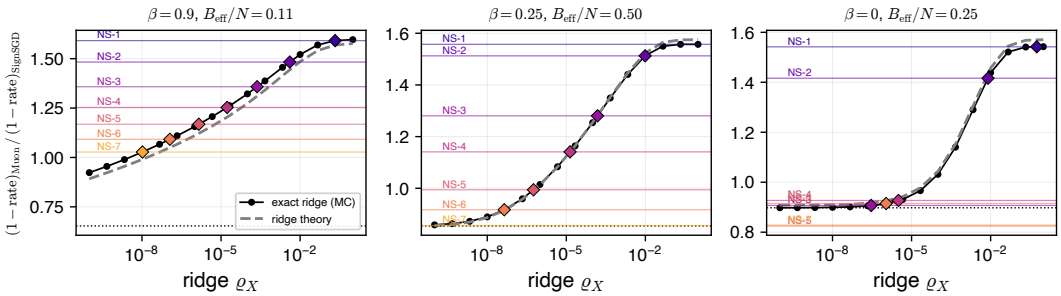


Figure 10: Rate statistic $\pi\gamma N I_1 = (1 - \text{rate})_{\rm Muon}/(1 - \text{rate})_{\rm SignSGD}$ (values above 1: MUON converges faster) of the exact ridge as a function of ϱ_X (black: frozen bank; dashed: ridge theory), the levels of NS-K (horizontal lines) and the calibrated ridges (diamonds), for three configurations. Dotted: exact polar factor. Right: at $\beta = 0$ the levels of NS-5 and NS-7 lie below every ridge, and no ridge matches.

Calibration curves. Fig. 10 shows the mechanism. The rate statistic of the ridge increases monotonically from the exact polar value at $\varrho_X \to 0$ to the value of a linear update at large ϱ_X, and each NS-K level crosses it once. The theory curve follows the bank to within a few percent; the gap at $\beta = 0.9$, $B = 3$ is the finite-N drift bias of the polar factor at large N/B noted in Appendix B. At $\beta = 0$ the NS-K levels with $K \geq 5$ lie below the exact polar value, which no ridge reaches.

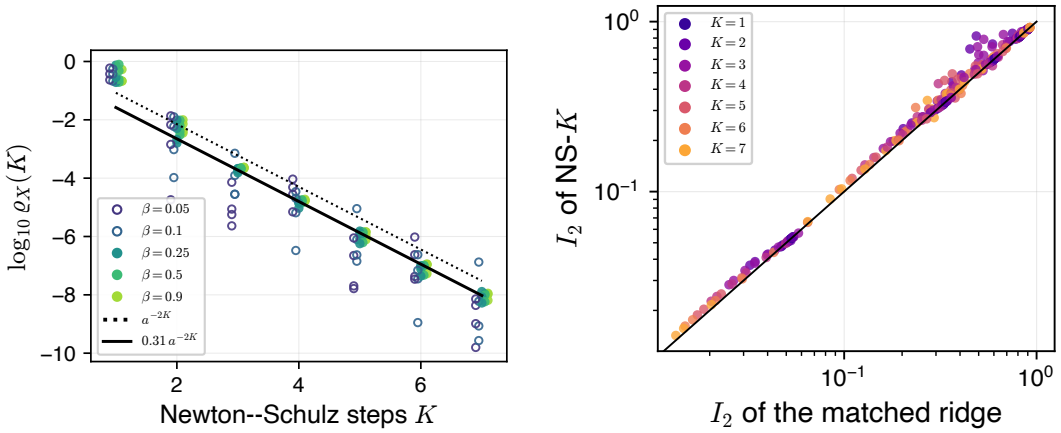


Figure 11: Left: calibrated $\log_{10} \varrho_X(K)$ for every configuration with $B_{\rm eff} < N$ and $\beta > 0$ (filled: $\beta \geq 0.25$; open: $\beta \in \{0.05, 0.1\}$), with the heuristic a^{-2K} (dotted) and the fit $0.31\, a^{-2K}$ (solid). Right: the lag share I_2 of NS-K against that of the matched ridge, all $\beta > 0$; the line is equality.

K	bank: median [range]	theory: median	$\log_{10} a^{-2K}$	$\log_{10} 0.31a^{-2K}$	I_2: NS/ridge-1
1	-0.58 $[-0.72, -0.10]$	-1.37	-1.07	-1.58	$+0.006$
2	-2.34 $[-2.80, -2.00]$	-2.36	-2.15	-2.65	$+0.11$
3	-3.67 $[-3.82, -3.63]$	-3.75	-3.22	-3.73	$+0.13$
4	-4.78 $[-5.03, -4.73]$	-4.84	-4.30	-4.80	$+0.10$
5	-5.92 $[-6.25, -5.81]$	-6.03	-5.37	-5.87	$+0.09$
6	-7.01 $[-7.35, -6.94]$	-7.14	-6.45	-6.95	$+0.07$
7	-8.01 $[-8.28, -7.89]$	-8.19	-7.52	-8.02	$+0.07$

Table 2: Calibrated $\log_{10} \varrho_X(K)$ at $N = 512$, $\beta \in \{0.25, 0.5, 0.9\}$, $B_{\text{eff}}/N \in \{1/16, \dots, 1/2\}$ (twelve configurations per K, nine at $K = 7$ where the polar end of the bank is not reached).

The effective ridge. For $\beta \geq 0.25$ and $B_{\text{eff}} < N$ the calibrated ridges collapse onto one law (Fig. 11, Table 2),

$$\varrho_X(K) \approx 0.31\, a^{-2K}, \qquad a = 3.4445,$$

with a scatter of 0.13 in $\log_{10} \varrho_X$ over the 57 calibrations with $K \geq 3$. A fit with a free slope gives -1.11 decades per step, against $-2 \log_{10} a = -1.07$. The heuristic a^{-2K} has the right dependence on K and is a factor of about three too large. The theory calibration agrees with the bank calibration to within 0.2 in $\log_{10} \varrho_X$ for $K \geq 2$, and lies slightly below it. At $K = 1$ the rate statistic of the ridge is nearly flat at large ϱ_X (one Newton–Schulz step on $M_t/\|M_t\|_{\text{F}}$ is almost linear), so $\varrho_X(1)$ is only weakly determined.

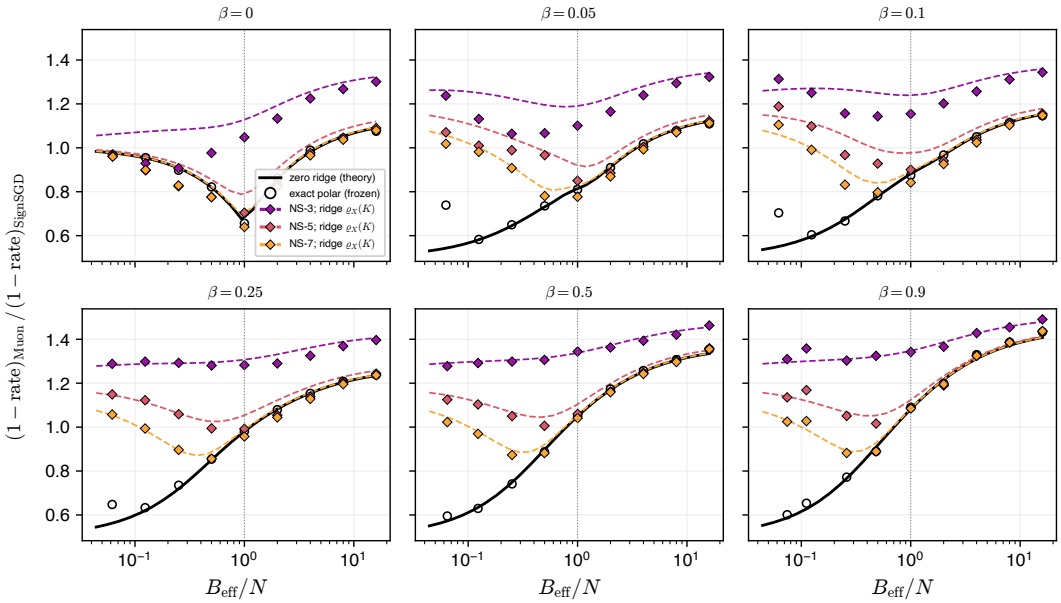


Figure 12: Frozen rate statistic $\pi\gamma N I_1 = (1 - \text{rate})_{\text{Muon}}/(1 - \text{rate})_{\text{SignSGD}}$ against B_{eff}/N for each momentum; values above 1 mean that MUON converges faster than SIGNSGD. Black: zero-ridge theory and frozen exact polar factor. Diamonds: frozen NS-3, NS-5, NS-7. Dashed: ridge theory at the single calibrated ridge $\varrho_X(K) = 0.31\, a^{-2K}$, the same for every panel.

Where the ridge describes Newton–Schulz. Fig. 12 compares every frozen NS-K measurement with the ridge theory at the calibrated $\varrho_X(K)$, with no further fitting.

(i) For $\beta \geq 0.25$ the calibrated ridge reproduces NS-3, NS-5 and NS-7 across the whole range $1/16 \leq B_{\text{eff}}/N \leq 16$, including the minimum near $B_{\text{eff}}/N \approx 1/2$ and the gain over the exact polar factor at small B_{eff}/N.

(ii) For $B_{\text{eff}} \geq N$ the effect of the ridge is small: for $K \geq 4$ the rate statistic of NS-K is within 2% of that of the exact polar factor, and the effective ridge is not determined there.

(iii) For $\beta \in \{0.05, 0.1\}$ a ridge still matches each NS-K, and Newton–Schulz still improves on the exact polar factor at small batch, but the matched ridge scatters over one to two decades and is not monotone in K at $\beta = 0.05$. With small momentum the weights β^j of old batches separate the singular values of the buffer into widely spaced groups, and any ridge between two groups has the same effect.

(iv) At $\beta = 0$ no ridge describes NS-K. For $K \geq 5$ and $B < N$, NS-K is slower than the exact polar factor (by 6% at $K = 5$), while every ridge is faster. The buffer is then a single minibatch gradient of rank $\min(B, N)$, with no small singular values for a ridge to suppress, and the remaining effect of Newton–Schulz is the spread of its outputs over $[0.68, 1.2]$, which adds volatility.

At the smallest batches with $\beta > 0$ the frozen exact polar factor lies above the zero-ridge curve. There the oldest directions of the buffer have weights below the rank threshold 10^{-13} of Appendix B, so the computed polar factor is itself a polar factor with a tiny ridge.

One Newton–Schulz step is SGD. In every panel the rate statistic decreases with K, and NS-1 is the fastest. One Newton–Schulz step is nearly linear on $M_t/\|M_t\|_{\mathrm{F}}$, whose bulk singular values are of order $N^{-1/2}$, so NS-1 is essentially momentum SGD with a normalized step. In the isotropic model there is nothing to precondition, and SGD is the most efficient of these updates: for nearly Gaussian gradient entries a linear update beats the sign update by the factor $\pi/2 \approx 1.57$, and NS-1 sits close to this value for every β and B_{eff}/N. In this model, taking the fewest steps therefore amounts to running SGD.

Consistency. The ridge fitted on I_1 predicts the lag share I_2 of NS-K to within about 10% (Fig. 11, right): NS-K has 7–13% more same-time volatility than the matched ridge for $2 \leq K \leq 7$, and the same at $K = 1$. The excess is the volatility added by the spread of the Newton–Schulz outputs, which a ridge does not have.

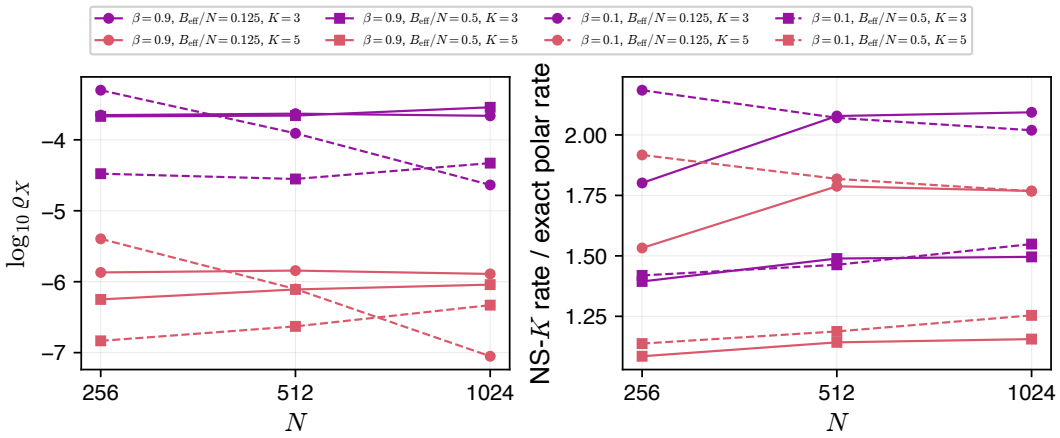


Figure 13: Dependence on N. Left: calibrated $\log_{10} \varrho_X$. Right: rate statistic of NS-K divided by that of the exact polar factor. Solid: $\beta = 0.9$; dashed: $\beta = 0.1$.

Dependence on the dimension. At $\beta = 0.9$ the calibrated ϱ_X does not depend on N:

$$\log_{10} \varrho_X(5) = -5.87, -5.84, -5.89, \qquad \log_{10} \varrho_X(3) = -3.65, -3.63, -3.66$$

at $N = 256, 512, 1024$ and $B_{\mathrm{eff}}/N = 1/8$ (Fig. 13). The Frobenius-normalized ridge is therefore the right unit, and the normalized ridge $\varrho_X N^2/B_{\mathrm{eff}}$ grows with N at a fixed B_{eff}/N. Accordingly the gain of NS-5 over the exact polar factor grows with N, from 1.09 to 1.14 to 1.16 at $B_{\mathrm{eff}}/N = 1/2$. At $\beta = 0.1$ the calibrated ridge drifts with N, as the scatter in (iii) suggests.

C.1.3 Dynamical validation

Protocol. The dynamical runs follow the protocol of Fig. 6, described in Appendix B. Each run starts from a Gaussian Δ_0 with $\mathscr{R}(0) = 6.25$ (4 at the smallest batches), warms the buffer at W_0 for

the burn-in of the frozen runs, and uses the constant step $\eta = 4\mathfrak{d}/\mathfrak{v}$ built from the frozen $(\mathfrak{d}, \mathfrak{v})$ of the same NS-K, so that its floor is near $\mathscr{R}_\star = 1$. Runs last $7/\kappa$ to $9/\kappa$ steps, $\kappa = 8\mathfrak{d}^2/\mathfrak{v}$, between $2 \cdot 10^2$ and $3.5 \cdot 10^5$ steps. Newton–Schulz is computed by matrix products in single precision, as in practice, not by the spectral formula. The grid is $N = 512$, $\beta \in \{0, 0.1, 0.25, 0.5, 0.9\}$, $B_{\text{eff}}/N = 2^k$ with $k = -4, \dots, 3$, $K = 1, \dots, 7$ and two seeds, 560 runs; the runs with different K share their initial errors. The rate is read from the recursion fit of Appendix B, and we report the rate statistic $(1 - \text{rate})_{\text{Muon}}/(1 - \text{rate})_{\text{SignSGD}}$ on the $\sqrt{B}$ scale of the main text.

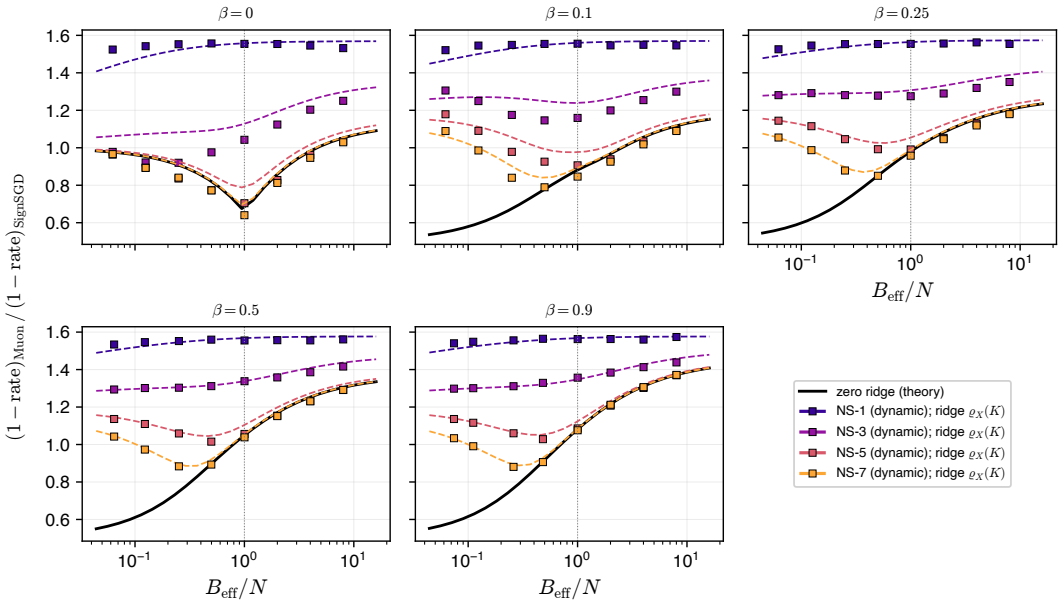


Figure 14: Dynamical rate statistic $(1 - \text{rate})_{\text{Muon}}/(1 - \text{rate})_{\text{SignSGD}}$ of NS-K (squares; values above 1: MUON converges faster) against the ridge theory at the frozen calibration $\varrho_X(K) = 0.31\, a^{-2K}$ (dashed), with no refitting. Solid: zero ridge.

Results. The dynamics reproduce the frozen calibration, and the calibrated ridge predicts the dynamical rates where the frozen calibration said it would (Fig. 14).

(i) The dynamical rate statistic agrees with the frozen $\pi\gamma N I_1$ of the same NS-K to 0.8% on average and 4.5% at most, over all 280 pairs of configuration and K; the two seeds differ by 0.3% in the median. Single-precision Newton–Schulz therefore behaves as its double-precision spectral formula.

(ii) For $\beta \geq 0.25$ the ridge theory at $\varrho_X(K)$ predicts the dynamical rate of NS-K to 0.6% on average (2.7% at most) for $B_{\text{eff}}/N \leq 1/4$, and to 2.1% on average at $B_{\text{eff}}/N = 1/2$. For $B_{\text{eff}} \geq N$ it is 2.6% too fast on average (8.6% at most): there NS-K is close to the exact polar factor, and the spread of its outputs, which a ridge does not have, is the larger effect.

(iii) For $\beta = 0.1$ the error is 4.8% on average and up to 11%, and for $\beta = 0$ it is 6.5% on average and up to 15%, with the ridge predicting too fast a rate, as the frozen results anticipated.

C.1.4 Learning rates

Sec. 3.1 gives the optimal learning rate $\eta_\star/\sqrt{\mathcal{R}} = 2\mathfrak{d}/\mathfrak{v}$ of exact whitening. Fig. 15 shows the same quantity for NS-K, measured on the frozen bank ($N = 512$), against the ridge theory at $\varrho_X(K)$.

Finite Newton–Schulz flattens the small-batch learning rate. With momentum and exact whitening, the learning rate grows like B for $B_{\text{eff}} < N$ (slope 0.95 in $\log B_{\text{eff}}$ between $B_{\text{eff}}/N = 1/16$ and 1), because the volatility of the whitened weak directions of the buffer grows like N^2/B. A ridge caps the gain on those directions: the slope falls to 0.25–0.34 for NS-5 at every $\beta \in [0.1, 0.9]$, and at $B_{\text{eff}}/N = 1/16$ the NS-5 learning rate is 6–8 times that of exact whitening. The picture is then close

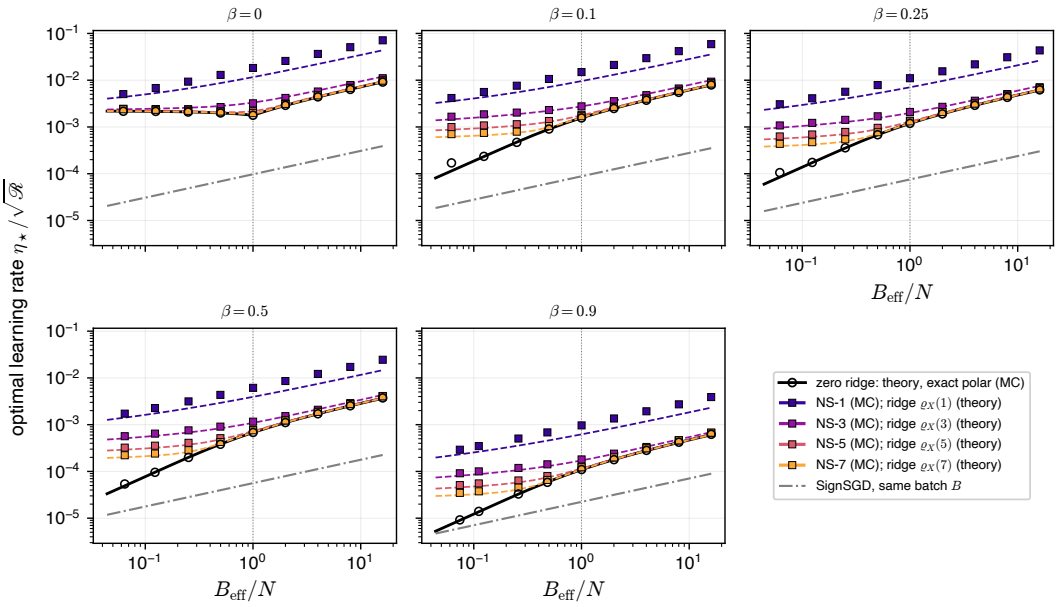
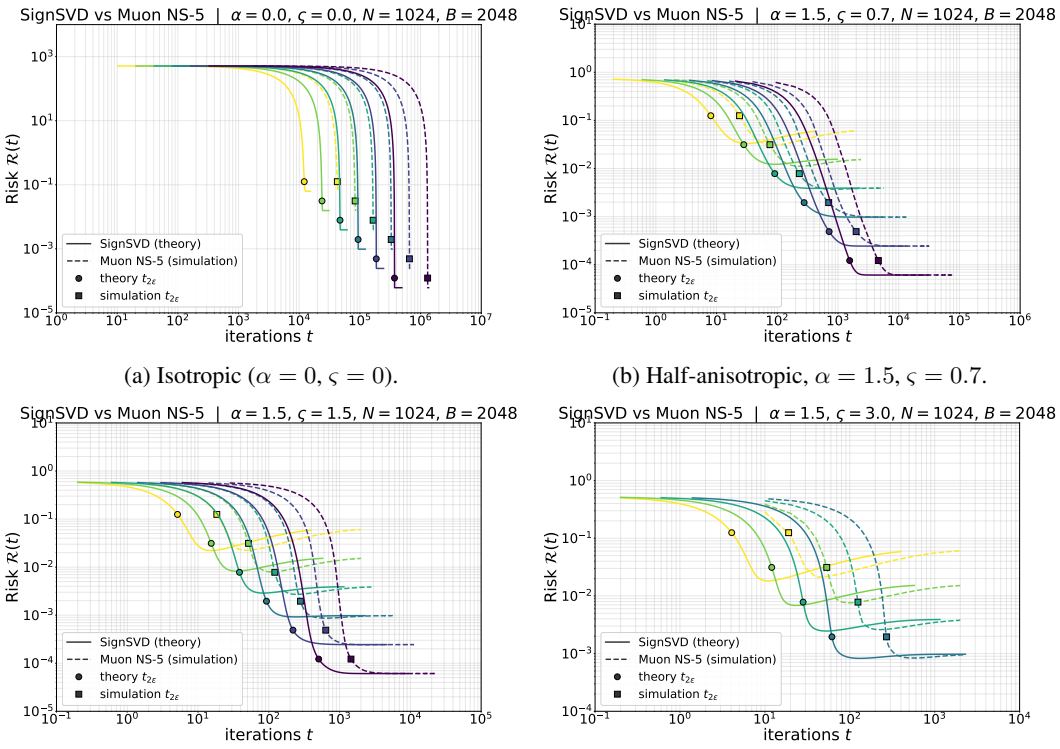


Figure 15: Optimal learning rate $\eta_\star/\sqrt{\mathcal{R}} = 2\mathfrak{d}/\mathfrak{v}$ against B_{eff}/N, $N = 512$. Black: zero ridge (theory) and exact polar factor (open circles, frozen bank). Dashed: ridge theory at the calibrated $\varrho_X(K) = 0.31\, a^{-2K}$; squares: NS-$K$ (frozen bank). Grey: SIGNSGD at the same batch B.

to that at $\beta = 0$, where the rank-B gradient has no weak directions, NS-5 and exact whitening nearly coincide, and the learning rate is nearly flat in B for $B < N$. For $B_{\text{eff}} \geq N$ the ridge matters little, and every curve rises with slope between $\frac{1}{2}$ and 0.6, as in Sec. 3.1. NS-K stays above SIGNSGD at the same batch in every configuration.

Amplitude. The calibration matches the scale-free statistic $I_1 = \mathfrak{d}^2/\mathfrak{v}$, while the learning rate $2\mathfrak{d}/\mathfrak{v}$ also carries the overall amplitude of the update. For $K \geq 3$ the NS-K learning rate is between 0.92 and 1.16 times the ridge theory at $\varrho_X(K)$ over all 54 frozen configurations at $N = 512$. NS-1 is nearly linear, $S_t \approx a\, M_t/\|M_t\|_{\text{F}}$, a smaller update than the ridge map at $\varrho_X(1)$, and its learning rate is 1.15–1.64 times higher. The exact-polar points at the smallest B_{eff}/N for $\beta \in \{0.1, 0.25\}$ lie above the zero-ridge curve, for the rank-threshold reason of Sec. C.1.2.

Scope. These measurements are in the isotropic model with the standard Muon coefficients. The calibration concerns the rate at the optimal step and the lag share; it does not claim that NS-K and the ridge have the same drift or volatility separately, since they differ by an overall amplitude.



(a) Isotropic ($\alpha = 0, \varsigma = 0$).

(b) Half-anisotropic, $\alpha = 1.5, \varsigma = 0.7$.

(c) Half-anisotropic, $\alpha = 1.5, \varsigma = 1.5$.

(d) Half-anisotropic, $\alpha = 1.5, \varsigma = 3.0$.

Figure 16: MUON without momentum (Newton–Schulz, dashed) against the deterministic time-to-ϵ model of Sec. 5 (solid), $N = 1024$, $B = 2048$, 16 trials per target, colour-coded by $\epsilon \in \{2^{-4}, \dots, 2^{-14}\}$. Circles: model $t_{2\epsilon}$; squares: simulated $t_{2\epsilon}$. The simulations follow the model's ϵ-exponent in every regime.

D Momentum in MUON: theory against an ablation

This appendix compares the momentum theory of the appendix with simulations of MUON at $\beta \in \{0, 0.9, 0.95, 0.99\}$ on the half-anisotropic power-law problem at $B = 2N$. The theory makes two predictions in this oversampled regime. The noise floor changes with β through an explicit noise constant $\mathcal{S}(\beta)$, which we compute and which matches the measurements to within 1–6% with momentum (Sec. D.2). The ϵ-exponent of the time to reach a target does not change with β; the time constants with momentum are outside the theory at these settings, because the buffer is dominated by the signal early in the descent (Sec. D.3).

Algorithms. Both updates are run on the half-anisotropic problem of Sec. 4, with $\Sigma_{\text{in}} = \text{Id}$ and $\Sigma_{\text{out}} = U \operatorname{diag}(\mu) U^\top$ for a Haar-random orthogonal U and $\mu_i = i^{-\alpha}$. Without momentum the update is $\Delta_{t+1} = \Delta_t - \eta \operatorname{orthog}(G_t)$, where orthog is five Newton–Schulz steps with the coefficients of [34]. With momentum a heavy-ball buffer $H_{t+1} = \beta H_t + G_t$ is orthogonalized instead, as in `optax.contrib.scale_by_muon` with `nesterov=False` [22]. The heavy-ball buffer is the normalized average M_t of definition F.1 divided by $1 - \beta$ (remark SA.2); the polar factor, and Newton–Schulz applied to a Frobenius-normalized input, do not see this factor, so the simulated algorithm is the algorithm of the appendix at zero ridge, up to the Newton–Schulz approximation.

Setup. We use the matched-floor protocol of Sec. 5: for each target ϵ the step is $\eta^\star = 2\sqrt{\epsilon}/\mathcal{S}$, which puts the risk floor at ϵ. In the simulations $\mathcal{S} = \mathcal{S}_{\text{eff}}$ is calibrated per (β, ς) from a short probe at $\epsilon = 0.25$, as $2\sqrt{\mathscr{R}_\infty}/\eta$ with $\mathscr{R}_\infty$ the measured plateau. The problems are $\alpha = 1.5$ with $\varsigma \in \{0.7, 1.5, 3.0\}$, one in each ϵ-exponent regime of Thm. 5.1; $N = 1024$, $B = 2048$, 16 trials, and $\epsilon \in \{2^{-4}, 2^{-6}, 2^{-8}, 2^{-10}\}$. SIGNSGD at $\beta = 0$ is run for reference. Fig. 16 first compares MUON at $\beta = 0$ with the deterministic model of Sec. 5 over a wider range of ϵ.

	$\beta = 0$	$\beta = 0.9$	$\beta = 0.95$	$\beta = 0.99$
theory $\mathcal{S}(\beta)$	11.99	39.10	54.79	116.56
measured $\mathcal{S}_{\text{eff}}$, $\varsigma = 0.7$	9.97	36.89	52.48	117.39
measured $\mathcal{S}_{\text{eff}}$, $\varsigma = 1.5$	9.97	36.91	52.49	117.38
measured $\mathcal{S}_{\text{eff}}$, $\varsigma = 3.0$	9.97	36.89	52.49	117.22
measured / theory	0.83	0.94	0.96	1.01

Table 3: Noise constant at $\alpha = 1.5$, $N = 1024$, $B = 2048$. Theory: exact polar factor, zero ridge, coefficients of Sec. D.1. Measured: MUON with five Newton–Schulz steps, calibrated $\mathcal{S}_{\text{eff}}$; the three target exponents agree to four significant figures at every β.

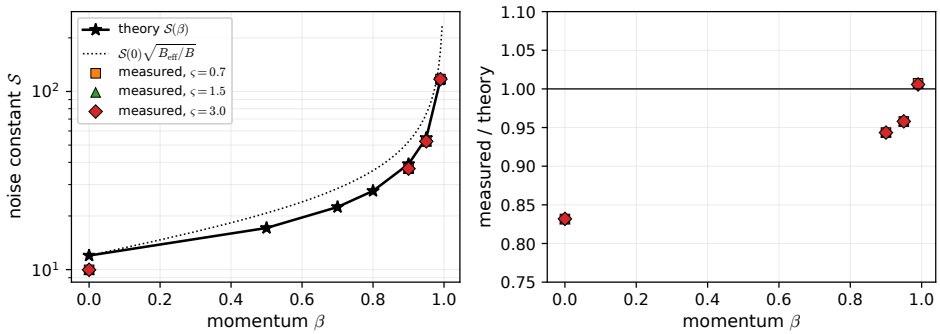


Figure 17: Noise constant against momentum. Left: theory $\mathcal{S}(\beta)$ (stars, line), the plain rescaling $\mathcal{S}(0)\sqrt{B_{\text{eff}}/B}$ (dotted), and the measured $\mathcal{S}_{\text{eff}}$ for the three target exponents, which coincide. Right: measured over theory.

D.1 What the theory predicts at $B = 2N$

Drift. At $B = 2N$ the small parameter of theorem SH.1 is $\epsilon_{\text{os}} = 3(1-\beta^2)N/B = 1.5(1-\beta^2)$, which equals 0.29, 0.15 and 0.03 at $\beta = 0.9, 0.95, 0.99$. In this window the theorem gives

$$\mathfrak{d}_i = \vartheta_i \sqrt{\frac{B_{\text{eff}}}{N}}\, \mathfrak{d}^{\text{os}}(\mu_i), \qquad \vartheta_i \in \left[1-\epsilon_{\text{os}}, (1-\epsilon_{\text{os}})^{-1}\right],$$

with a profile $\mathfrak{d}^{\text{os}}$ that does not depend on β: momentum enters the drift only through the effective batch $B_{\text{eff}} = B(1+\beta)/(1-\beta)$. We evaluate the zero-ridge drift kernel of each row exactly (eq. (SC.4), single integral over the branch). Relative to $\sqrt{B_{\text{eff}}/B}$ times the drift at $\beta = 0$, the ratio lies in $[1.00, 1.19]$, $[1.00, 1.14]$ and $[0.95, 1.02]$ over all rows at $\beta = 0.9, 0.95, 0.99$.

Volatility. The zero-ridge row volatility is explicit at a finite spectrum: it is 1 plus twice the sum over lags of double integrals of the two-time row kernel (propositions SG.3 and SG.4), and without momentum it equals 1, since the polar factor of a full-rank square gradient is orthogonal. With momentum the head rows have $\mathfrak{v}_1 = 19.0, 38.9, 198.2$, essentially the factor $B_{\text{eff}}/B = 19, 39, 199$ of a linear buffer, while the tail rows have less, $\mathfrak{v}_{1000} = 9.2, 16.7, 56.7$. Our code reproduces the isotropic constants of Fig. 4 to 0.2% when all eigenvalues are equal.

Noise constant. The noise constant of the matched-floor protocol is (eq. (16))

$$\mathcal{S}(\beta) = \sum_{i=1}^{N} \frac{\mu_i\, \mathfrak{v}_i}{2\, \mathfrak{d}_i},$$

and it does not depend on the target exponent ς, because neither coefficient does. The theory values are in Table 3.

D.2 The noise floor

The theory predicts the noise floor with momentum (Table 3, Fig. 17). For $\beta \in \{0.9, 0.95, 0.99\}$ the measured noise constants are within 6%, 4% and 1% of $\mathcal{S}(\beta)$, and, as predicted, they do not depend

Figure 18: $\alpha = 1.5$, $\varsigma = 0.7$, $N = 1024$, $B = 2048$, 16 trials. Left: risk at the matched step for each target ϵ (colour) and β (linestyle); markers at the measured $t_{2\epsilon}$. Right: $t_{2\epsilon}$ against ϵ. Blue dashed: the row recursion with the theory coefficients of Sec. D.1 at $\beta = 0$, at its own matched step (no line is drawn for $\beta > 0$; see Sec. D.3). Grey dashed: the same recursion with the SIGNSGD kernels of Table 1. The asymptotic exponents are 0.625 (SIGNSVD) and 1.25 (SIGNSGD).

Figure 19: $\alpha = 1.5$, $\varsigma = 1.5$; conventions as in Fig. 18. Asymptotic exponents 0.5 (SIGNSVD) and 0.75 (SIGNSGD).

on ς. The floor at a given step is $(\eta\,\mathcal{S}(\beta)/2)^2$, so the noise-floor adjustment for momentum is the ratio $\mathcal{S}(\beta)/\mathcal{S}(0)$, equal to 3.26, 4.57 and 9.72. It is smaller than the plain rescaling $\sqrt{B_{\text{eff}}/B} = 4.36$, 6.24, 14.1 by the factors 0.75, 0.73 and 0.69: the drift grows by $\sqrt{B_{\text{eff}}/B}$ on every row, but the volatility of the tail rows grows by less than B_{eff}/B, and the tail rows dominate $\mathcal{S}$.

Without momentum the measured constant is 17% below the theory. Here the gradient is a square matrix with aspect ratio $N/B = 1/2$ and spread-out singular values, and five Newton–Schulz steps differ from the exact polar factor by the spread of their outputs, the effect identified in Appendix C at $\beta = 0$. With momentum the buffer averages many batches, its singular values are concentrated, and the difference is small. The empirical prefactor $\hat{\rho} \approx 0.84$ of an earlier version of this appendix, defined relative to the measured $\beta = 0$ value, combines this 17% with the factors 0.69–0.75 above.

D.3 Time to ϵ

The exponent. The ϵ-exponent of the time to reach a target is set by the power-law orders in i of $\mathfrak{d}_i$ and $\mathfrak{v}_i$ (Thm. 5.1). Momentum multiplies the drift of every row by $\sqrt{B_{\text{eff}}/B}$ up to a bounded factor, and the volatility by factors bounded above and below uniformly in i (between 9.1 and 19 at $\beta = 0.9$,

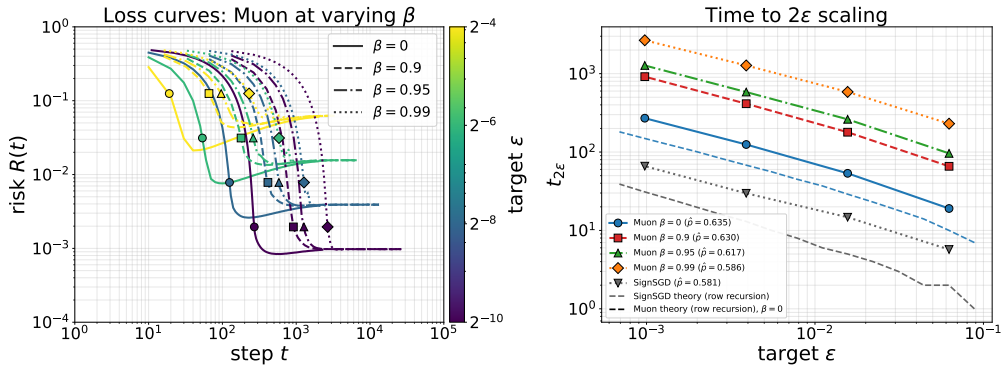


Figure 20: $\alpha = 1.5$, $\varsigma = 3.0$; conventions as in Fig. 18. Both algorithms are saturated, with asymptotic exponent 0.5.

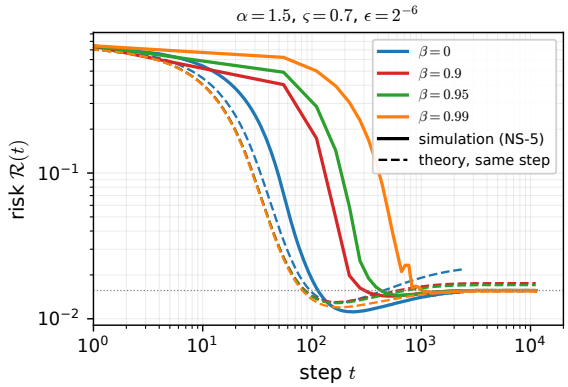


Figure 21: Risk at $\alpha = 1.5$, $\varsigma = 0.7$, $\epsilon = 2^{-6}$: simulations (solid) against the row recursion with the theory coefficients at the simulations' own step (dashed). With momentum the simulated descent is slower than the theory, increasingly with β; the floors agree to the accuracy of Table 3.

between 55 and 198 at $\beta = 0.99$), so it does not change these orders, and the exponent is the same for every β. The measured slopes (Figs. 18–20, right panels) agree: they lie within 10–30% of the SIGNSVD exponent for every β, do not depend systematically on β, and show the same finite-N deviation already present at $\beta = 0$.

The time constants. Without momentum the theory follows the simulations (blue dashed lines in the right panels: the row recursion with the coefficients of Sec. D.1 at its own matched step): within 0.8–1.06 at $\varsigma = 0.7$ and 0.7–0.84 at $\varsigma = 1.5$, and about 0.55 at $\varsigma = 3.0$, where the gap includes the Newton–Schulz offset of Sec. D.2. With momentum the theory predicts a curve close to the $\beta = 0$ one, while the simulations are slower by large factors (2–8 at $\beta = 0.9$, 5–30 at $\beta = 0.99$), so the panels show no theory line for $\beta > 0$. Fig. 21 shows the gap on the risk itself.

Why the momentum transient is outside the theory. The gain of the drift by $\sqrt{B_{\text{eff}}/B}$ (Sec. D.1) is a linear response: the polar factor of a buffer that is mostly noise responds linearly to the signal in it, and averaging more batches raises the signal-to-noise ratio. It requires the noise to dominate the buffer (remark SH.1, which asks for $B_{\text{eff}} \ll N^2$ for a spread-out error). Here $B_{\text{eff}}/N^2 = 0.04$ at $\beta = 0.9$ and 0.39 at $\beta = 0.99$, and the initial error is concentrated: at $\varsigma = 0.7$ the top mode carries about two thirds of the weighted risk. Early in the descent the buffer is therefore close to its signal $\Sigma_{\text{out}}\Delta_t$, whose polar factor has bounded entries, and the extra signal-to-noise ratio buys no extra

	theory	fresh buffer		momentum buffer		MUON run
		polar	NS-5	polar	NS-5	
$\beta = 0$	58	68	71	—	—	76
$\beta = 0.9$	51	120	160	141	179	220
$\beta = 0.99$	50	357	482	1039	787	605

Table 4: Steps to 2ϵ at $\alpha = 1.5$, $\varsigma = 0.7$, $\epsilon = 2^{-6}$, with the step and initial error of the ablation. Gaussian gradient model $G_t = \Sigma_{\text{out}}\Delta_t + \sqrt{2\mathscr{R}(t)/B}\,{\Sigma_{\text{out}}}^{1/2}W_t$ in the eigenbasis; the fresh buffer has the noise level of the momentum buffer but is formed at the current state. One run per entry.

speed. Near the floor the error is small and spread out, the noise dominates again, and the linear response holds; this is why the noise floor and the exponent are predicted.

A controlled comparison separates this effect from two others (Table 4). A Gaussian model of the gradient reproduces the simulations to within about 30%. In it, a fresh buffer at the current state with the exact polar factor, which has neither a stale buffer nor Newton–Schulz, is already 2.4 and 7 times slower than the theory at $\beta = 0.9$ and 0.99; this is the loss of linear response. Replacing the exact polar factor by five Newton–Schulz steps slows the descent by a further 1.3–1.35: the effective ridge $\varrho_X\|M_t\|_{\text{F}}^2$ of Appendix C grows with the signal in $\|M_t\|_{\text{F}}$ and suppresses the noise directions through which the tail rows descend. Using the actual momentum buffer, whose signal lags the state, slows it by 1.1–1.2 at $\beta = 0.9$ and by 1.6–2.9 at $\beta = 0.99$; with this stale buffer Newton–Schulz is faster than the exact polar factor, because its ridge down-weights the old directions, as in Appendix C.

Takeaway. At $B = 2N$ the theory predicts how momentum changes the noise floor of MUON, through the explicit noise constant $\mathcal{S}(\beta)$, to within a few percent, and it predicts that the ϵ-exponent of the time to target does not change, as observed. The transient with momentum is outside the theory at these settings: the buffer is signal-dominated early in the descent, so the linear-response gain of the drift is not realized, and the stale buffer and the Newton–Schulz ridge add smaller slowdowns. At matched floor, momentum therefore costs time in this ablation.

E Roadmap of the results and their proofs

The preceding sections give additional related work, numerical methods, finite-ridge experiments and momentum ablations. The sections that follow present the results and their proofs. This proof appendix is written for a reader who has finished the main text. It restates, verbatim, every appendix result that the main text cites, and follows each with a paragraph on what the result says and which claim of the main text it supports, and, for theorems, propositions, lemmas and corollaries, a sketch of the proof. The line printed after each restated result gives its number in the supplement, which contains the complete appendix with full proofs, in sections SA to SJ; references such as Lemma SE.4 point into the supplement. The statements of the supplement are formalized in Lean, except the remarks and the numerical statements. The standard Gaussian results they use are proved in the accompanying foundations library, listed in appendix SK of the supplement; the line after each restated result names the corresponding Lean declarations and their status.

The proof appendix has three parts. Appendix F fixes the model, the algorithm and the proportional regime, which are standing assumptions for everything that follows. Appendix G gives an informal overview of six proof strategies and how they fit together: the deterministic equivalent for the Gram resolvent of the momentum buffer (appendix G.1), the expected regularized update (appendix G.2), the volatility and the lag kernel (appendix G.3), the passage from the algorithm to the deterministic dynamics (appendix G.4), the power-law and large-batch asymptotics (appendix G.5), and the Volterra analysis behind the phase map (appendix G.6). The seven remaining chapters state the precise results and their formalization status, each opening with the setting its results share.

The first four strategies form a chain, and together they carry the central claim of the main text, theorem 2.1. Appendix G.1 computes the resolvent of the Gram matrix of the momentum buffer at a fixed state through the fixed-point system of fig. 2. Appendix G.2 integrates that resolvent into the mean of the regularized update, which gives the drift kernel eq. (7), and appendix G.3 computes the covariance of consecutive updates, which gives the volatility kernel. Appendix G.4 compares the algorithm with the deterministic recursion eq. (6) over blocks of steps, using the two kernels and the fact that a diffuse error stays diffuse. Appendix H restates this comparison and its half-anisotropic form, which is the row recursion eq. (14); it rests on all four strategies. It comes first because the later chapters evaluate its coefficients: appendix I restates the fixed-point system and the drift kernel (appendices G.1 and G.2), and appendix J reads the volatility kernel for the half-anisotropic rows and without momentum (appendices G.1 to G.3).

The next three chapters turn the kernels into the quantitative claims of the main text. Appendix K treats identity covariances, where the fixed-point system becomes one scalar equation and the dynamics closes on the risk; it supplies the drift and volatility of section 3, the learning-rate ratio and the speedup, at large and at small batch. It uses the first four strategies and, for its limits in the aspect ratio, appendix G.5. Appendix L computes the half-anisotropic coefficients at zero ridge, for power-law spectra at thin batch and for every spectrum at oversampled batch, which gives table 1 and the square-root law of theorem 4.1; it is appendix G.5 applied to the formulas of appendices G.1 to G.3. Appendix M takes the kernels as fixed coefficients of the row recursion and derives the times to accuracy, their ratio and the phase diagram of section 5; it rests on appendix G.6, and its isotropic comparison also on appendix G.1.

The last chapter, appendix N, concerns the comparator. Its results are one-step Gaussian computations for the entrywise sign and use none of these six strategies. They give the SIGNSGD kernels of sections 3 and 4, on which the SIGNSGD arrays of appendix M are modelled.

F Model, algorithm and standing assumptions

The model is the matrix linear regression of section 1.1: a parameter W of size $N_{\text{out}} \times N_{\text{in}}$, a teacher $W^\star$, independent Gaussian features with covariances $\Sigma_{\text{out}} = \sum_i \mu_i \, u_i \otimes u_i$ and $\Sigma_{\text{in}} = \sum_j \lambda_j \, v_j \otimes v_j$, and the risk $\mathscr{R}(W) = \frac{1}{2}\|{\Sigma_{\text{out}}}^{1/2}(W - W^\star){\Sigma_{\text{in}}}^{1/2}\|_{\text{F}}^2 = \frac{1}{2}\sum_{i,j} \mu_i \lambda_j \, \mathscr{Q}_{ij}$. The minibatch gradient eq. (2) at the error $\Delta_t = W_t - W^\star$ is written G°_{t+1}; the superscript $^\circ$ marks physical objects, as opposed to the normalized ones divided by $\sqrt{2\mathscr{R}(t)}$. For $\varrho > 0$ the regularized polar map is $Q_\varrho(K) = (KK^\top + \varrho \, \mathrm{Id})^{-1/2}K$ (eq. (SA.12)). The three statements below are standing for the whole appendix: the first fixes the law of the data, the second the algorithm whose risk the dynamics

tracks, and the third the limit in which every asymptotic statement is made. The random-matrix results add one condition on the state, that the error be spectrally diffuse, which the dynamical theorem shows persists, from a diffuse initial state, over the time window it covers.

Assumption 2 (Data-target normalization)**.** *The input and output features are independent and Gaussian,* $x_{\mathrm{in}} \sim \mathcal{N}(0, \Sigma_{\mathrm{in}})$ *and* $x_{\mathrm{out}} \sim \mathcal{N}(0, \Sigma_{\mathrm{out}})$*, with positive semidefinite covariances* $\Sigma_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}} \times N_{\mathrm{in}}}$ *and* $\Sigma_{\mathrm{out}} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{out}}}$*. They are represented in whitened form as* $x_{\mathrm{in}} = {\Sigma_{\mathrm{in}}}^{1/2} w$ *and* $x_{\mathrm{out}} = {\Sigma_{\mathrm{out}}}^{1/2} z$ *with* $w \sim \mathcal{N}(0, \mathrm{Id}_{N_{\mathrm{in}}})$ *and* $z \sim \mathcal{N}(0, \mathrm{Id}_{N_{\mathrm{out}}})$ *independent; the whitened columns of a batch are written* $(w_a, z_a)_{a \le B}$*.*

Full statement and proof: Assumption S1 of the supplement. Lean: proved: `Muon.Setup.assum_data_target`*,* `Muon.Setup.dataTarget`*, +2 more.*

This is the data model of section 1.1. In the whitened form the residual of a sample is $d = z^\top A_t w$ with the rotated error $A_t = {\Sigma_{\mathrm{out}}}^{1/2} \Delta_t {\Sigma_{\mathrm{in}}}^{1/2}$, and $\mathscr{R}(t) = \frac{1}{2}\|A_t\|_{\mathrm{F}}^2$ (eq. (SA.9)). Every random-matrix statement uses the normalized residual $d/\sqrt{2\mathscr{R}(t)} = z^\top \widehat{A}_t w$, which has unit variance and is close to a standard Gaussian when the normalized error $\widehat{A}_t$ has small operator norm (lemma SA.7). Gaussianity is used for these computations (Wick pairings and Stein's identity); the exact one-step risk recursion needs only second moments.

Definition F.1 (The algorithm)**.** *Fix a momentum* $\beta \in [0, 1)$*, a normalized ridge* $\varrho > 0$*, a deterministic step sequence* $(\eta_t)_{t \ge 0}$ *with* $\eta_t \ge 0$*, an initial parameter* W_0 *and an initial buffer* $M^\circ_{-1} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}}$*, both given before the first batch is drawn (the default is* $M^\circ_{-1} = 0$*). Let* $t_{\mathrm{stop}} \overset{\mathrm{def}}{=} \inf\{t \ge 0 : \mathscr{R}(t) = 0\} \in \{0, 1, \dots\} \cup \{\infty\}$*. The algorithm is, for* $t \ge 0$*,*

$$\begin{aligned} M^\circ_t &= \beta\, M^\circ_{t-1} + (1-\beta)\, G^\circ_{t+1}, \qquad W_{t+1} = W_t - \eta_t\, \widetilde{G}^\circ_{t+1}, \\ \widetilde{G}^\circ_{t+1} &\overset{\mathrm{def}}{=} \begin{cases} Q_{\varrho^\circ_t}(M^\circ_t), & t < t_{\mathrm{stop}}, \\ 0, & t \ge t_{\mathrm{stop}}, \end{cases} \qquad \varrho^\circ_t \overset{\mathrm{def}}{=} 2\mathscr{R}(t)\, \varrho. \end{aligned} \tag{19}$$

Full statement and proof: definition SA.2 of the supplement. Lean: proved: `Muon.Setup.def_app_setup_algorithm`*,* `Muon.Setup.AlgParams`*, +17 more.*

This is MUON $_\varphi$ of section 1.2 with the regularized polar map Q_ϱ, applied at the physical ridge $2\mathscr{R}(t)\varrho$. Since $Q_{2\mathscr{R}\varrho}(\sqrt{2\mathscr{R}}\, K) = Q_\varrho(K)$ (proposition SA.3), the update is the regularized polar factor, at the fixed normalized ridge ϱ, of the buffer divided by $\sqrt{2\mathscr{R}(t)}$; this is why the laws computed later do not depend on the risk. The stopping time only matters with momentum, where the buffer can be nonzero at a state of zero risk. A heavy-ball buffer at ridge ϱ' is the same algorithm at ridge $(1-\beta)^2\varrho'$ (remark SA.2), and no statement concerns $\varrho \downarrow 0$ or Newton–Schulz iterations.

Assumption 3 (Proportional regime)**.** *The dimensions* N_{in}*,* N_{out}*,* B *tend to infinity jointly, with aspect ratios converging to constants:*

$$\frac{N_{\mathrm{in}}}{B} \to \gamma_{\mathrm{in}} \in (0, \infty), \qquad \frac{N_{\mathrm{out}}}{B} \to \gamma_{\mathrm{out}} \in (0, \infty).$$

The covariance operator norms remain uniformly bounded: there exists a constant $C > 0$*, independent of* N_{in} *and* N_{out}*, such that*

$$\|\Sigma_{\mathrm{in}}\|_{\mathrm{op}} \;\le\; C, \qquad \|\Sigma_{\mathrm{out}}\|_{\mathrm{op}} \;\le\; C.$$

We refer to the joint limit as the proportional regime*.*

Full statement and proof: Assumption S2 of the supplement. Lean: proved: `Muon.Setup.assum_proportional_regime`*,* `Muon.ProportionalRegime`*.*

This is the proportional limit of theorem 2.1; the main text's $N_{\mathrm{in}} = N_{\mathrm{out}} = N$ is the case $\gamma_{\mathrm{in}} = \gamma_{\mathrm{out}}$. No lower bound on the covariances and no rank condition is assumed. The random-matrix results of the supplement are proved under the stronger Assumption S3, for Gaussian data in the frozen-state model of appendix G.1: explicit bounds $\max(N_{\mathrm{in}}, N_{\mathrm{out}}) \le \gamma_\star B$ and $\|\Sigma_{\mathrm{in}}\|, \|\Sigma_{\mathrm{out}}\| \le C_\Sigma$ along the sequence, a normalized ridge in a fixed compact subset of $(0, \infty)$, and a spectrally diffuse normalized error, $\|\widehat{A}_t\| \to 0$ with $\|\widehat{A}_t\|_{\mathrm{F}} = 1$.

G Proof strategies

This section gives an informal overview of the proof strategies and how they fit together. The arguments are grouped by the methods they share: the resolvent calculation leads to the drift and volatility kernels, which enter the comparison with deterministic dynamics; asymptotic analysis of these kernels and a Volterra equation then give the predictions for learning rates and times to accuracy. The following sections are organized by result. They state the precise hypotheses and conclusions, identify the corresponding Lean formalizations and their status, and give a proof sketch for each result. The strategies described here explain the connections between those proofs; the supplement contains the complete proofs.

G.1 Deterministic equivalent for the EMA Gram resolvent

Every kernel of the spectral family is a spectral integral against the resolvent of the Gram matrix of the momentum buffer, and this engine computes that resolvent (appendix SB). It works in the frozen-state model of definition SA.3, in which the state is held fixed and every gradient in the buffer is recomputed at it with its own independent batch. The normalized buffer is then

$$M_t = \sum_{j\geq 0} w_j\, G_{t+1-j}, \qquad w_j = (1-\beta)\beta^j, \qquad G_{t+1-j} = \frac{1}{B}\sum_{a=1}^{B} d_{ja}\, {\Sigma_{\mathrm{out}}}^{1/2}(z_{ja}\otimes w_{ja}){\Sigma_{\mathrm{in}}}^{1/2},$$

with independent standard Gaussian vectors z_{ja}, w_{ja} and residuals $d_{ja} = z_{ja}^\top \widehat{A}_t w_{ja}$, where $\widehat{A}_t = {\Sigma_{\mathrm{out}}}^{1/2}\Delta_t {\Sigma_{\mathrm{in}}}^{1/2}/\sqrt{2\mathscr{R}(t)}$ is the normalized error, of Frobenius norm one. The engine assumes the proportional regime and that the error is diffuse, $\|\widehat{A}_t\| \to 0$: no singular direction of the error carries a fixed fraction of the risk. Its output is that, uniformly for s in compact subsets of $(0,\infty)$ and in operator norm (theorem SB.2),

$$\mathbb{E}\left(M_t M_t^\top + s\,\mathrm{Id}\right)^{-1} \longrightarrow \overline{R}(-s) = \left(s\,\mathrm{Id} + \tilde{s}_1(-s)\,\Sigma_{\mathrm{out}}\right)^{-1},$$

with the same statement for $M_t^\top M_t$ and the limit $\overline{\widetilde{R}}(-s) = \left(s\,\mathrm{Id} + s_1(-s)\Sigma_{\mathrm{in}}\right)^{-1}$, where $\tilde{s}_1, s_1$ solve the weighted fixed-point system of definition I.1; normalized traces against bounded test matrices converge in mean square (corollary SB.1).

The resolvent equation is first written exactly through one deleted sample. The buffer is a sum of rank-one terms, one per sample, $M_t = \sum_{j,a} \mathsf{h}_{ja}\,\mathsf{u}_{ja}\otimes\mathsf{v}_{ja}$ with $\mathsf{h}_{ja} = w_j d_{ja}$, $\mathsf{u}_{ja} = {\Sigma_{\mathrm{out}}}^{1/2} z_{ja}/\sqrt{B}$ and $\mathsf{v}_{ja} = {\Sigma_{\mathrm{in}}}^{1/2} w_{ja}/\sqrt{B}$, so the identity $\mathrm{Id} - sR = RM_tM_t^\top$ for $R = R(-s)$ is a sum over samples of $(\mathsf{h}RM_t\mathsf{v})\otimes\mathsf{u}$. Removing one sample changes M_t by a rank-one matrix, and the Woodbury formula for a 2×2 block computes the change exactly (lemma SB.4). With M_0 the buffer without the sample, R_0 its resolvent and $\mathsf{X}_0 = R_0 M_0$,

$$\mathsf{h}\,RM_t\mathsf{v} = -\alpha_3\, R_0\mathsf{u} + \alpha_1\,\mathsf{X}_0\mathsf{v}, \qquad \alpha_1 = \frac{\mathsf{h}(1+\mathsf{h}\mathsf{C})}{\Lambda}, \quad \alpha_3 = -\frac{s\mathsf{h}^2\mathsf{V}}{\Lambda}, \quad \Lambda = (1+\mathsf{h}\mathsf{C})^2 + s\mathsf{h}^2\mathsf{U}\mathsf{V},$$

where $\mathsf{U} = \mathsf{u}^\top R_0\mathsf{u}$, $\mathsf{V} = \mathsf{v}^\top\widetilde{R}_0\mathsf{v}$ and $\mathsf{C} = \mathsf{u}^\top\mathsf{X}_0\mathsf{v}$. The deleted resolvents do not depend on the deleted sample, so after taking expectations each term is a Gaussian computation in one sample.

The coefficients then become functions of the residual of the deleted sample alone. The scalars U and V are quadratic forms of a Gaussian vector independent of R_0, so they concentrate, with errors $O(B^{-1/2})$, at the expected normalized traces $\sigma_B = \frac{1}{B}\operatorname{Tr}(\Sigma_{\mathrm{out}}\,\mathbb{E}\,R)$ and $\tilde{\sigma}_B$, and C concentrates at 0 (lemma SB.11); the traces themselves fluctuate by $O(B^{-1/2})$, by the Efron–Stein inequality (lemma SB.10). For a sample of age j this gives, in mean square (lemma SB.12),

$$\alpha_1 \approx \frac{w_j\, d}{1 + w_j^2 q_B\, d^2}, \qquad \alpha_3 \approx -\frac{s\, w_j^2\, \tilde{\sigma}_B\, d^2}{1 + w_j^2 q_B\, d^2}, \qquad q_B = s\,\sigma_B\tilde{\sigma}_B,$$

and the age of a sample enters only through its squared weight in the denominator.

Gaussian integration produces the weight function, and diffuseness removes the cross term. In the term carrying α_3 one needs $\mathbb{E}[\psi(d)\, z\otimes z]$ with $\psi(d) = d^2/(1+xd^2)$ and $x = w_j^2 q_B$. Stein's identity gives the exact formula

$$\mathbb{E}\left[(z\otimes z - \mathrm{Id})\psi(d)\right] = \widehat{A}_t\ \mathbb{E}\left[(w\otimes w)\psi''(d)\right]\widehat{A}_t^\top,$$

whose Frobenius norm is at most $6\|\widehat{A}_t\|$ (lemma SB.8). The residual d is asymptotically standard Gaussian because $\|\widehat{A}_t\| \to 0$ (lemma SA.7), so $\mathbb{E}\,\psi(d)$ tends to $J(x)$. In the term carrying α_1 the response is $\mathbb{E}[f(d)\,w \otimes z] = \widehat{A}_t^\top\,\mathbb{E}\,f'(d) + O(\|\widehat{A}_t\|^2)$, of Frobenius norm of order one but of operator norm at most $\|\widehat{A}_t\|$, which vanishes. Summing over ages gives, in operator norm (eq. (SB.44)),

$$\mathrm{Id} - s\;\mathbb{E}\,R - a_B\;\mathbb{E}\,R\,\Sigma_{\mathrm{out}} \longrightarrow 0, \qquad a_B = s\,\tilde{\sigma}_B\,F_\beta(q_B), \qquad F_\beta(q) = \sum_{j\geq 0} w_j^2\,J(w_j^2 q),$$

that is, $\mathbb{E}\,R \approx (s\,\mathrm{Id} + a_B\Sigma_{\mathrm{out}})^{-1}$. Taking the trace against Σ_{out} closes the loop: $(\sigma_B, \tilde{\sigma}_B, a_B, c_B)$, with c_B the input-side analogue of a_B, solves the weighted system up to vanishing residuals.

An approximate root is close to the root. The weighted system has exactly one nonnegative solution, and it is stable under approximate roots uniformly over spectral measures of bounded mass and support (theorem I.1(iii)). Hence $a_B \to \tilde{s}_1$, and the resolvent identity gives $\mathbb{E}\,R \to \overline{R}$. The Gaussian limit of d and the replacement of the coefficients are qualitative, no rate is assumed for $\|\widehat{A}_t\| \to 0$, and the stability is proved by compactness, so the engine gives no rate. The system depends on the whole weight sequence: with $s_p = \sum_j w_j^p$, the EMA weights with $\beta > 0$ have $s_4/s_2^3 = (1+\beta)^2/(1+\beta^2) > 1$, so no single effective batch size B/s_2 reproduces F_β beyond its value at $q = 0$ (remark SB.2).

G.2 Drift: the expected regularized update

This engine computes the mean of the update in the frozen-state model (appendix SC), which is the input of every drift kernel. It works in the setting of appendix G.1 and uses its output. It produces a deterministic matrix $\overline{\widetilde{G}}_\varrho$, whose entries in the eigenbases of the covariances are multiples of the corresponding entries of Δ_t, with, for normalized ridges in a fixed interval $[\varrho_-, \varrho_+] \subset (0, \infty)$ (theorem SC.2),

$$\sup_{\varrho \in [\varrho_-, \varrho_+]} \|\,\mathbb{E}\,Q_\varrho(M_t) - \overline{\widetilde{G}}_\varrho\|_{\mathrm{F}} \longrightarrow 0.$$

The norm is the unnormalized Frobenius norm, the norm in which eq. (26) controls the drift errors of all modes together.

The update is an integral of the mixed resolvent along the negative axis. For a singular value σ_k of M_t, the scalar integral $\frac{2}{\pi}\int_0^\infty \sigma_k(\sigma_k^2 + \varrho + u^2)^{-1}du$ equals $\sigma_k(\sigma_k^2+\varrho)^{-1/2}$, the filter of the regularized polar map, so (lemma SA.3)

$$\mathbb{E}\,Q_\varrho(M_t) = \frac{2}{\pi}\int_0^\infty \mathcal{H}(\varrho + u^2)\,du, \qquad \mathcal{H}(s) \overset{\mathrm{def}}{=} \mathbb{E}\left[(M_t M_t^\top + s\,\mathrm{Id})^{-1} M_t\right].$$

It suffices to find a deterministic equivalent of $\mathcal{H}(s)$ in Frobenius norm, uniformly for s in compact sets, and a bound that controls the integral at large s.

The mixed resolvent solves a linear response equation. Deleting one sample as in appendix G.1, now through the identity $\mathsf{h}R\mathsf{u} = \alpha_1 R_0\mathsf{u} - \alpha_2\mathsf{X}_0\mathsf{v}$ with $\alpha_2 = \mathsf{h}^2\mathsf{U}/\Lambda$, writes $\mathcal{H}$ exactly as a sum over ages of a source term carrying α_1 and a feedback term carrying α_2 (lemma SC.6). In the source term the Gaussian response of a sample with residual function f is

$$\mathbb{E}\left[f(d)\,z \otimes w\right] = \widehat{A}_t\;\mathbb{E}\,f'(d) + O\big(\|\widehat{A}_t\|^2\big),$$

and multiplying by ${\Sigma_{\mathrm{out}}}^{1/2}$ and ${\Sigma_{\mathrm{in}}}^{1/2}$ turns $\widehat{A}_t$ into the mean gradient

$$\mathbb{E}\,G_{t+1} = \Sigma_{\mathrm{out}}\Delta_t\Sigma_{\mathrm{in}}/\sqrt{2\mathscr{R}(t)}.$$

By Stein's identity $\mathbb{E}\,f'(d) \to J(w_j^2 q)$, so the weights now enter to the first power, through L_β. In the feedback term α_2 is replaced by its conditional mean, whose expected sum over ages, $b_B(s)$, tends to $\sigma F_\beta(q)$. The result is the response equation (proposition SC.3)

$$\mathcal{H}(s)\big(\mathrm{Id} + b_B(s)\,\Sigma_{\mathrm{in}}\big) = L_\beta(q)\,\overline{R}(-s)\;\mathbb{E}\,G_{t+1} + o(1) \quad \text{in Frobenius norm.}$$

Since $s(\mathrm{Id} + \sigma F_\beta(q)\Sigma_{\mathrm{in}}) = s\,\mathrm{Id} + s_1\Sigma_{\mathrm{in}} = \overline{\widetilde{R}}(-s)^{-1}$ by the fixed-point system, the candidate below solves the same equation exactly with b_B replaced by its limit, and comparing the two equations shows that $\mathcal{H}(s)$ is close to it (theorem SC.1(i)):

$$T(s) = s\,L_\beta\big(q(-s)\big)\,\overline{R}(-s)\;\mathbb{E}\,G_{t+1}\,\overline{\widetilde{R}}(-s).$$

Because the Frobenius norm is not normalized by the dimension, every replacement must cost $o(1)$ there, which is more than the trace statements of appendix G.1 give. The delicate replacement is the covariance between the conditional mean of α_2 and the deleted mixed resolvent. It is expanded along a martingale over the other samples; replacing one of them changes the buffer by rank two, which moves the conditional mean by $O(w_j^2/B)$ (lemma SC.4), and the covariance is $O(w_j^2 B^{-1/2})$ in Frobenius norm (lemma SC.5).

The integral is split at a cutoff. The exact identity, without any asymptotic replacement, bounds $\|\mathcal{H}(s)\|_{\mathrm{F}} \le C/s$ for $s \ge 1$ at every dimension and for every state, and $\|T(s)\|_{\mathrm{F}} \le C_\Sigma/s$ (theorem SC.1(ii)). With $\overline{\widetilde{G}}_\varrho = \frac{2}{\pi}\int_0^\infty T(\varrho + u^2)\,du$ this gives, for every $\mathsf{R} \ge 1$,

$$\|\,\mathbb{E}\,Q_\varrho(M_t) - \overline{\widetilde{G}}_\varrho\|_{\mathrm{F}} \le \frac{2}{\pi}\Big[\mathsf{R}\sup_{\varrho_- \le s \le \varrho_+ + \mathsf{R}^2}\|\mathcal{H}(s) - T(s)\|_{\mathrm{F}} + \frac{C}{\mathsf{R}}\Big],$$

and the dimension limit followed by $\mathsf{R} \to \infty$ gives the output.

The candidate acts entrywise in the eigenbases. Since $\overline{R}(-s)$ and $\overline{\widetilde{R}}(-s)$ are functions of Σ_{out} and Σ_{in},

$$u_i^\top T(s) v_j = \frac{s\,L_\beta(q(-s))\,\mu_i\lambda_j}{(s + \tilde{s}_1(-s)\mu_i)(s + s_1(-s)\lambda_j)} \cdot \frac{u_i^\top \Delta_t v_j}{\sqrt{2\mathscr{R}(t)}},$$

and integrating at $s = \varrho + u^2$ gives $u_i^\top \overline{\widetilde{G}}_\varrho v_j = \mathfrak{d}(\mu_i, \lambda_j)\,u_i^\top \Delta_t v_j/\sqrt{\mathscr{R}(t)}$ with the kernel eq. (SC.4). The same argument applied to the buffer truncated to its first J ages shows that, after the dimension limit, ages beyond J contribute a vanishing amount as $J \to \infty$ (proposition SC.1); the dynamics uses this.

G.3 Volatility and the lag kernel

The volatility is computed in the frozen-state model of definition SA.3: the error is held at a state whose normalized error $\widehat{A}_t$ has vanishing operator norm, and the momentum buffer $M_{t'} = \sum_{j\ge 0} w_j G_{t'+1-j}$ is a stationary average of independent batches at that state. The updates $\mathsf{S}_{t'} = Q_\varrho(M_{t'})$ are then stationary, and $\mathsf{S}^{\mathrm{c}}_{t'}$ is the update minus its mean. For matrices F, G diagonal in the eigenbases, $C_\ell[F,G] = B^{-1}\,\mathbb{E}\,\mathrm{Tr}[F\mathsf{S}^{\mathrm{c}}_t G\,\mathsf{S}^{\mathrm{c}\top}_{t+\ell}]$ is the covariance of two updates ℓ steps apart; with $F = u_i \otimes u_i$ and $G = v_j \otimes v_j$ it is that of the entries of mode (i,j). The engine produces the forcing measure ζ_ϱ. Its atom at a point (μ,λ) is the deterministic long-run covariance of the update tested on the eigenspaces of μ and λ, and each mode (i,j) contributes the mass $\mathfrak{v}_{ij}/B$ at (μ_i,λ_j). The engine shows that the covariance measure of T consecutive updates, divided by BT, is within an explicit error of its long-run value and within $o(1)$ of ζ_ϱ, and that $\mathfrak{v}_{ij}$ is a series of nonnegative double integrals of a closed-form kernel.

Proof sketch. Over a block of T steps the updates accumulate their lag covariances, which decay geometrically in the lag. Expanding the square of $S_T = \sum_{t'\le T}\mathsf{S}^{\mathrm{c}}_{t'}$ over pairs of times, stationarity makes all pairs at lag ℓ equal. This gives an exact identity, and the geometric decay bounds its distance to the long-run variance $\Gamma[F,G] = C_0[F,G] + 2\sum_{\ell\ge1} C_\ell[F,G]$ (theorem SD.2(i),(iii)):

$$\begin{aligned}\frac{1}{BT}\,\mathbb{E}\,\mathrm{Tr}\left[F\,S_T\,G\,S_T^\top\right] &= C_0[F,G] + 2\sum_{\ell=1}^{T-1}\Big(1 - \frac{\ell}{T}\Big)C_\ell[F,G] \\ &= \Gamma[F,G] + O\Big(\frac{\beta^T}{1-\beta} + \frac{\beta}{T(1-\beta)^2}\Big).\end{aligned}$$

The decay comes from the recursion $M_{t+\ell} = \beta^\ell M_t + \sum_{r<\ell} w_r G_{t+\ell+1-r}$. Replacing M_t in it by a copy built from independent batches gives a buffer with the law of $M_{t+\ell}$, independent of M_t, at distance β^ℓ times the difference of two independent buffers. The regularized map is $\varrho^{-1/2}$-Lipschitz in Frobenius norm, so $|C_\ell| \le \mathsf{K}_\star\beta^\ell$. The error term of the display holds also in total variation for the covariance measure of T consecutive updates (theorem SD.8(iv)), the form the dynamics use.

At each fixed lag, the covariance is unchanged in the limit when every residual $d = z^\top \widehat{A}_t w$, which depends on its own sample, is replaced by an independent Gaussian weight. Call a *filter* an odd function applied to the singular values of a buffer: Q_ϱ is the filter $x/\sqrt{x^2+\varrho}$, and it is an average

of the resolvent filters $x/(x^2+s)$, which give $(KK^\top + s\,\mathrm{Id})^{-1}K$ (lemma SA.3). Adding one sample is a rank-one change of each buffer. By the Woodbury formula, with a 2×2 matrix to invert, the change of a tested product of two resolvents is a function of d and of quadratic forms that concentrate at normalized traces, so it depends on the sample only through the law of d, which is asymptotically Gaussian because $\|\widehat{A}_t\| \to 0$. Summed over the samples, one at a time, the errors vanish (lemma SD.6).

In the resulting model, which has no state, the lag covariances have deterministic limits in which only tree-shaped Gaussian pairings survive. For polynomial filters the tested product of the two buffers is a polynomial in Gaussian variables, and its expectation is a sum over pairings, each gluing the cyclic product into a connected graph. Pairings whose graph is a tree contribute order one and the others $O(1/B)$ (lemma SD.8). The tree values are products of normalized traces $B^{-1}\sum_i \mu_i^p$, $B^{-1}\sum_j \lambda_j^p$ and of moments of the weights, hence deterministic, and polynomial approximation extends the limit to all odd bounded continuous filters (lemma SD.10). The resolvent filters give the lag kernel $\mathfrak{k}_\ell[F,G](s,u)$, and the filter of Q_ϱ gives the deterministic covariance $\overline{C}_\ell[F,G](\varrho,\varrho)$, the double integral of $\frac{4}{\pi^2}\mathfrak{k}_\ell[F,G](\varrho+x^2,\varrho+y^2)$ over $x,y>0$ (theorems SD.3 and SD.6). With the geometric decay, Γ is within $o(1)$ of $\overline{\Gamma} = \overline{C}_0 + 2\sum_{\ell\geq 1}\overline{C}_\ell$, without a rate (theorem SD.4(iv)).

The trees can be summed, and the lag kernel is an explicit expression in the solution of the fixed-point system of fig. 2 at two spectral parameters and in two scalar functions of the momentum. In every tree, the subtrees hanging off the path from the vertex carrying the test F to the vertex carrying the test G involve one buffer only, and they sum to its one-time deterministic equivalent (appendix G.1): the branch $(\sigma,\tilde{\sigma},\tilde{s}_1,s_1,q)$ at $-s$ for the earlier buffer and, primed, at $-u$ for the later one (lemma SD.18). Only batches on the path are shared; a batch of age j enters the two buffers with weights w_j and $\beta^\ell w_j$. The sum over paths is a geometric series of 2×2 matrices, and, first for large s,u and then by analyticity for all $s,u>0$ (theorem SD.5),

$$\mathfrak{k}_\ell[F,G](s,u) = \frac{su\,b_\ell\ \mathrm{Tr}\left(F\Sigma_{\mathrm{out}}\overline{R}(-s)\overline{R}(-u)\right)\mathrm{Tr}\left(G\Sigma_{\mathrm{in}}\overline{\widetilde{R}}(-s)\overline{\widetilde{R}}(-u)\right)}{B^2\,\Delta_\ell(s,u)},$$

$$b_\ell = \sum_{j\geq 0}\mathbb{E}\,\frac{\beta^\ell w_j^2 g^2}{(1+qw_j^2g^2)(1+\beta^{2\ell}q'w_j^2g^2)},$$

where $\overline{R}(-s) = (s\,\mathrm{Id}+\tilde{s}_1\Sigma_{\mathrm{out}})^{-1}$, $\overline{\widetilde{R}}(-s) = (s\,\mathrm{Id}+s_1\Sigma_{\mathrm{in}})^{-1}$, and $\Delta_\ell>0$ is the determinant of the 2×2 series, built from the branch, the traces of the same form with $F=\Sigma_{\mathrm{out}}$ and $G=\Sigma_{\mathrm{in}}$, b_ℓ and a companion scalar e_ℓ. The lag enters only through b_ℓ and e_ℓ, and b_ℓ carries the factor β^ℓ. With spectral projections as tests the kernel factors into multiplicities and a nonnegative mode kernel, $\mathfrak{k}_\ell[P_\mu,P'_\lambda](s,u) = r_\mu t_\lambda B^{-2}\,\mathfrak{h}_\ell(\mu,\lambda;s,u)$ for the spectral projections P_μ, P'_λ of $\Sigma_{\mathrm{out}},\Sigma_{\mathrm{in}}$ with ranks r_μ,t_λ, and the atom of ζ_ϱ at (μ_i,λ_j) splits into equal mode coefficients (corollary SD.3, proposition SD.3)

$$\mathfrak{v}_{ij}(\varrho) = \frac{1}{B}\Big(\mathfrak{c}_0(\mu_i,\lambda_j;\varrho) + 2\sum_{\ell\geq 1}\mathfrak{c}_\ell(\mu_i,\lambda_j;\varrho)\Big),$$

where $\mathfrak{c}_\ell(\mu,\lambda;\varrho)\geq 0$ is the double integral of $\frac{4}{\pi^2}\mathfrak{h}_\ell(\mu,\lambda;\varrho+x^2,\varrho+y^2)$ over $x,y>0$. ◇

G.4 From the algorithm to the deterministic dynamics

This is the argument of theorem H.1, sketched in full after that theorem. It takes in the drift kernel $\mathfrak{d}$ (appendix G.2), the forcing measure ζ with the covariance of consecutive updates (appendix G.3), and an initial error that is *diffuse*: the normalized error $\widehat{A}_0$, of Frobenius norm one, has vanishing operator norm. It produces the comparison, up to a time $T_\star$, of the spectral energy measure $\chi(\Delta_t)$, which places $\mathscr{Q}_{ij}$ at (μ_i,λ_j), with the solution of the deterministic measure equation (eq. (SE.9)).

The argument has four ideas. The algorithm is compared, over blocks of $L=\lfloor B^{1/8}\rfloor$ steps, with the model in which the state is held at its value at the start of the block and the momentum buffer is a stationary average of independent batches; because a test of the energy measure is a quadratic form in the error, its expected change over a block has an exact expansion (eq. (SE.28)). The two leading terms of that expansion are the drift and the volatility, and because the block is long the

momentum correlations of consecutive updates are summed in full. The motion of the state inside a block changes the expected update only through a causal response, which Stein's identity reduces to projections that are small because the error is diffuse (proposition SE.4). Finally, diffuseness persists, by an operator-norm bound on the mean update that holds at every state and a matrix martingale inequality (lemma SE.24). Gronwall's inequality and routine estimates close the comparison.

G.5 Power-law and large-batch asymptotics

This engine computes the half-anisotropic coefficients $\mathfrak{d}_i$ and $\mathfrak{v}_i$ when $\Sigma_{\text{in}} = \mathrm{Id}_N$ and $N_{\text{out}} = N_{\text{in}} = N$, in two regimes: the thin batch $N/B \to \infty$ with a power-law spectrum, and the oversampled batch $B_{\text{eff}} > 3(1+\beta)^2 N$ with an arbitrary spectrum, where only the drift is computed. It takes in the scalar fixed-point system of the deterministic equivalent (appendix G.1) and the integral formulas for the drift and the volatility (appendices G.2 and G.3), and it produces explicit profiles. At each finite dimension the coefficients are deterministic functions of the spectrum, of B and of β; the zero-ridge coefficient is their limit as $\varrho \downarrow 0$, and nothing is claimed about the algorithm at zero ridge.

With an isotropic input side, the fixed-point system reduces to a single scalar, the participation m. Put

$$m_N(\Lambda) = \frac{1}{B}\sum_{i=1}^{N} \frac{\Lambda\mu_i}{1+\Lambda\mu_i}, \qquad \Lambda_N \overset{\text{def}}{=} m_N^{-1}, \qquad f_i(m) = \frac{\Lambda_N(m)\mu_i}{1+\Lambda_N(m)\mu_i}.$$

Each eigenvalue enters with a soft indicator of $\mu_i \geq 1/\Lambda$, so $Bm_N(\Lambda)$ is a soft count of the output directions above the level $1/\Lambda$, and $\sum_i f_i(m) = Bm$. On the branch at the spectral parameter $-s$ the participation equals $r_\beta(q) \overset{\text{def}}{=} qF_\beta(q)$, which counts, up to a bounded error, the past batches that the normalized update whitens (lemma SG.1(iii)), and s fixes m through the exact relation $s = (\gamma - m)m/(Q_m\Lambda_N(m))$, with $Q_m = r_\beta^{-1}(m)$; this is a decreasing bijection of $(0,\gamma)$ onto $(0,\infty)$ (proposition SG.1). In this coordinate the drift is a single integral and the zero-lag volatility is f_i itself (eq. (SG.7)):

$$\mathfrak{d}_i(\varrho) = \frac{\sqrt{2}}{\pi\gamma}\int_0^\infty \frac{L_\beta(Q_{m(\varrho+u^2)})}{F_\beta(Q_{m(\varrho+u^2)})}\, f_i\big(m(\varrho+u^2)\big)\,\mathrm{d}u, \qquad \mathfrak{v}_i^{(0)}(\varrho) = f_i\big(m(\varrho)\big).$$

The volatility is a double integral over two participations with an explicit lag kernel (eq. (SG.22)). With momentum r_β is unbounded, so $m \to \gamma$ as $s \downarrow 0$, and at zero ridge the zero-lag coefficient of every row is one, $\mathfrak{v}_i^{(0)}(0) = 1$ (propositions SG.2 and SG.4).

The drift integral splits at participation of order one into a head, where few past batches are whitened, and a tail, where many are. In the head q is small, the weight functions take their small-argument values $F_\beta \approx s_2$ and $L_\beta \approx 1$, and the integrand is a Lorentzian in u of width $\sqrt{\gamma s_2 \mu_i}$. When the row is resolved, that is $m_N(1/\mu_i) \to 0$, the head part of the drift is (lemma SG.11)

$$\sqrt{2}\,\mathfrak{d}_i^{\leq M} \sim \sqrt{\frac{\mu_i}{\gamma s_2}} = \sqrt{\frac{\mu_i B_{\text{eff}}}{N}}.$$

In the tail the weight function $\sqrt{q}\,L_\beta(q)$ oscillates with the geometric spacing of the weights and has no limit, but its average in the participation is exact. With $\mathcal{A}(m) \overset{\text{def}}{=} \sqrt{Q_m}L_\beta(Q_m)/E(m)$ and $E(m) \overset{\text{def}}{=} Q_m r'_\beta(Q_m)$ (lemma SG.2),

$$\sup_{t \geq m_0}\Big|\int_{m_0}^{t}\big(\mathcal{A}(m) - \sqrt{2\pi}\big)\,\mathrm{d}m\Big| \leq C_{\beta,m_0}.$$

In the variable q each past batch contributes $\int_0^\infty x^{-1/2}J(x)\,\mathrm{d}x = \pi\,\mathbb{E}\,|g| = \sqrt{2\pi}$ once it is whitened, and m counts the whitened batches up to a bounded error. An integration by parts then replaces $\mathcal{A}$ by $\sqrt{2\pi}$ against any slowly varying weight (eq. (SG.9)). This is where the constant $c_0 = \sqrt{2\pi}/\pi = \mathbb{E}\,|g|$ of the profiles comes from.

For a power law the participation is a Riemann sum of a continuum function, and the small-argument law of that function separates the cases $\alpha < 1$, $\alpha = 1$ and $\alpha > 1$. With $V = c_N\Lambda N^{-\alpha}$ one has $m_N(\Lambda) = \gamma S_N(V)$, where $S_N(V) = N^{-1}\sum_j V/(V + (j/N)^\alpha)$ is within $1/N$ of

$S_\alpha(V) = \int_0^1 V(V+t^\alpha)^{-1}\mathrm{d}t$, so that $\Lambda_N(m) = (N^\alpha/c_N)V_N(m/\gamma)$ with $V_N = S_N^{-1}$ (eq. (SG.42), lemma SG.6). As $V \downarrow 0$ (lemma SG.5),

$$S_\alpha(V) \sim \frac{V}{1-\alpha}\ (\alpha<1), \qquad S_1(V) \sim V\log\frac{1}{V}, \qquad S_\alpha(V) \sim K_\alpha V^{1/\alpha}\ (\alpha>1),$$

with $K_\alpha = \pi/(\alpha\sin(\pi/\alpha))$. For $\alpha > 1$ this says that a participation m resolves about Bm/K_α rows: $c_N\Lambda_N(m) \sim (Bm/K_\alpha)^\alpha$ when $Bm \to \infty$ and $m/\gamma \to 0$ (lemma SG.6(iv)). The amplitude c_N multiplies the drift by $\sqrt{c_N}$ and leaves the volatility unchanged (lemma SG.4), so the proofs take $c_N = 1$. A uniform statement over the rows reduces, by passing to subsequences, to row sequences along which i/N and the relevant ratios converge.

In the oversampled regime the branch is the momentum-free profile equation up to a factor that stays in $[1-\epsilon_{\mathrm{os}}, 1]$, with $\epsilon_{\mathrm{os}} = 3(1+\beta)^2 N/B_{\mathrm{eff}}$. In the rescaled variables $\hat{s} = s/(s_2\gamma_{\mathrm{in}})$ and $b = \tilde{s}_1(-s)/(s_2\gamma_{\mathrm{in}})$, where $\gamma_{\mathrm{in}} = N/B$, the branch satisfies exactly (lemma SH.4)

$$\frac{b}{\widehat{F}} + m_{\hat{\nu}}(\hat{s}/b) = 1, \qquad \widehat{F} = \frac{F_\beta(q)}{s_2} \in (0,1], \qquad m_\nu(c) = \int \frac{x\,\nu(\mathrm{d}x)}{c+x},$$

where $\hat{\nu}$ is the empirical spectral distribution of Σ_{out}; this is the same soft count as above, since $m_N(\Lambda) = \gamma\, m_{\hat{\nu}}(1/\Lambda)$. The coupling gives $qF_\beta(q) < \gamma_{\mathrm{in}}$, which bounds q, and the Gaussian inequality $J(x) \geq 1/(1+3x)$ (Cauchy–Schwarz with $\mathbb{E}\, g^4 = 3$) then keeps $\widehat{F}$ and L_β above $1-\epsilon_{\mathrm{os}}$ for every s (lemma SH.3). A monotone comparison places b between $\widehat{F}b_0$ and b_0, where b_0 is the root at $\widehat{F} = 1$, and hence the drift integrand within the factors $(1-\epsilon_{\mathrm{os}})^{\pm 1}$ of the integrand of the profile $\mathfrak{d}^{\mathrm{os}}$ (lemma SH.5, theorem SH.1). The substitution $c = \hat{s}/b_0$ turns the profile equation into $\Upsilon_\nu(c) \overset{\mathrm{def}}{=} c(1-m_\nu(c)) = \hat{s}$, and one integration by parts gives the exact representation (lemma SH.8)

$$\mathfrak{d}^{\mathrm{os}}_\nu(\mu) = \frac{\sqrt{2}}{\pi}\,\mu\int_0^\infty \frac{\sqrt{\Upsilon_\nu(c)}}{(c+\mu)^2}\,\mathrm{d}c = \sqrt{\frac{\mu}{2}}\,\chi_\nu(\mu), \qquad \chi_\nu(\mu) = \int_0^\infty \varpi(y)\sqrt{1-m_\nu(\mu y)}\,\mathrm{d}y,$$

with $\varpi(y) = \frac{2}{\pi}\sqrt{y}\,(1+y)^{-2}$. Here ϖ is a probability density and $1-m_\nu(c)$ is a soft fraction of the spectrum below c, so the efficiency $\chi_\nu(\mu) \in (0,1]$ averages the square root of the fraction of the spectrum below the level of the mode, on a logarithmic scale around it (remark SH.2). The results on head, packed and tail modes are bounds on this average.

G.6 Volterra analysis and the phase map

This is the argument of theorem M.3, and the common part of every result of appendix M. It takes in the deterministic row model of that chapter: the row recursion eq. (14) in continuous time, with weights μ_i, drifts $\mathfrak{d}_i$, volatilities $\mathfrak{v}_i$ and initial row energies that are fixed numbers (definition SI.1), run at the matched step $\eta = 2\sqrt{\epsilon}/\mathcal{S}$. It produces an exact formula for the time $t_{T\epsilon}$ as a prefactor times a dimensionless integral, lower and upper bounds for that integral, and, for the power-law arrays, its order in each range of exponents.

Proof sketch. The risk of the model solves a scalar equation once time is measured in units of the effective step. In the row equations the drift of every row is divided by $\sqrt{\mathcal{R}}$, so the natural time is $s = \phi(t) = \int_0^t \eta\,\mathcal{R}(t')^{-1/2}\,\mathrm{d}t'$. In this time each row solves a linear equation driven by $\sqrt{Q}$, where $Q(s)$ is the risk at time s, and summing the rows with weights $\mu_i/2$ gives exactly (theorems SI.1 and SI.2)

$$Q = F + k * \sqrt{Q}, \qquad t(s) = \frac{1}{\eta}\int_0^s \sqrt{Q(v)}\,\mathrm{d}v,$$
$$F(s) = \tfrac{1}{2}\sum_i \mu_i \mathcal{Q}_i(0) e^{-2\mathfrak{d}_i s}, \qquad k(s) = \tfrac{\eta}{2}\sum_i \mu_i \mathfrak{v}_i e^{-2\mathfrak{d}_i s}.$$

The forcing F is the risk the model would have without noise, and the kernel k, of total mass $\eta\mathcal{S}/2$, feeds the volatility back in. Monotone iteration gives a solution, unique by the comparison principle below, and its limit is $\eta^2\mathcal{S}^2/4$, which is ϵ at the matched step; there the kernel has mass $\sqrt{\epsilon}$.

The time to a level is the integral of $\sqrt{Q}$ up to the first passage, and it is squeezed between the forcing and any function that dominates the equation. Since $k * \sqrt{Q} \geq 0$, the risk is at least the forcing, so it cannot reach a level before the forcing does, and $t_a \geq \eta^{-1}\int_0^{s_F}\sqrt{F}$, with s_F the first time $F \leq a$

(theorem SI.3). Conversely, a positive function U with $U \geq F + k * \sqrt{U}$ lies above Q: where $Q > U$ the difference is bounded by a fixed multiple of its own convolution with k; since k has small mass on short intervals, this forces it to vanish step by step (lemma SI.3).

For the power-law arrays a change of time scale separates the dimension from the accuracy. With the time scale T_0^j the decay rates $2\mathfrak{d}_i T_0^j$ become $2i^{-q_j}$, and $Y_j(u) = Q_j(T_0^j u)$ solves $Y_j = f_{N,q_j} + \sqrt{\epsilon}\, g_j * \sqrt{Y_j}$, where $f_{N,q}(u) = \frac{1}{2}\sum_{i \leq N} i^{-p} e^{-2ui^{-q}}$ and g_j is a decreasing kernel of mass one; neither depends on B (proposition SI.4). If $\hat{u}_j$ is the first time $Y_j \leq T\epsilon$, then exactly (eqs. (SI.26) and (SI.30))

$$t^j_{T\epsilon} = \frac{\mathcal{S}_j T_0^j}{2\sqrt{\epsilon}}\, \mathsf{I}_j, \qquad \mathsf{I}_j \stackrel{\text{def}}{=} \int_0^{\hat{u}_j} \sqrt{Y_j(u)}\, \mathrm{d}u, \qquad \frac{\mathcal{S}_{\text{s-SGD}} T_0^{\text{s-SGD}}}{\mathcal{S}_{\text{s-SVD}} T_0^{\text{s-SVD}}} = D_N.$$

The order of the clock integral I_j is decided by whether $\sqrt{f_{N,q_j}}$ is integrable uniformly in N. The forcing lies between two power laws (lemma SI.7), the lower one for $N \geq 4$ and $u \leq (N/4)^q$:

$$A_p (1+u)^{-2r_q} \leq f_{N,q}(u) \leq C^2_{p,q} (1+u)^{-2r_q}, \qquad r_q = \frac{p-1}{2q}.$$

Write $\mathsf{I}_j = \int_0^{\hat{u}_j} \sqrt{f_{N,q_j}} + \mathcal{E}_j$, where $\mathcal{E}_j \geq 0$ is the noise part. When $r_j \leq 1$, suppose the first passage occurs before a horizon $H \leq (N/4)^{q_j}$. There the lower envelope holds, and $f_{N,q_j}(\hat{u}_j) \leq T\epsilon$ gives $1 + \hat{u}_j \geq (A_p/T\epsilon)^{1/(2r_j)}$. Integrating the envelopes then gives (lemma SI.10)

$$\int_0^{\hat{u}_j} \sqrt{f_{N,q_j}} \begin{cases} \text{lies between } c\,\epsilon^{-(1-r_j)/(2r_j)} \text{ and } C_{p,q_j} I_{r_j}(H), & r_j < 1, \\ = \dfrac{\log(1/\epsilon)}{2\sqrt{8q_j}} + O\big(1 + \log_+((1+H)\sqrt{\epsilon}) + HN^{-q_j}\big), & r_j = 1, \\ \longrightarrow J(p,q_j) \stackrel{\text{def}}{=} \int_0^\infty \sqrt{f_{\infty,q_j}(u)}\, \mathrm{d}u \quad (N \to \infty,\ \epsilon \downarrow 0), & r_j > 1, \end{cases}$$

with $I_r(x) = \int_0^x (1+w)^{-r} \mathrm{d}w$, so that $C_{p,q} I_r(H)$ is of order $\epsilon^{-(1-r)/(2r)}$ when H is of order $\epsilon^{-1/(2r)}$. Since $r_{\text{s-SVD}} = (p-1)/\alpha$ and $r_{\text{s-SGD}} = (p-1)/(2\alpha)$, the boundary $r_j = 1$ is $\varsigma = 1$ for SIGNSVD and $\varsigma = \alpha + 1$ for SIGNSGD.

The noise part is controlled by comparison with $(h + \sqrt{\epsilon})^2$, and checking the kernel condition is what produces the windows. Take for h a power-law envelope of $\sqrt{f_{N,q_j}}$ that decreases to $\sqrt{\epsilon}$ at a horizon H. If $g_j * h \leq 2h$ or $g_j * h \leq \sqrt{\epsilon}$ on $[0, H]$, then $(h + \sqrt{\epsilon})^2$ dominates the equation, the first passage occurs before H, and $\mathcal{E}_j$ is at most a fixed multiple of $\epsilon^{(r_0-1)/(2r_0)}$, where $r_0 \leq r_j$ is the exponent of the envelope and $r_0 > 1$ exactly when $r_j > 1$ (lemma SI.9). For $r_j < 1$ this is of the order of the main term, so these models get orders; for $r_j > 1$ it tends to zero, and the time has an equivalent. (For SIGNSVD at $\alpha = 2$, $\varsigma > 1$, lemma SI.18 adds a correction term to the square and bounds the noise part by a constant that tends to zero.) The kernel condition holds by small local mass of a decreasing kernel, by an integrable power tail, or by a small kernel height, each on a range of times that the window of the theorem keeps above the horizon (see the note to theorem M.3). $\diamond$

H Dynamics: from the algorithm to the deterministic recursion

This chapter connects the algorithm to the deterministic recursion eq. (6). The algorithm runs at the step $\eta = h/B$, so a unit of the macroscopic time $\tau = t/B$ is B steps. The error is $\Delta_t = W_t - W^\star$, and its normalized form $\widehat{A}_t = {\Sigma_{\text{out}}}^{1/2} \Delta_t {\Sigma_{\text{in}}}^{1/2} / \sqrt{2\mathscr{R}(t)}$ has Frobenius norm one. With $\mathfrak{d}_{ij} = \mathfrak{d}(\mu_i, \lambda_j)$ the drift kernel (corollary I.1) and $\mathfrak{v}_{ij}$ the volatility kernel, which is B times the mass of the forcing measure ζ at (μ_i, λ_j) divided equally among the modes there (appendix G.3), the deterministic mode-energy equations eq. (SE.7), divided by B^2, and the two spectral energy measures are

$$\frac{\mathrm{d}\Omega_{ij}}{\mathrm{d}\tau} = -\frac{2h\,\mathfrak{d}_{ij}}{\sqrt{\mathcal{R}}}\,\Omega_{ij} + \frac{h^2}{B}\,\mathfrak{v}_{ij}, \qquad \mathcal{R} = \frac{1}{2}\sum_{i,j} \mu_i \lambda_j \Omega_{ij},$$

$$\chi^{\text{act}}(\tau) = \sum_{i,j} \mathscr{Q}_{ij}(\lfloor B\tau \rfloor)\, \delta_{(\mu_i,\lambda_j)}, \quad \chi^{\text{ode}}(\tau) = \sum_{i,j} \Omega_{ij}(\tau)\, \delta_{(\mu_i,\lambda_j)},$$

with $\mathcal{Q}_{ij}(0) = \mathscr{Q}_{ij}(0)$; the measures live on $[0, C_\Sigma]^2$. The norm $\|m\|_{\mathrm{BL}}$ of a signed measure is the supremum of $|\int \phi \,\mathrm{d}m|$ over tests with $\|\phi\|_\infty \le 1$ that are 1-Lipschitz for the ℓ^1 distance. Let $\gamma_\star$, C_Σ, $E_\star$ and $r_0 \le r_1$ be the constants fixed in theorem H.1: bounds on the aspect ratios, on the covariances, on $\|\Delta_0\|_{\mathrm{F}}^2$, and on the initial risk from below and above. The horizon $T_\star = \min(r_0, 1)/(8(1+C_E+C_{\mathscr{R}}))$ (eq. (SE.10)) is explicit, with C_E and $C_{\mathscr{R}}$ bounds on the rates of change of the energy and of the risk along the equations, depending only on $(h, \varrho, \beta, \gamma_\star, C_\Sigma, E_\star, r_0, r_1)$. The theorem compares χ^{act} with χ^{ode} for general covariances, by the argument of appendix G.4; the corollary sums the equations over the columns when $\Sigma_{\mathrm{in}} = \mathrm{Id}$.

Theorem H.1 (Dynamical comparison)**.** *Fix $h > 0$, $\varrho > 0$, $\beta \in [0, 1)$, $\gamma_\star < \infty$, $C_\Sigma \ge 1$, $E_\star < \infty$, $c_M < \infty$ and $0 < r_0 \le r_1 < \infty$. Along a sequence of dimensions with $B \to \infty$ assume:*

- *(H1) Assumptions 2 and 3 hold with deterministic covariances, $\max(N_{\mathrm{in}}, N_{\mathrm{out}}) \le \gamma_\star B$ and $\|\Sigma_{\mathrm{in}}\|, \|\Sigma_{\mathrm{out}}\| \le C_\Sigma$;*
- *(H2) the algorithm of definition F.1 is run with momentum β, normalized ridge ϱ and step $\eta_t = h/B$;*
- *(H3) W_0 is deterministic, $\|\Delta_0\|_{\mathrm{F}}^2 \le E_\star$, $r_0 \le \mathscr{R}(0) \le r_1$, and $\|\widehat{A}_0\| \to 0$;*
- *(H4) the initial buffer M_{-1}° is independent of all batches and $\mathbb{E}\,\|M_{-1}^\circ\|_{\mathrm{F}}^2 \le c_M B$.*

Let $T_\star$ be given by eq. (SE.10) *and let $(\mathcal{Q}_{ij})$ be the solution of eq.* (SE.7) *of proposition SE.1. Then the following hold.*

- *(i) With probability tending to one,*
$$t_{\mathrm{stop}} > BT_\star, \qquad \inf_{t \le BT_\star} \mathscr{R}(t) \ge r_0/2, \qquad \sup_{t \le BT_\star} \|\Delta_t\|_{\mathrm{F}}^2 \le E_\star + 1.$$
- *(ii) $\sup_{\tau \le T_\star} \|\chi^{\mathrm{act}}(\tau) - \chi^{\mathrm{ode}}(\tau)\|_{\mathrm{BL}} \to 0$ in probability.*
- *(iii) $\sup_{\tau \le T_\star} \|\mu\lambda\,(\chi^{\mathrm{act}}(\tau) - \chi^{\mathrm{ode}}(\tau))\|_{\mathrm{BL}} \to 0$ and $\sup_{\tau \le T_\star} |\mathscr{R}(\lfloor B\tau \rfloor) - \mathcal{R}(\tau)| \to 0$ in probability.*
- *(iv) On the event of (i), $\sup_{t \le BT_\star} \|\widehat{A}_t\| \to 0$ in probability.*

Full statement and proof: theorem SE.1 of the supplement. Lean: proved: `Muon.Dynamics.thm_app_dyn_main_i`, `Muon.Dynamics.thm_app_dyn_main_ii`, +4 more; foundation results: `GaussianOpNorm`, `GaussianOpNormL6`, `GaussianPoincare`, `GaussianQuadFormLp`.

This theorem is the precise form of theorem 2.1. At the step $\eta = h/B$, the projected risks $\mathscr{Q}_{ij}$ follow, in aggregate, the recursion eq. (6) with the drift kernel of corollary I.1 and the volatility coefficient of proposition SD.3; the equations eq. (SE.7) are that recursion in the macroscopic time $\tau = t/B$. "In aggregate" refers to the spectral energy measures χ^{act} and χ^{ode} defined at the start of this section, which place the projected risks at the points (μ_i, λ_j). The risk is half the integral of $\mu\lambda$ against such a measure, so (iii) is the risk statement of the main text. Nothing is claimed about a single mode.

The kernels are computed in the frozen-state model of definition SA.3, in which the momentum buffer is a stationary average of independent batches, all evaluated at one fixed state. That computation assumes the state is *diffuse*: the normalized error $\widehat{A}_t$ has vanishing operator norm, so no single singular direction carries a fixed fraction of the risk. Part (iv) says that this property, assumed at time zero in (H3), persists along the trajectory. This is what allows the frozen-state kernels to be used there.

Proof sketch. The algorithm is compared, over blocks of $L = \lfloor B^{1/8} \rfloor$ steps, with the frozen-state model at the state reached at the start of the block. Over a block the state moves by at most a constant times $L/\sqrt{B}$ in Frobenius norm (lemma SE.14), so we hold it at its block-start value D. We run the same batches at the fixed state D, with the momentum buffer started from an independent stationary past. The resulting updates V_r, $r < L$, are those of the frozen-state model at D (lemma SE.2), and $\overline{V}$ denotes their common mean. The spectral energy measure is tested by $X(\Delta) = \int f \otimes g \,\mathrm{d}\chi(\Delta)$ for bounded f, g. This is a quadratic form in Δ, so the expected change of X over a block has an exact

expansion. Write $\mathcal{W}X = f(\Sigma_{\text{out}})\, X\, g(\Sigma_{\text{in}})$, let S_{t+r} be the actual updates of the block starting at step t, and put

$$X(\Delta) = \langle \Delta, \mathcal{W}\Delta \rangle, \qquad S = \sum_{r<L} V_r, \qquad \mathcal{E} = \sum_{r<L} \mathsf{S}_{t+r} - S.$$

Then, exactly (eq. (SE.28)),

$$\begin{aligned}\mathbb{E}_t\left[X(\Delta_{t+L}) - X(D)\right] = & -2\eta \Big\langle \mathcal{W}D, \sum_{r<L} \mathbb{E}_t\, \mathsf{S}_{t+r} \Big\rangle + \eta^2 L^2 \langle \overline{V}, \mathcal{W}\overline{V} \rangle \\ & + \eta^2 LB \int f \otimes g \,\mathrm{d}\mathfrak{F}_L(D) + \eta^2\, \mathbb{E}_t\left[2\langle S, \mathcal{W}\mathcal{E}\rangle + \langle \mathcal{E}, \mathcal{W}\mathcal{E}\rangle\right],\end{aligned}$$

where $\mathfrak{F}_L(D)$ is the covariance measure of the sum of L consecutive updates of the frozen-state model (theorem SD.8(iv)).

The two leading terms of this expansion are the drift and the volatility of the main text. Replace $\mathbb{E}_t\, \mathsf{S}_{t+r}$ by $\overline{V}$ in the first term. The deterministic equivalent of corollary I.1 gives $u_i^\top \overline{V} v_j = \mathfrak{d}_{ij}\, u_i^\top D v_j / \sqrt{\mathscr{R}(D)}$ up to an error of vanishing Frobenius norm. The first term is therefore $-2\eta L\, \mathscr{R}(D)^{-1/2} \int f \otimes g\, \mathfrak{d}\, \mathrm{d}\chi(D)$ up to $o(L/B)$. The third term carries the covariance of a sum of L consecutive updates. Momentum correlates these updates, and because the block is long the correlations are summed in full: $\mathfrak{F}_L(D)$ is within $O(\beta^L + L^{-1})$ in total variation of the integrated lag covariance. That covariance is close to the forcing measure ζ, which carries the mass $\mathfrak{v}_{ij}/B$ at (μ_i, λ_j) (theorem SD.8(iii),(iv), proposition SD.3). The third term is therefore $(L/B)\, h^2 \int f \otimes g \,\mathrm{d}\zeta$ up to $o(L/B)$. The second term, the square of the mean step, is only $O(L^2/B^2)$. Over a block, then, the expected change of X is L/B times the test of the measure $\mathsf{F}_B(\chi(D))$ below. Summed over the blocks, this reproduces the deterministic equations, which in measure form read exactly (eq. (SE.9))

$$\chi^{\text{ode}}(\tau) = \chi^{\text{ode}}(0) + \int_0^\tau \mathsf{F}_B\big(\chi^{\text{ode}}(\sigma)\big)\,\mathrm{d}\sigma, \qquad \mathsf{F}_B(\chi) = -\frac{2h}{\sqrt{\mathcal{R}(\chi)}}\, \mathfrak{d}\,\chi + h^2\zeta,$$

with $\mathcal{R}(\chi)$ half the integral of $\mu\lambda$ against χ, and $\mathfrak{d}\,\chi$ the measure with density $\mathfrak{d}$ with respect to χ.

The motion of the state within a block does not change the expected update at leading order, because the error is diffuse. Apart from the effect of the batches carried in the buffer from before the block, replacing $\mathbb{E}_t\, \mathsf{S}_{t+r}$ by $\overline{V}$ neglects, to first order in η, the response $\eta\, \mathbb{E}\, \dot{V}_r$. Here $\dot{V}_r$ is the derivative of the frozen-state update with respect to the step size, at step size zero (lemma SE.8). This derivative is causal: in $\dot{V}_r$ the batch of step $t + j$ acts only on the updates V_k, $k < j$, which are built from earlier batches and are independent of it. Removing each sample of the acting batch from the buffer and integrating it out by Stein's identity (proposition SE.2) leaves mean products of the projections $\langle \Sigma_{\text{out}} D \Sigma_{\text{in}}, V_k \rangle$ with scalars of the reduced buffer. The product of their means is bounded without diffuseness (lemma SE.7). Their covariance is bounded through the change caused by a single sample, which for such a projection is $O(\delta(D))$ in L^2, where $\delta(D) = \|\widehat{A}(D)\|$ is the operator norm of the normalized error (proposition SE.3). The ridge also moves with the risk, and its contribution to $\dot{V}_r$ is bounded in the same way (lemma SE.13). For a test matrix Π of bounded Frobenius norm the result is (proposition SE.4)

$$\eta\, \big|\, \mathbb{E}\langle \Pi, \dot{V}_r \rangle \big| \le C\Big\{ \frac{r}{\sqrt{B}} + \frac{1}{B} + \delta(D)\big(\|\Pi\| + \delta(D)\big)\Big\}.$$

The relevant test is $\Pi = \mathcal{W}D$, whose operator norm is bounded but not small, so the response vanishes only because $\delta(D) \to 0$.

The error stays diffuse, by a drift bound and a matrix martingale for the weighted state $Y_b = {\Sigma_{\text{out}}}^{1/2} \Delta_{t_b} {\Sigma_{\text{in}}}^{1/2}$ at the start t_b of block b. The drift is controlled by an operator-norm bound on the mean update that holds at every state of bounded risk, diffuse or not, so the argument is not circular (lemma SE.7(iv)):

$$\|{\Sigma_{\text{out}}}^{1/2} \overline{V}\, {\Sigma_{\text{in}}}^{1/2}\| \le C \|{\Sigma_{\text{out}}}^{1/2} D\, {\Sigma_{\text{in}}}^{1/2}\| + C r_B, \qquad r_B \to 0.$$

The centered increments of Y_b have operator norm $O(L/B)$ in dimension $O(B)$, so their partial sums vanish by a matrix martingale inequality (lemma SE.23). A discrete Gronwall inequality then bounds $\max_b \|Y_b\|$ by $e^{Ch(T+1)}(\|Y_0\| + o_B)$, with $o_B \to 0$ in probability, and $\|Y_0\| \to 0$ by (H3).

The remaining errors are routine, with two exceptions. During the first $\lceil C_0 \log(2B) \rceil$ steps of a block, the momentum buffer still remembers the previous block, so these updates are not close to the stationary model. They are compared with the model started that many steps earlier, and their total cost is of order $\log^2(2B)/L$ (lemma SE.21). The deterministic equivalents for the drift and the volatility come without a rate. They are made uniform over diffuse states by a subsequence argument (lemma SE.25), and this is one reason the theorem gives no rate. The fluctuations of the block increments form a martingale whose maximum vanishes (lemma SE.20). Energy and risk then move at bounded rates, so they stay close to their initial values up to $T_\star$ (lemma SE.27), which is (i). Gronwall's inequality, with the Lipschitz continuity of F_B (lemma SE.1), and routine estimates give (ii) and (iii). ◇

When $\Sigma_{\text{in}} = \text{Id}_{N_{\text{in}}}$, the projected row risk is $\mathscr{Q}_i(t) = \|\Delta_t^\top u_i\|^2 = \sum_j \mathscr{Q}_{ij}(t)$, and the row coefficient $\mathfrak{v}_i(\varrho)$ of eq. (SD.75) is the deterministic counterpart, in aggregate over the rows, of the long-run variance of the row $\mathsf{S}^\top u_i$ of the update: B times the deterministic integrated covariance $\overline{\Gamma}[u_i \otimes u_i, \text{Id}](\varrho, \varrho)$ of appendix G.3.

Corollary H.1 (Half-anisotropic input)**.** *Assume the hypotheses of theorem H.1 with* $\Sigma_{\text{in}} = \text{Id}_{N_{\text{in}}}$. *Let* $\mathfrak{d}_i \overset{\text{def}}{=} \mathfrak{d}(\mu_i, 1)$ *and let* $\mathfrak{v}_i \overset{\text{def}}{=} \mathfrak{v}_i(\varrho)$ *be the row coefficient of eq.* (SD.75)*. Then the row equations*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\big(B\mathcal{Q}_i(\tau)\big) = -\frac{2h\,\mathfrak{d}_i}{\sqrt{\mathcal{R}(\tau)}}\,B\mathcal{Q}_i(\tau) + h^2\mathfrak{v}_i, \qquad \mathcal{R}(\tau) = \frac{1}{2}\sum_i \mu_i \mathcal{Q}_i(\tau), \qquad \mathcal{Q}_i(0) = \|\Delta_0^\top u_i\|^2, \tag{20}$$

have a unique solution on $[0, T_\star]$*, given by* $\mathcal{Q}_i = \sum_j \mathcal{Q}_{ij}$ *with* $(\mathcal{Q}_{ij})$ *the solution of eq.* (SE.7)*. The row measures* $\chi^{\text{act}}_{\text{row}}(\tau) \overset{\text{def}}{=} \sum_i \mathscr{Q}_i(\lfloor B\tau \rfloor)\,\delta_{\mu_i}$ *and* $\chi^{\text{ode}}_{\text{row}}(\tau) \overset{\text{def}}{=} \sum_i \mathcal{Q}_i(\tau)\,\delta_{\mu_i}$ *on* $[0, C_\Sigma]$ *satisfy* $\sup_{\tau \le T_\star} \|\chi^{\text{act}}_{\text{row}}(\tau) - \chi^{\text{ode}}_{\text{row}}(\tau)\|_{\text{BL}} \to 0$ *in probability, and the same holds for the measures weighted by* μ*.*

Full statement and proof: corollary SE.1 of the supplement. Lean: proved: `Muon.Dynamics.cor_app_dyn_half_i`*,* `Muon.Dynamics.cor_app_dyn_half_ii`*, +1 more.*

The corollary is the dynamical statement behind the half-anisotropic section of the main text, where $\Sigma_{\text{in}} = \text{Id}_N$ and only the output covariance is anisotropic. There the natural unknowns are the projected row risks $\mathcal{Q}_i = \sum_j \mathcal{Q}_{ij}$, the deterministic counterparts of $\mathscr{Q}_i$, and the risk $\mathcal{R} = \frac{1}{2}\sum_i \mu_i \mathcal{Q}_i$ involves only them. The corollary says that the row risks obey a closed system, the row recursion eq. (14) in the macroscopic time $\tau = t/B$: dividing eq. (20) by B and taking B steps of size $\eta = h/B$ per unit of time, the change over one step is $-2\eta\,\mathfrak{d}_i\mathcal{Q}_i/\sqrt{\mathcal{R}} + \eta^2\mathfrak{v}_i$. The algorithm follows this system in aggregate over the rows: the row measure, which places $\mathscr{Q}_i$ at μ_i, is close in the bounded-Lipschitz distance to its deterministic counterpart, and its μ-weighted version gives the risk. As in theorem H.1, nothing is claimed about a single row.

Proof sketch. With an isotropic input side, both kernels are constant along each row, so summing the mode equations over the columns closes them on the rows. The drift kernel $\mathfrak{d}_{ij} = \mathfrak{d}(\mu_i, \lambda_j)$ is evaluated at $\lambda_j = 1$ for every j, so $\mathfrak{d}_{ij} = \mathfrak{d}_i$ (corollary I.2). The mode volatilities of a row are equal, $\mathfrak{v}_{ij} = \mathfrak{v}_i/N_{\text{in}}$, and sum to the row coefficient $\mathfrak{v}_i$ (proposition J.1). Summing eq. (SE.7) over j and dividing by B therefore gives, exactly,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Big(B\sum_j \mathcal{Q}_{ij}\Big) = -\frac{2h\,\mathfrak{d}_i}{\sqrt{\mathcal{R}}}\,B\sum_j \mathcal{Q}_{ij} + h^2\sum_j \mathfrak{v}_{ij} = -\frac{2h\,\mathfrak{d}_i}{\sqrt{\mathcal{R}}}\,B\mathcal{Q}_i + h^2\mathfrak{v}_i,$$

with $\mathcal{R} = \frac{1}{2}\sum_{i,j}\mu_i\mathcal{Q}_{ij} = \frac{1}{2}\sum_i \mu_i\mathcal{Q}_i$ and $\mathcal{Q}_i(0) = \sum_j (u_i^\top \Delta_0 v_j)^2 = \|\Delta_0^\top u_i\|^2$ by Parseval's identity in the basis (v_j) (lemma SA.1).

The row system has a unique solution on $[0, T_\star]$. The mode system has a solution on $[0, T_\star]$ (proposition SE.1), so its row sums solve the row system there. The row system is locally Lipschitz where its risk is positive, so this solution is its only one (Picard–Lindelöf).

The row measures are images of the mode measures, and the projection $(\mu, \lambda) \mapsto \mu$ preserves the bounded-Lipschitz norm on measures carried by the segment $\{\lambda = 1\}$. Both $\chi^{\text{act}}(\tau)$ and $\chi^{\text{ode}}(\tau)$ live on that segment, and their images under the projection are the row measures, since $\mathscr{Q}_i = \sum_j \mathscr{Q}_{ij}$. A test on the rectangle restricts to a test $\phi(\cdot, 1)$ of the same size and Lipschitz constant, and a test ψ of μ

extends to the test $\psi(\mu)$ on the rectangle, so the two norms coincide. Theorem H.1(ii) therefore gives the row statement, and theorem H.1(iii) gives the μ-weighted one, because $\mu\lambda = \mu$ on the segment. All the probabilistic work is that of theorem H.1, sketched in appendix G.4: the block comparison of the algorithm with the frozen-state model, the identification of the drift and the volatility, and the persistence of diffuseness. ◇

I Drift

The drift kernel $\mathfrak{d}_{ij}$ of eq. (6) is computed in the frozen-state model of definition SA.3: the state Δ_t is held fixed, the momentum buffer is replaced by $M_t = \sum_{j\geq 0} w_j\, G_{t+1-j}$, with weights $w_j = (1-\beta)\beta^j$, of independent minibatch gradients all recomputed at Δ_t and divided by $\sqrt{2\mathscr{R}(t)}$, and the update is $\widetilde{G}_{t+1} = Q_\varrho(M_t)$. Its drift on the mode (i,j) is $\mathscr{D}_{ij}(t) = (u_i^\top \Delta_t v_j)\ \mathbb{E}[u_i^\top \widetilde{G}_{t+1} v_j]$ (eq. (SA.19)). Assumption S3 is the proportional regime in this model, with a normalized ridge in a fixed interval $[\varrho_-, \varrho_+] \subset (0,\infty)$ and a diffuse normalized error, $\|\widehat{A}_t\| \to 0$. For a nonnegative summable weight sequence w, not all zero, and a standard Gaussian g, put

$$J(x) = \mathbb{E}\,\frac{g^2}{1+xg^2}, \qquad F_w(q) = \sum_{j\geq 0} w_j^2\, J(w_j^2 q), \qquad L_w(q) = \sum_{j\geq 0} w_j\, J(w_j^2 q),$$

written F_β, L_β for the EMA weights. The weighted fixed-point system and its unique solution give the deterministic equivalent of the Gram resolvent of the buffer (appendix G.1), from which the expected update (appendix G.2) and the drift kernel are read off.

In the next definition w is a weight sequence and F_w the function of the chapter introduction, and the spectral measures of the covariances are $\nu_{\mathrm{out}} = \frac{1}{B}\sum_i \delta_{\mu_i}$ and $\nu_{\mathrm{in}} = \frac{1}{B}\sum_j \delta_{\lambda_j}$ (eq. (SA.41)). The spectral parameter $z = -s$ is negative.

Definition I.1 (Weighted fixed-point system)**.** *Let $(\nu,\tilde\nu)$ be finite positive Borel measures on $[0,\infty)$ with compact support, let w be as above, and let $z = -s$ with $s > 0$. The weighted fixed-point system at z is the system of equations in the four unknowns $(\sigma,\tilde\sigma,\tilde s_1, s_1) \in [0,\infty)^4$*

$$\begin{aligned} \sigma &= \int \frac{x}{\tilde s_1 x - z}\,\nu(dx), & \tilde\sigma &= \int \frac{x}{s_1 x - z}\,\tilde\nu(dx), \\ q &= -z\,\sigma\tilde\sigma, & \tilde s_1 &= -z\,\tilde\sigma\, F_w(q), \qquad s_1 = -z\,\sigma\, F_w(q). \end{aligned} \tag{21}$$

For the spectral measures of eq. (SA.41) *the first two equations read*

$$\sigma = \frac{1}{B}\,\mathrm{Tr}\left(\Sigma_{\mathrm{out}}(\tilde s_1\Sigma_{\mathrm{out}} - z\,\mathrm{Id}_{N_{\mathrm{out}}})^{-1}\right), \qquad \tilde\sigma = \frac{1}{B}\,\mathrm{Tr}\left(\Sigma_{\mathrm{in}}(s_1\Sigma_{\mathrm{in}} - z\,\mathrm{Id}_{N_{\mathrm{in}}})^{-1}\right),$$

and we define the deterministic equivalents

$$\overline{R}(z) \overset{\mathrm{def}}{=} \left(\tilde s_1(z)\,\Sigma_{\mathrm{out}} - z\ \mathrm{Id}_{N_{\mathrm{out}}}\right)^{-1}, \qquad \overline{\widetilde{R}}(z) \overset{\mathrm{def}}{=} \left(s_1(z)\,\Sigma_{\mathrm{in}} - z\ \mathrm{Id}_{N_{\mathrm{in}}}\right)^{-1}, \tag{22}$$

where $(\sigma,\tilde\sigma,\tilde s_1, s_1)(z)$ is the solution given by theorem I.1.

Full statement and proof: definition SB.1 of the supplement. Lean: proved: `Muon.DetEquiv.IsBranchSol`, `Muon.DetEquiv.branchSol`, *+5 more.*

For the EMA weights and the spectral measures of the covariances this is the system of fig. 2; the definition allows any weight sequence and any pair of measures. The unknowns $\sigma,\tilde\sigma$ are the deterministic values of the normalized traces $\frac{1}{B}\mathrm{Tr}(\Sigma_{\mathrm{out}}R(z))$ and $\frac{1}{B}\mathrm{Tr}(\Sigma_{\mathrm{in}}\widetilde{R}(z))$ of the Gram resolvents of the normalized buffer, and $\tilde s_1, s_1$ are the scalings with which Σ_{out} and Σ_{in} enter their deterministic equivalents $\overline{R}, \overline{\widetilde{R}}$. Multiplying the last two equations by σ and $\tilde\sigma$ gives the coupling $\tilde s_1\sigma = s_1\tilde\sigma = qF_w(q)$ of fig. 2. The system depends on the covariances only through their eigenvalues, and at $\beta = 0$ the weight function is $F_w = J$.

In the next theorem $s_2 = \sum_j w_j^2$, and $\mathcal{K}$ is the class of pairs $(\nu,\tilde\nu)$ of positive measures on $[0, C_\Sigma]$, each of total mass at most $\gamma_\star$ (eq. (SB.5)); it contains $(\nu_{\mathrm{out}}, \nu_{\mathrm{in}})$ at every dimension under Assumption S3.

Theorem I.1 (The weighted branch)**.** *Let w be nonnegative, summable and not all zero, and let $C_\Sigma, \gamma_\star > 0$.*

(i) (Existence and uniqueness.) For every $(\nu, \tilde\nu) \in \mathcal{K}$ and every $s > 0$ the system eq. (21) at $z = -s$ has exactly one solution in $[0,\infty)^4$. It satisfies

$$\sigma, \tilde\sigma \le \frac{\gamma_\star C_\Sigma}{s}, \qquad \tilde s_1, s_1 \le \gamma_\star C_\Sigma\, s_2, \qquad q \le \frac{(\gamma_\star C_\Sigma)^2}{s}. \tag{23}$$

(ii) (Continuity.) Equip $\mathcal{K}$ with the topology of weak convergence against $C([0, C_\Sigma])$ in each coordinate. The solution map $(s, \nu, \tilde\nu) \mapsto (\sigma, \tilde\sigma, \tilde s_1, s_1)$ is continuous on $(0,\infty) \times \mathcal{K}$, and for every compact $I \subset (0,\infty)$ it is uniformly continuous in $s \in I$, uniformly over $(\nu, \tilde\nu) \in \mathcal{K}$.

(iii) (Stability under approximate roots.) Let $I \subset (0,\infty)$ be compact with $s_- \stackrel{\text{def}}{=} \min I$. For every $\epsilon > 0$ there is $\delta > 0$, depending only on $(\epsilon, I, C_\Sigma, \gamma_\star, w)$, with the following property. If $(\nu, \tilde\nu) \in \mathcal{K}$, $s \in I$, and $(\sigma', \tilde\sigma', a', c') \in [0, \gamma_\star C_\Sigma/s_-]^2 \times [0, \gamma_\star C_\Sigma s_2]^2$ satisfy, with $q' = s\sigma'\tilde\sigma'$,

$$\left|\sigma' - \int \frac{x}{s + a'x}\,\nu(dx)\right| \le \delta, \qquad \left|\tilde\sigma' - \int \frac{x}{s + c'x}\,\tilde\nu(dx)\right| \le \delta,$$
$$|a' - s\tilde\sigma' F_w(q')| \le \delta, \qquad |c' - s\sigma' F_w(q')| \le \delta,$$

then $|\sigma' - \sigma| + |\tilde\sigma' - \tilde\sigma| + |a' - \tilde s_1| + |c' - s_1| \le \epsilon$, where $(\sigma, \tilde\sigma, \tilde s_1, s_1)$ is the solution of eq. (21) at $z = -s$ for $(\nu, \tilde\nu)$.

(iv) (Smoothness in s.) Fix $(\nu, \tilde\nu) \in \mathcal{K}$ with $\int x\,\nu(dx) > 0$ and $\int x\,\tilde\nu(dx) > 0$. Then $s \mapsto (\sigma, \tilde\sigma, \tilde s_1, s_1, q)(-s)$ is C^∞ on $(0,\infty)$, and $|q'(s)| \le q(s)/s$.

Full statement and proof: theorem SB.1 of the supplement. Lean: proved: `Muon.DetEquiv.thm_app_de_branch_i`, `Muon.DetEquiv.thm_app_de_branch_i_weight`, *+3 more.*

The fixed-point system of fig. 2 has exactly one nonnegative solution on the negative axis, for every momentum and every spectrum, so the kernels built from it are well defined. Since $\mathcal{K}$ contains the spectral measures at every dimension, continuity uniform over $\mathcal{K}$ makes the dimension dependence of the kernels harmless, and (iii) identifies the limit of the random traces in appendix G.1.

Proof sketch. The system reduces exactly to one equation in q. If ν or $\tilde\nu$ has no mass on $(0, C_\Sigma]$ the solution is explicit, so assume both have. Put $H(x) = \int x\lambda/(s + x\lambda)\,\nu(d\lambda)$, which increases from 0 to the mass p of ν on $(0, C_\Sigma]$, and let X be its inverse; define $\widetilde H, \widetilde X, \tilde p$ from $\tilde\nu$ in the same way. The coupling reads $H(\tilde s_1) = \widetilde H(s_1) = r(q)$ with $r(q) = qF_w(q)$, which is strictly increasing from 0. Hence

$$\tilde s_1 = X(r(q)), \qquad s_1 = \widetilde X(r(q)), \qquad \sigma = \frac{r(q)}{\tilde s_1}, \qquad \tilde\sigma = \frac{r(q)}{s_1},$$

and the equation $q = s\sigma\tilde\sigma$ becomes (eq. (SB.8))

$$\Phi(q) \stackrel{\text{def}}{=} \frac{X(r(q))}{r(q)} \cdot \frac{\widetilde X(r(q))}{r(q)} - \frac{s}{q} = 0, \qquad r(q) < \min(p, \tilde p);$$

conversely every root of Φ gives a solution. Since $H(x)/x$ decreases, $X(\rho)/\rho$ increases, so Φ is strictly increasing. It tends to $-\infty$ as $q \downarrow 0$, where the product stays bounded, and it becomes positive at the right end, where either one factor blows up or $s/q \to 0$. So Φ has exactly one root. The bounds follow from $x/(s + ax) \le C_\Sigma/s$ and $F_w \le s_2$.

Continuity and stability follow from uniqueness by compactness. The class $\mathcal{K}$ is weakly compact (lemma SB.2), so a sequence of approximate roots that stays away from the exact roots would have a limit solving the limit system and differing from its unique solution. This is why (iii) has no rate. The implicit function theorem, applied to the root equation written in logarithmic form, whose q-derivative is at least $1/q$, gives (iv). ◇

With the solution $(\sigma, \tilde{\sigma}, \tilde{s}_1, s_1, q)$ of the weighted system for the EMA weights and $\mathbb{E}\, G_{t+1} = \Sigma_{\text{out}} \Delta_t \Sigma_{\text{in}} / \sqrt{2\mathscr{R}(t)}$, the candidate update and the kernel function are (eqs. (SC.2) to (SC.4))

$$\overline{\widetilde{G}}_\varrho = \frac{2}{\pi} \int_0^\infty T(\varrho + u^2)\, du, \qquad T(s) = s\, L_\beta\big(q(-s)\big)\, \overline{R}(-s)\; \mathbb{E}\, G_{t+1}\, \overline{\widetilde{R}}(-s),$$

$$\mathfrak{d}(\mu, \lambda) = \frac{\sqrt{2}}{\pi} \int_0^\infty \frac{z\, L_\beta(q(-z))\, \mu\lambda}{\big(z + \tilde{s}_1(-z)\mu\big)\big(z + s_1(-z)\lambda\big)}\, du, \qquad z = \varrho + u^2.$$

Here, as in eq. (7), the letter z changes sign: it is the positive number $\varrho + u^2$, and the solution of the system is evaluated at the negative spectral parameter $-z$.

Corollary I.1 (The drift kernel)**.** *Under Assumption S3, let $\widetilde{G}_{t+1} = Q_\varrho(M_t)$ be the frozen-state update and put $(E_\varrho)_{ij} \overset{\text{def}}{=} u_i^\top \big(\mathbb{E}\, \widetilde{G}_{t+1} - \overline{\widetilde{G}}_\varrho\big) v_j$.*

(i) For every pair (i, j) and every $\varrho \in [\varrho_-, \varrho_+]$,

$$\mathbb{E}\left[u_i^\top \widetilde{G}_{t+1} v_j\right] = \frac{\sqrt{2}\, \mathfrak{d}(\mu_i, \lambda_j)}{\sqrt{2\mathscr{R}(t)}}\, (u_i^\top \Delta_t v_j) + (E_\varrho)_{ij} = \frac{\mathfrak{d}(\mu_i, \lambda_j)}{\sqrt{\mathscr{R}(t)}}\, (u_i^\top \Delta_t v_j) + (E_\varrho)_{ij}, \tag{24}$$

and $\sup_{\varrho \in [\varrho_-, \varrho_+]} \big(\sum_{i,j} (E_\varrho)_{ij}^2\big)^{1/2} = \varepsilon_B \to 0$. Consequently the frozen-state drift eq. (SA.19) *of the update $\widetilde{G}_{t+1}$ is*

$$\mathscr{D}_{ij}(t) = \frac{\mathfrak{d}_{ij}}{\sqrt{\mathscr{R}(t)}}\, \mathscr{Q}_{ij}(t) + (u_i^\top \Delta_t v_j)(E_\varrho)_{ij}, \qquad \mathfrak{d}_{ij} \overset{\text{def}}{=} \mathfrak{d}(\mu_i, \lambda_j), \tag{25}$$

and for every bounded function ϕ on $[0, \infty)^2$,

$$\Big| \sum_{i,j} \phi(\mu_i, \lambda_j)\, (u_i^\top \Delta_t v_j)\, (E_\varrho)_{ij} \Big| \le \|\phi\|_\infty\, \|\Delta_t\|_{\text{F}}\, \varepsilon_B. \tag{26}$$

(ii) In the frozen-state model with $\max(N_{\text{out}}, N_{\text{in}}) \le \gamma_\star B$ and $\|\Sigma_{\text{out}}\|, \|\Sigma_{\text{in}}\| \le C_\Sigma$, at every finite dimension and for every $\varrho > 0$, the function $\mathfrak{d}$ is continuous on $[0, \infty)^2$, vanishes when $\mu = 0$ or $\lambda = 0$, and satisfies

$$0 \le \mathfrak{d}(\mu, \lambda) \le \frac{\mu\lambda}{\sqrt{2\varrho}}, \qquad |\partial_\mu \mathfrak{d}(\mu, \lambda)| \le \frac{\lambda}{\sqrt{2\varrho}}, \qquad |\partial_\lambda \mathfrak{d}(\mu, \lambda)| \le \frac{\mu}{\sqrt{2\varrho}}, \tag{27}$$

with one-sided derivatives at 0; in particular, for $\varrho \in [\varrho_-, \varrho_+]$, $\mathfrak{d}$ is Lipschitz on $[0, C_\Sigma]^2$ with constant $C_\Sigma / \sqrt{2\varrho_-}$ in each variable.

Full statement and proof: corollary SC.1 of the supplement. Lean: proved: `Muon.Drift.Erho`, `Muon.Drift.dIdx`, *+5 more.*

This corollary identifies the drift kernel of eq. (6) and justifies eq. (7) at a positive ridge. Under Assumption S3, that is, in the frozen-state model with a diffuse error, the expected update projected on the eigenpair (u_i, v_j) is $\mathfrak{d}_{ij}\, u_i^\top \Delta_t v_j / \sqrt{\mathscr{R}(t)}$ up to an error $(E_\varrho)_{ij}$, so the drift of the mode is the projected risk times the kernel divided by the square root of the risk. The error matrix E_ϱ is small in Frobenius norm over all pairs together, not mode by mode; this is the form eq. (26) in which the dynamics uses it. Part (ii), valid at every dimension, gives the Lipschitz bound used by the deterministic equations. Nothing is claimed as $\varrho \downarrow 0$: the zero-ridge kernel of the main text is the formal specialization.

Proof sketch. The expected update has the deterministic equivalent $\overline{\widetilde{G}}_\varrho$ in Frobenius norm (theorem SC.2, appendix G.2). Writing the regularized polar map as the integral $\frac{2}{\pi} \int_0^\infty (M_t M_t^\top + \varrho + u^2)^{-1} M_t\, du$ of mixed resolvents, removing one sample at a time from the buffer and integrating it out by Stein's identity gives a linear equation whose solution is $T(s)$ up to $o(1)$ in Frobenius norm (theorem SC.1), and a bound of order $1/s$ at large s, valid at every dimension, controls the tail of the integral.

The candidate acts entrywise in the eigenbases, with the kernel as multiplier. Since $\overline{R}(-s)$ is a function of Σ_{out} and $\overline{\widetilde{R}}(-s)$ of Σ_{in}, and $u_i^\top \Sigma_{\text{out}} \Delta_t \Sigma_{\text{in}} v_j = \mu_i \lambda_j \, u_i^\top \Delta_t v_j$,

$$u_i^\top T(s) v_j = \frac{s \, L_\beta(q(-s)) \, \mu_i \lambda_j}{\big(s + \tilde{s}_1(-s)\mu_i\big)\big(s + s_1(-s)\lambda_j\big)} \cdot \frac{u_i^\top \Delta_t v_j}{\sqrt{2\mathscr{R}(t)}}.$$

Integrating at $s = \varrho + u^2$ against $\frac{2}{\pi}\, du = \sqrt{2} \cdot \frac{\sqrt{2}}{\pi}\, du$ gives exactly $u_i^\top \overline{\widetilde{G}}_\varrho v_j = \mathfrak{d}_{ij} \, u_i^\top \Delta_t v_j / \sqrt{\mathscr{R}(t)}$, which is eq. (24). Because the bases are orthonormal, $\sum_{i,j} (E_\varrho)_{ij}^2 = \| \mathbb{E}\, \widetilde{G}_{t+1} - \overline{\widetilde{G}}_\varrho \|_{\text{F}}^2 \le \varepsilon_B^2$. Multiplying by $u_i^\top \Delta_t v_j$ gives eq. (25), and the Cauchy–Schwarz inequality over the pairs gives eq. (26).

The bounds of (ii) are read off the integrand. Since $\tilde{s}_1, s_1 \ge 0$ and $0 \le L_\beta \le 1$, the integrand of eq. (SC.4) lies between 0 and $\mu\lambda/z$, and its μ-derivative $z^2 L_\beta \lambda / ((z + \tilde{s}_1\mu)^2 (z + s_1\lambda))$ lies between 0 and λ/z. With

$$\frac{\sqrt{2}}{\pi} \int_0^\infty \frac{du}{\varrho + u^2} = \frac{1}{\sqrt{2\varrho}},$$

this gives eq. (27); the λ-derivative is symmetric. ◇

In the next corollary $\mathscr{Q}_i = \sum_j \mathscr{Q}_{ij} = \|\Delta_t^\top u_i\|^2$ is the projected row risk, and its drift is the row sum $\sum_j \mathscr{D}_{ij}(t)$ (eq. (SA.21)). The letter $z = \varrho + u^2$ is positive, as in corollary I.1, and ρ_i in (iv) is a measure, not the ridge ϱ.

Corollary I.2 (Half-anisotropic input)**.** *Under Assumption S3, assume* $\Sigma_{\text{in}} = \text{Id}_{N_{\text{in}}}$*, let* $\varrho \in [\varrho_-, \varrho_+]$*, write* $z = \varrho + u^2$*, and put* $\mathfrak{d}_i \overset{\text{def}}{=} \mathfrak{d}(\mu_i, 1)$*.*

(i) For every pair (i, j) *the kernel is* $\mathfrak{d}_{ij} = \mathfrak{d}_i$*, and*

$$\mathfrak{d}_i = \frac{\sqrt{2}\, B}{\pi\, N_{\text{in}}} \int_0^\infty \frac{z \, L_\beta(q(-z)) \, \tilde{\sigma}(-z) \, \mu_i}{z + \tilde{s}_1(-z)\mu_i} \, du.$$

(ii) The row drift of eq. (SA.21) *is, with* E_ϱ *as in corollary I.1,*

$$\sum_j \mathscr{D}_{ij}(t) = \frac{\mathfrak{d}_i}{\sqrt{\mathscr{R}(t)}} \, \mathscr{Q}_i(t) + \sum_j (u_i^\top \Delta_t v_j)(E_\varrho)_{ij}.$$

(iii) At $\beta = 0$*,*

$$\mathfrak{d}_i = \frac{\sqrt{2}\, B}{\pi\, N_{\text{in}}} \int_0^\infty \frac{\tilde{s}_1(-z)\mu_i}{z + \tilde{s}_1(-z)\mu_i} \, du.$$

(iv) At $\beta = 0$*, if* ρ_i *is a probability measure on* $[0, \infty)$ *with*

$$\big(z + \tilde{s}_1(-z)\mu_i\big)^{-1} = \int \frac{\rho_i(dx)}{x + z} \qquad (z > 0),$$

then

$$\mathfrak{d}_i = \frac{B}{\sqrt{2}\, N_{\text{in}}} \int_{[0,\infty)} \frac{x}{\sqrt{x + \varrho}} \, \rho_i(dx).$$

Full statement and proof: corollary SC.2 of the supplement. Lean: proved: `Muon.Drift.dIdx`*,* `Muon.Drift.cor_app_drift_halfaniso_i`*, +4 more.*

This is the drift of the half-anisotropic setting of section 4. When $\Sigma_{\text{in}} = \text{Id}$ every λ_j equals 1, so the kernel $\mathfrak{d}_{ij}$ does not depend on j, and the projected row risks $\mathscr{Q}_i$ have the drift of eq. (14) with $\mathfrak{d}_i$ the common value of the $\mathfrak{d}_{ij}$, not their sum. Part (i) writes $\mathfrak{d}_i$ through the input trace $\tilde{\sigma}$, and (iii) simplifies it at $\beta = 0$. Part (iv) writes it as a moment of a row spectral measure ρ_i; setting $\varrho = 0$ there formally gives the $\sqrt{x}$-moment $\mathfrak{d}_i = \frac{B}{\sqrt{2}\, N_{\text{in}}} \int \sqrt{x}\, \rho_i(dx)$, in the normalization of the main-text kernel eq. (7). Neither the existence of ρ_i nor the limit $\varrho \downarrow 0$ is proved; the phase-comparison chapter adopts the zero-ridge kernels as assumptions.

Proof sketch. With identity input covariance the input trace equation is algebraic. The second equation of eq. (21) becomes $\tilde{\sigma}(-s) = (N_{\text{in}}/B)/(s + s_1(-s))$, so the input factor of the kernel eq. (SC.4) at $\lambda = 1$ is

$$\frac{1}{z + s_1(-z)} = \frac{B\,\tilde{\sigma}(-z)}{N_{\text{in}}},$$

and substituting it gives (i). Part (ii) is the sum over j of eq. (25), since $\mathfrak{d}_{ij} = \mathfrak{d}_i$ and $\mathscr{Q}_i = \sum_j \mathscr{Q}_{ij}$; its error terms, summed over the rows against a bounded function $\psi(\mu_i)$, are controlled by eq. (26) with $\phi(\mu, \lambda) = \psi(\mu)$.

At $\beta = 0$ the source factor equals the coupling. There is one weight, so $L_\beta = F_\beta = J$, and with $q(-z) = z\sigma\tilde{\sigma}$ and the coupling $\tilde{s}_1\sigma = qJ(q)$ of fig. 2,

$$z\,J\big(q(-z)\big)\,\tilde{\sigma}(-z) = \frac{qJ(q)}{\sigma} = \tilde{s}_1(-z),$$

which turns (i) into (iii).

Part (iv) integrates a Stieltjes representation in u. If $(z + \tilde{s}_1(-z)\mu_i)^{-1} = \int \rho_i(dx)/(x + z)$ with ρ_i a probability measure, then the integrand of (iii) is $\tilde{s}_1\mu_i/(z + \tilde{s}_1\mu_i) = \int x\,\rho_i(dx)/(x + z)$, and the scalar integral

$$\frac{2}{\pi}\int_0^\infty \frac{du}{x + \varrho + u^2} = \frac{1}{\sqrt{x + \varrho}}$$

gives (iv) after exchanging the two integrals. ◇

J Volatility

The volatility is the noise that the updates inject into each mode. It is computed in the frozen-state model of appendix G.3, where the updates $\mathsf{S}_{t'}(\varrho) = Q_\varrho(M_{t'})$ are stationary and $\mathsf{S}^{\mathrm{c}}_{t'}(\varrho)$ is the update minus its mean. With momentum, consecutive buffers share most of their batches, so over the B steps of a unit of macroscopic time the quadratic terms accumulate as a long-run variance. For symmetric F, G the lag covariances and their integrated sum are

$$C_\ell[F, G](\varrho, \varrho') = \frac{1}{B}\,\mathbb{E}\,\mathrm{Tr}\left[F\,\mathsf{S}^{\mathrm{c}}_t(\varrho)\,G\,\mathsf{S}^{\mathrm{c}}_{t+\ell}(\varrho')^\top\right],$$
$$\Gamma[F, G](\varrho, \varrho') = C_0[F, G](\varrho, \varrho') + \sum_{\ell \geq 1}\big(C_\ell[F, G](\varrho, \varrho') + C_\ell[F, G](\varrho', \varrho)\big),$$

and $C_\ell = C_\ell[\Sigma_{\text{out}}, \Sigma_{\text{in}}]$, $\Gamma = \Gamma[\Sigma_{\text{out}}, \Sigma_{\text{in}}]$. Their deterministic equivalents $\overline{C}_\ell$, $\overline{\Gamma}$ are given by the lag kernels $\mathfrak{k}_\ell$ of appendix G.3. With P_μ, P'_λ the spectral projections of $\Sigma_{\text{out}}, \Sigma_{\text{in}}$ and r_μ, t_λ their ranks, the forcing measure and the mode and row coefficients are

$$\zeta_\varrho = \sum_{\mu,\lambda} \overline{\Gamma}[P_\mu, P'_\lambda](\varrho, \varrho)\,\delta_{(\mu,\lambda)}, \qquad \mathfrak{v}_{ij}(\varrho) = \frac{B\,\zeta_\varrho(\{(\mu_i, \lambda_j)\})}{r_{\mu_i} t_{\lambda_j}},$$
$$\mathfrak{v}_i(\varrho) = B\,\overline{\Gamma}[u_i \otimes u_i, \mathrm{Id}](\varrho, \varrho) \quad \text{when } \Sigma_{\text{in}} = \mathrm{Id},$$

and $\mathfrak{v}_{ij}$ is the volatility kernel of eq. (6). The two results of this chapter read $\mathfrak{v}_{ij}$ in the two cases the main text uses: the half-anisotropic rows, and no momentum.

Proposition J.1 (Recovery of the rows)**.** *If $\Sigma_{\text{in}} = \mathrm{Id}_{N_{\text{in}}}$, then $\zeta_\varrho = B^{-1}\sum_i \mathfrak{v}_i(\varrho)\,\delta_{(\mu_i,1)}$, and $\mathfrak{v}_{ij}(\varrho) = \mathfrak{v}_i(\varrho)/N_{\text{in}}$, so that $\sum_j \mathfrak{v}_{ij}(\varrho) = \mathfrak{v}_i(\varrho)$ for every i.*

Full statement and proof: proposition SD.4 of the supplement. Lean: proved: `Muon.Volatility.prop_app_vol_half_recovery`.

The proposition is the relation $\mathfrak{v}_i = \sum_j \mathfrak{v}_{ij}$ that the main text uses in the row recursion eq. (14) of the half-anisotropic setting $\Sigma_{\text{in}} = \mathrm{Id}_{N_{\text{in}}}$. There the volatility of row i is the long-run variance of the whole row $\mathsf{S}^\top u_i$ of the update, and its deterministic counterpart $\mathfrak{v}_i(\varrho)$ is defined directly from the row test. The proposition says that this row coefficient is the sum of the mode coefficients of the row, and that all N_{in} modes of a row carry the same share $\mathfrak{v}_i/N_{\text{in}}$. It is an exact identity between deterministic

quantities at every finite dimension; no limit is taken. With the closed form of appendix G.3 it gives the explicit row coefficient $\mathfrak{v}_i = (N_{\text{in}}/B)\big(\mathfrak{c}_0(\mu_i, 1; \varrho) + 2\sum_{\ell\geq 1} \mathfrak{c}_\ell(\mu_i, 1; \varrho)\big)$ (eq. (SD.91)), whose same-step term is $\mathfrak{v}_i^{(0)}(\varrho)$ (corollary SD.4), the deterministic equivalent of the expected row energy of the update (theorem SD.1).

Proof sketch. The deterministic covariances are linear in each of the two tests, so the mass of the forcing measure at a point is the sum of the row coefficients of the rows sitting there. The deterministic lag covariance $\overline{C}_\ell[F, G]$ is a limit of expected traces $B^{-1}\,\mathbb{E}\,\mathrm{Tr}[F\mathsf{S}_t^{\mathrm{c}} G\,\mathsf{S}_{t+\ell}^{\mathrm{c}\,\top}]$, and in the pairing expansion of appendix G.3 each test enters every tree through exactly one factor. Hence $\overline{C}_\ell$ and its lag sum $\overline{\Gamma}$ are linear in F and in G. When $\Sigma_{\text{in}} = \mathrm{Id}$ the input side has the single eigenvalue 1, whose spectral projection is Id, with multiplicity N_{in}. The atom of ζ_ϱ at $(\mu, 1)$ is therefore, splitting the projection P_μ onto the eigenvalue μ of Σ_{out} into its rows,

$$\zeta_\varrho\big(\{(\mu,1)\}\big) = \overline{\Gamma}[P_\mu, \mathrm{Id}](\varrho,\varrho) = \sum_{i:\,\mu_i=\mu} \overline{\Gamma}[u_i \otimes u_i, \mathrm{Id}](\varrho,\varrho) = \frac{1}{B}\sum_{i:\,\mu_i=\mu} \mathfrak{v}_i(\varrho),$$

where the last step is the definition of $\mathfrak{v}_i$. This is the first claim, $\zeta_\varrho = B^{-1}\sum_i \mathfrak{v}_i\,\delta_{(\mu_i,1)}$.

The row coefficient depends on the row only through its eigenvalue, which turns the atom into equal mode coefficients. With the test $u_i \otimes u_i$, the row index enters every tree value only through one normalized trace $B^{-1}\mu_i^p$ with $p \geq 1$, so $\mathfrak{v}_i$ is a function of μ_i alone (theorem SD.7(iv)). The atom at $(\mu, 1)$ is thus $r_\mu \mathfrak{v}_i/B$ for any row i with $\mu_i = \mu$, where r_μ is the multiplicity of μ. The mode coefficient divides the atom equally among the $r_\mu N_{\text{in}}$ modes (i, j) that sit at $(\mu, 1)$ (proposition SD.3), so

$$\mathfrak{v}_{ij}(\varrho) = \frac{B}{r_\mu N_{\text{in}}} \cdot \frac{r_\mu\,\mathfrak{v}_i(\varrho)}{B} = \frac{\mathfrak{v}_i(\varrho)}{N_{\text{in}}}, \qquad \sum_{j=1}^{N_{\text{in}}} \mathfrak{v}_{ij}(\varrho) = \mathfrak{v}_i(\varrho).$$

$\diamond$

Here $r_\beta(q) = qF_\beta(q)$, and b_ℓ, e_ℓ, $\mathfrak{k}_\ell = \mathfrak{k}_\ell[\Sigma_{\text{out}}, \Sigma_{\text{in}}]$ and the mode integrals $\mathfrak{c}_\ell(\mu, \lambda; \varrho)$ are the closed-form objects of appendix G.3; $\overline{\mathfrak{C}}_\ell(\varrho, \varrho)$ is the measure with atoms $\overline{C}_\ell[P_\mu, P'_\lambda](\varrho, \varrho)$. The same-step row and column energies are

$$\mathfrak{v}_i^{(0)}(\varrho) = \frac{\tilde{s}_1(-\varrho)\,\mu_i}{\varrho + \tilde{s}_1(-\varrho)\,\mu_i}, \qquad \widetilde{\mathfrak{v}}_j^{(0)}(\varrho) = \frac{s_1(-\varrho)\,\lambda_j}{\varrho + s_1(-\varrho)\,\lambda_j},$$

and $\overline{Z}_{\mathfrak{k}}(\varrho)$ is $\frac{4}{\pi^2}$ times the double integral of the lag-zero kernel $\mathfrak{k}_0(\varrho + x^2, \varrho + y^2)$ over $x, y > 0$ (eq. (SD.87), where the kernel is written in the variables of one resolvent pair and identified with $\mathfrak{k}_0$ by corollary SD.5(i)).

Remark J.1 (No momentum)**.** *At $\beta = 0$ the only weight is $w_0 = 1$, the buffer is the current gradient, $F_0 = L_0 = J$, $r'_0(q) = \mathbb{E}\; g^2(1+qg^2)^{-2}$ and $F'_0(q) = -\,\mathbb{E}\; g^4(1+qg^2)^{-2}$, and the branch is the no-momentum system of remark SB.1. In the closed form, b_ℓ carries the factor β^ℓ, so $b_\ell = 0$ and $\mathfrak{k}_\ell = 0$ for every $\ell \geq 1$; the same-step scalars are the divided differences $b_0 = \big(r_0(q) - r_0(q')\big)/(q - q')$ and $e_0 = -\big(F_0(q) - F_0(q')\big)/(q - q')$. All lag covariances and their deterministic equivalents vanish for $\ell \geq 1$ (theorem SD.2(ii), theorem SD.4(ii)), so $\Gamma = C_0$, $\overline{\Gamma} = \overline{C}_0$, the forcing measure is $\zeta_\varrho = \overline{\mathfrak{C}}_0(\varrho, \varrho)$, and*

$$\mathfrak{v}_{ij}(\varrho) = \frac{1}{B}\,\mathfrak{c}_0(\mu_i, \lambda_j; \varrho), \qquad \sum_{j=1}^{N_{\text{in}}} \mathfrak{v}_{ij}(\varrho) = \mathfrak{v}_i^{(0)}(\varrho), \qquad \sum_{i=1}^{N_{\text{out}}} \mathfrak{v}_{ij}(\varrho) = \widetilde{\mathfrak{v}}_j^{(0)}(\varrho),$$

by corollaries SD.3 and SD.4, for every pair of covariances. For $\Sigma_{\text{in}} = \mathrm{Id}$ this is $\mathfrak{v}_i(\varrho) = \mathfrak{v}_i^{(0)}(\varrho)$ of theorem SD.7(iv), with the spectral form of remark SD.1. The covariance-weighted same-step total is $\overline{C}_0(\varrho, \varrho) = \overline{Z}_{\mathfrak{k}}(\varrho)$ of corollary SD.5, with the kernel eq. (SD.86) *formed from F'_0 and r'_0.*

Full statement and proof: remark SD.3 of the supplement. Lean: not formalized (remark).

Without momentum the buffer is the current gradient, so updates at different steps are built from disjoint batches and are independent; every lag covariance vanishes and the volatility is the same-step variance of the update. In the closed form of appendix G.3 this is the factor β^ℓ carried by b_ℓ,

and the same-step scalars become divided differences of the no-momentum functions r_0 and F_0. Summed over a row, the same-step term is $\mathfrak{v}_i^{(0)}(\varrho)$ (corollary SD.4), the deterministic equivalent of the expected row energy $\mathbb{E}\,\|\mathsf{S}^\top u_i\|^2$: the exact identity $\mathsf{S}\mathsf{S}^\top = \mathrm{Id} - \varrho(M_t M_t^\top + \varrho\,\mathrm{Id})^{-1}$ reduces it to the deterministic equivalent of the Gram resolvent (theorem SD.1). This is the coefficient behind the main text's statement that without momentum, at $B \geq N$, the update is orthogonal and $\mathfrak{v}_i = 1$. The remark itself is at a positive ridge. As $\varrho \downarrow 0$ the integrand of the row energy tends to the indicator of $(0,\infty)$, whose integral against the spectral measure ρ_i of u_i for $M_t M_t^\top$ is $\rho_i((0,\infty)) = 1 - \|P_0 u_i\|^2$, with P_0 the projection on the left null space of M_t (remark SD.1); this is 1 when the gradient has full row rank. The appendix makes no statement at $\varrho = 0$.

K The isotropic case

In this chapter the data are isotropic, $\Sigma_{\mathrm{in}} = \mathrm{Id}_{N_{\mathrm{in}}}$ and $\Sigma_{\mathrm{out}} = \mathrm{Id}_{N_{\mathrm{out}}}$, as in section 3; $\gamma_{\mathrm{in}} = N_{\mathrm{in}}/B$ and $\gamma_{\mathrm{out}} = N_{\mathrm{out}}/B$ are the two aspect ratios, $\gamma_\wedge$ and $\gamma_\vee$ their minimum and maximum, and $\gamma = N/B$ in the square case. Every eigenvalue is one, so every mode has the same drift kernel $\mathfrak{d}_{\mathrm{iso}}(\varrho)$ and the same share $\mathfrak{v}_{\mathrm{iso}}(\varrho)/(N_{\mathrm{out}}N_{\mathrm{in}})$ of the volatility. The fixed-point system of fig. 2 at $z = -s$ reduces to one scalar equation for its root $q = q(-s)$ (proposition SF.1):

$$s\,q = \big(\gamma_{\mathrm{out}} - r_\beta(q)\big)\big(\gamma_{\mathrm{in}} - r_\beta(q)\big),$$
$$r_\beta(q) \overset{\mathrm{def}}{=} qF_\beta(q) = \sum_{j\geq 0} \Phi(w_j^2 q), \qquad \Phi(v) \overset{\mathrm{def}}{=} \mathbb{E}\,\frac{vg^2}{1+vg^2} \quad (g \sim \mathcal{N}(0,1)).$$

The value $\psi(s) \overset{\mathrm{def}}{=} r_\beta(q(-s)) < \gamma_\wedge$ is the participation level: $B\psi(\varrho)$ is the deterministic equivalent of the squared Frobenius norm of the update $Q_\varrho(M_t)$, which for an exact polar factor would be its rank (theorem SF.2(i)). The drift is the integral eq. (SF.10) of $qL_\beta(q)$ along the root, written $\mathfrak{d}_{\mathrm{iso}}(\varrho;\beta,\gamma_{\mathrm{in}},\gamma_{\mathrm{out}})$ when the arguments matter; its value at zero ridge is

$$C_\beta(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}) \overset{\mathrm{def}}{=} \frac{\sqrt{2}}{\pi\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}} \int_0^\infty q(-u^2)\,L_\beta\big(q(-u^2)\big)\,\mathrm{d}u, \qquad C_\beta(\gamma) \overset{\mathrm{def}}{=} C_\beta(\gamma,\gamma), \qquad C \overset{\mathrm{def}}{=} C_0,$$

and $s_k \overset{\mathrm{def}}{=} \sum_{j\geq 0} w_j^k$, which is $(1-\beta)^k/(1-\beta^k)$ for $\beta > 0$, so that $B/s_2 = B_{\mathrm{eff}}$. The chapter closes the dynamics on the risk, computes the two coefficients of eq. (8) at zero ridge at large and small batch, and ends with a small ridge at small batch.

Proposition K.1 (Zero-ridge isotropic volatility without momentum)**.** *Let* $\beta = 0$ *and* $m_\gamma \overset{\mathrm{def}}{=} \min\{1,\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}\}$. *Then* $\psi(\varrho) \to m_\gamma$ *as* $\varrho \downarrow 0$*, and*

$$\lim_{\varrho\downarrow 0} \mathfrak{v}_{\mathrm{iso}}(\varrho) = B\,m_\gamma = \min\{B, N_{\mathrm{in}}, N_{\mathrm{out}}\}.$$

Full statement and proof: proposition SF.2 of the supplement. Lean: proved: `Muon.Isotropic.prop_app_iso_volatility_zero`*; foundation results:* `GaussianOpNorm`*.*

Without momentum the zero-ridge volatility is $\min\{B, N_{\mathrm{in}}, N_{\mathrm{out}}\}$, the rank of one minibatch gradient. This is the value $\mathfrak{v}_0 = \min\{B, N\}$ of fig. 4, and it gives the learning-rate ratio $N/\sqrt{B}$ at $\beta = 0$ and $B \leq N$ in eq. (11). For an exact polar factor the squared Frobenius norm is the rank, $\|Q_0(G_{t+1})\|_{\mathrm{F}}^2 = \operatorname{rank} G_{t+1}$ (lemma SA.2(i)), so the value is the expected one; the proposition shows that the deterministic coefficients of the regularized algorithm have this limit as $\varrho \downarrow 0$. No statement is made about the algorithm at zero ridge.

Proof sketch. Without momentum the volatility is the same-time value $B\psi(\varrho)$, exactly. The updates of two different steps are built from independent batches, so the lag covariances of the volatility vanish (theorem SD.4(ii)), and the integrated coefficient is the expected squared Frobenius norm of one update. Its deterministic equivalent is read off the anchors of the scalar equation (eq. (SF.18)):

$$\mathfrak{v}_{\mathrm{iso}}(\varrho) = B\,\psi(\varrho) = B\,\Phi\big(q(-\varrho)\big) = B\big(\gamma_{\mathrm{out}} - \varrho\,\sigma(-\varrho)\big).$$

At $\beta = 0$ there is a single weight $w_0 = 1$, so the participation level is one Gaussian average, $\psi = \Phi(q)$.

The participation level has two ceilings. Both factors on the right of the scalar equation are positive, so $\psi < \min(\gamma_{\text{in}}, \gamma_{\text{out}})$ (proposition SF.1(ii)). And $\Phi < 1$, because the integrand $vg^2/(1+vg^2)$ is below one (lemma SF.1(ii)): one batch contributes at most one unit of participation. Hence $\psi(\varrho) < m_\gamma = \min\{1, \gamma_{\text{in}}, \gamma_{\text{out}}\}$.

As the ridge vanishes, the level reaches the lower of the two ceilings. Suppose that $\psi(\varrho_k) \le m_\gamma - \delta$ along a sequence $\varrho_k \downarrow 0$. Then both factors of the scalar equation stay at least δ, and the root is forced to infinity:

$$q(-\varrho_k) = \frac{\big(\gamma_{\text{out}} - \psi(\varrho_k)\big)\big(\gamma_{\text{in}} - \psi(\varrho_k)\big)}{\varrho_k} \ge \frac{\delta^2}{\varrho_k} \longrightarrow \infty.$$

But $\Phi(v) \to 1$ as $v \to \infty$, so $\psi(\varrho_k) = \Phi(q(-\varrho_k)) \to 1 \ge m_\gamma$, a contradiction. Hence $\psi(\varrho) \to m_\gamma$, and $\mathfrak{v}_{\text{iso}}(\varrho) \to B\, m_\gamma = \min\{B, N_{\text{in}}, N_{\text{out}}\}$ because $B\gamma_{\text{in}} = N_{\text{in}}$ and $B\gamma_{\text{out}} = N_{\text{out}}$. With momentum the ceiling one disappears, since r_β is unbounded, and the same-time total tends to $\min\{N_{\text{in}}, N_{\text{out}}\}$ instead (proposition SF.3(i)). ◇

Definition K.1 (Zero-ridge isotropic volatility)**.** *Let $\beta \in [0,1)$ and let $\mathfrak{v}_{\text{iso}}(\varrho)$ be the number of theorem SF.2(ii) at momentum β. The zero-ridge isotropic volatility at momentum β is*

$$\mathfrak{v}_{\text{iso},\beta} \overset{\text{def}}{=} \lim_{\varrho \downarrow 0} \mathfrak{v}_{\text{iso}}(\varrho),$$

the limit being taken at this fixed β, whenever it exists.

Full statement and proof: definition SF.1 of the supplement. Lean: proved: `Muon.Isotropic.def_app_iso_vol_zero`, `Muon.Isotropic.isoVol`, *+1 more; foundation results:* `GaussianOpNorm`.

The definition names the volatility $\mathfrak{v}_\beta$ of the main text's scalar recursion eq. (8) at zero ridge, written $\mathfrak{v}_{\text{iso},\beta}$ here. It is the limit of the deterministic coefficient of theorem SF.2(ii) as $\varrho \downarrow 0$ at a fixed momentum, it depends on B, N_{in} and N_{out}, and it says nothing about the algorithm at zero ridge. At $\beta = 0$ the limit is $\min\{B, N_{\text{in}}, N_{\text{out}}\}$ (proposition K.1). For $\beta > 0$ the supplement proves that the limit exists in the square case, where it equals $N(1 + 2\sum_{\ell \ge 1} \overline{C}_{1,\ell}(0))$ with $\overline{C}_{1,\ell}(0)$ the correlation of two updates ℓ steps apart (proposition SF.3(ii)).

Proposition K.2 (Scalar reduction)**.** *Let $\Sigma_{\text{in}} = \mathrm{Id}_{N_{\text{in}}}$ and $\Sigma_{\text{out}} = \mathrm{Id}_{N_{\text{out}}}$, let $\beta \in [0,1)$, $\varrho > 0$ and $h > 0$, and write $\mathfrak{d}_{\text{iso}}$ and $\mathfrak{v}_{\text{iso}} = \mathfrak{v}_{\text{iso}}(\varrho)$ for the numbers of theorems SF.1 and SF.2.*

(i) The deterministic mode-energy equations eq. (SE.7) *close on the risk: if (Ω_{ij}) solves them on an interval on which $\mathcal{R} > 0$, then $\mathcal{R} = \frac{1}{2}\sum_{i,j} \Omega_{ij}$ solves*

$$\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\tau} = -2h\, \mathfrak{d}_{\text{iso}} \sqrt{\mathcal{R}} + \frac{h^2}{2B}\, \mathfrak{v}_{\text{iso}}, \qquad \mathcal{R}(0) = \mathscr{R}(0), \tag{28}$$

and eq. (28) *has at most one solution with $\mathcal{R} > 0$ on any interval $[0,T]$.*

(ii) Under Assumption S3 and for $\varrho \in [\varrho_-, \varrho_+]$, one step of size $\eta \ge 0$ from a frozen state, $\Delta^+ = \Delta_t - \eta\, Q_\varrho(M_t)$, satisfies

$$\begin{aligned} \mathbb{E}\, \mathscr{R}(\Delta^+) - \mathscr{R}(t) &= -2\eta\, \mathfrak{d}_{\text{iso}} \sqrt{\mathscr{R}(t)} + \frac{\eta^2 B}{2}\, \psi(\varrho) + r_t, \\ |r_t| &\le \eta \sqrt{2\mathscr{R}(t)}\, \varepsilon_B + \frac{\eta^2 B}{2}\, \gamma_\star \varrho_+ e_B, \end{aligned} \tag{29}$$

with ε_B of theorem SC.2 and e_B of theorem SD.1.

Full statement and proof: proposition SF.4 of the supplement. Lean: proved: `Muon.Isotropic.IsIsoScalarOde`, `Muon.Isotropic.prop_app_iso_scalar_i`, *+2 more; foundation results:* `GaussianOpNorm`.

With isotropic data the deterministic mode-energy equations close on the risk. Part (i) is the continuous-time form of the main text's recursion eq. (8): with $\eta = h/B$ and $t = B\tau$ it reads $\mathrm{d}\mathcal{R}/\mathrm{d}t = -2\eta\, \mathfrak{d}_{\text{iso}}\sqrt{\mathcal{R}} + \frac{\eta^2}{2}\mathfrak{v}_{\text{iso}}$, so the drift and volatility of the main text are $\mathfrak{d} = \mathfrak{d}_{\text{iso}}$ and $\mathfrak{v} = \mathfrak{v}_{\text{iso}}$. Part (ii) is the balance for one step from a fixed state. Its quadratic term is the same-time value $B\psi(\varrho)$,

which equals $\mathfrak{v}_{\mathrm{iso}}$ at $\beta = 0$ and differs from it by the lag terms for $\beta > 0$; over many steps the lag terms are added (corollary K.1).

Proof sketch. Every mode has the same drift kernel, and the expected update is a multiple of the error itself. At $\mu_i = \lambda_j = 1$ the denominator of the drift kernel eq. (7) is $(z + \tilde{s}_1)(z + s_1) = z\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}/q$ by the anchors of the scalar equation, so the kernel loses its dependence on the mode and becomes an integral along the root (theorem SF.1(i),(iii)):

$$\mathfrak{d}_{\mathrm{iso}} = \frac{\sqrt{2}}{\pi\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}} \int_0^\infty q(-z)\, L_\beta\big(q(-z)\big)\,\mathrm{d}u \quad (z = \varrho + u^2), \qquad \mathbb{E}\, Q_\varrho(M_t) = \frac{\mathfrak{d}_{\mathrm{iso}}}{\sqrt{\mathscr{R}(t)}}\,\Delta_t + E_\varrho,$$

with $\|E_\varrho\|_{\mathrm{F}} \le \varepsilon_B \to 0$. The volatility is shared equally, $\mathfrak{v}_{ij} = \mathfrak{v}_{\mathrm{iso}}/(N_{\mathrm{out}}N_{\mathrm{in}})$ (theorem SF.2(ii)). Since $\mu_i\lambda_j = 1$, the risk is $\mathcal{R} = \frac{1}{2}\sum_{i,j}\mathcal{Q}_{ij}$, and summing the mode-energy equations eq. (SE.7) over the pairs gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\big(2B^2\mathcal{R}\big) = -\frac{2h\,\mathfrak{d}_{\mathrm{iso}}}{\sqrt{\mathcal{R}}}\,2B^2\mathcal{R} + h^2B\,\mathfrak{v}_{\mathrm{iso}},$$

which is eq. (28) after division by $2B^2$. The right side is continuously differentiable in $\mathcal{R} > 0$, so the Picard–Lindelöf theorem gives uniqueness. This is (i).

For one step from a fixed state the change of the risk is exactly a linear term, the pairing of the error with the mean update, plus half the second moment of the update (proposition SC.2). By the display above, the linear term is $-\eta\langle\Delta_t, \mathbb{E}\, Q_\varrho(M_t)\rangle = -2\eta\,\mathfrak{d}_{\mathrm{iso}}\sqrt{\mathscr{R}(t)}$ up to $\eta\sqrt{2\mathscr{R}(t)}\,\varepsilon_B$, by $\|\Delta_t\|_{\mathrm{F}}^2 = 2\mathscr{R}(t)$ and the Cauchy–Schwarz inequality. For the quadratic term, the regularized polar map of a matrix K satisfies, exactly (eq. (SD.27)),

$$\|Q_\varrho(K)\|_{\mathrm{F}}^2 = \mathrm{Tr}\; KK^\top\big(KK^\top + \varrho\,\mathrm{Id}\,\big)^{-1} = N_{\mathrm{out}} - \varrho\;\mathrm{Tr}\,\big(KK^\top + \varrho\,\mathrm{Id}\,\big)^{-1}.$$

For the normalized buffer, the deterministic equivalent of the resolvent trace divided by B is $\sigma(-\varrho)$, so the second moment is $B(\gamma_{\mathrm{out}} - \varrho\,\sigma(-\varrho)) = B\psi(\varrho)$ up to $B\gamma_\star\varrho_+ e_B$ (theorem SD.1, theorem SF.2(i)). This is (ii). ◇

Corollary K.1 (Isotropic dynamics)**.** *Let* $\Sigma_{\mathrm{in}} = \mathrm{Id}_{N_{\mathrm{in}}}$ *and* $\Sigma_{\mathrm{out}} = \mathrm{Id}_{N_{\mathrm{out}}}$*, fix* $h > 0$*,* $\varrho > 0$*,* $\beta \in [0, 1)$*,* $\gamma_\star < \infty$*,* $E_\star < \infty$*,* $c_M < \infty$ *and* $0 < r_0 \le r_1 < \infty$*, and along a sequence of dimensions with* $B \to \infty$ *assume:*

- *(a) Assumption 3 holds and* $\max(N_{\mathrm{in}}, N_{\mathrm{out}}) \le \gamma_\star B$*;*
- *(b) the algorithm of definition F.1 is run with momentum* β*, normalized ridge* ϱ *and step* $\eta_t = h/B$*;*
- *(c)* W_0 *is deterministic,* $\|\Delta_0\|_{\mathrm{F}}^2 \le E_\star$*,* $r_0 \le \mathscr{R}(0) \le r_1$ *and* $\|\Delta_0\|/\|\Delta_0\|_{\mathrm{F}} \to 0$*;*
- *(d) the initial buffer* M_{-1}° *is independent of all batches and* $\mathbb{E}\,\|M_{-1}^\circ\|_{\mathrm{F}}^2 \le c_M B$*.*

Let $T_\star$ *be given by eq.* (SE.10) *at* $C_\Sigma = 1$*. Then eq.* (28) *has a unique solution* $\mathcal{R}$ *on* $[0, T_\star]$*, and*

- *(i) with probability tending to one,* $t_{\mathrm{stop}} > BT_\star$ *and* $\inf_{t \le BT_\star}\mathscr{R}(t) \ge r_0/2$*;*
- *(ii)* $\sup_{\tau \le T_\star}|\mathscr{R}(\lfloor B\tau\rfloor) - \mathcal{R}(\tau)| \to 0$ *in probability;*
- *(iii) on the event of (i),* $\sup_{t \le BT_\star}\|\Delta_t\|/\|\Delta_t\|_{\mathrm{F}} \to 0$ *in probability.*

Full statement and proof: corollary SF.1 of the supplement. Lean: proved: `Muon.Isotropic.cor_app_iso_dynamics_scalar`*,* `Muon.Isotropic.cor_app_iso_dynamics_i`*, +3 more; foundation results:* `GaussianOpNorm`*,* `GaussianOpNormL6`*,* `GaussianPoincare`*,* `GaussianQuadFormLp`*.*

The corollary is the precise form of the main text's claim that, with isotropic data, the risk of MUON follows the scalar recursion eq. (8). At the step $\eta = h/B$ and up to the macroscopic time $T_\star$, the risk at step $\lfloor B\tau\rfloor$ is uniformly close to the solution of eq. (28), whose coefficients are the integrated volatility $\mathfrak{v}_{\mathrm{iso}}(\varrho)$ and the drift $\mathfrak{d}_{\mathrm{iso}}(\varrho)$ at the fixed ridge $\varrho > 0$. Hypothesis (c) asks that the initial error be diffuse: its operator norm is a vanishing fraction of its Frobenius norm, so no single singular direction carries a fixed fraction of the risk. Part (iii) says that this persists. The time $T_\star > 0$ does

not depend on the dimension. The corollary does not reach the zero-ridge coefficients C_β and $\mathfrak{v}_{\mathrm{iso},\beta}$, which enter only as limits of the coefficients.

Proof sketch. The corollary is theorem H.1 at identity covariances. The covariance bound of (H1) holds with $C_\Sigma = 1$, the weighted pairing is the Frobenius pairing, and the normalized error is $\Delta_t/\|\Delta_t\|_{\mathrm{F}}$, so the diffuseness of (H3) is the ratio $\|\Delta_0\|/\|\Delta_0\|_{\mathrm{F}}$ of hypothesis (c) (eq. (SF.2)). The mode-energy equations have a unique solution on $[0, T_\star]$ with risk at least $r_0/2$ (proposition SE.1), and by proposition K.2(i) its risk solves eq. (28), which has only one positive solution. The spectral energy measure is the single atom $\|\Delta\|_{\mathrm{F}}^2\,\delta_{(1,1)} = 2\mathscr{R}(\Delta)\,\delta_{(1,1)}$, so the measure statement of the theorem is the risk statement (ii).

In the isotropic case the block argument behind the theorem (appendix G.4) is a computation for the risk alone. The algorithm is compared, over blocks of $L = \lfloor B^{1/8} \rfloor$ steps, with the model in which the state is held at its value D at the start of the block and the momentum buffer is a stationary average of independent batches. Test the exact block expansion eq. (SE.28) of lemma SE.26 against the constant function. The test is then $\|\Delta\|_{\mathrm{F}}^2 = 2\mathscr{R}(\Delta)$. Its drift term is the pairing of D with the mean update, which is $\mathfrak{d}_{\mathrm{iso}} D/\sqrt{\mathscr{R}(D)}$ up to a vanishing error (theorem SF.1(iii)). Its volatility term is the total mass $\mathfrak{v}_{\mathrm{iso}}/B$ of the forcing measure (eq. (SF.16)). Halving gives

$$\mathbb{E}_t\left[\mathscr{R}(\Delta_{t+L})\right] - \mathscr{R}(D) = \frac{L}{B}\Big(- 2h\,\mathfrak{d}_{\mathrm{iso}}\sqrt{\mathscr{R}(D)} + \frac{h^2}{2B}\,\mathfrak{v}_{\mathrm{iso}}\Big) + \frac{L}{B}\,\epsilon_t,$$

where the mean of $|\epsilon_t|$ tends to zero, uniformly over the blocks that start at a diffuse state whose risk and energy stay within the bounds of the hypotheses (proof of lemma SE.26). The main term is L/B times the right side of eq. (28). Its volatility is the integrated coefficient, not the same-time value $B\psi(\varrho)$ of eq. (29). A block is long compared with the memory $1/(1-\beta)$ of the buffer, so the covariance of the sum of its L updates contains the correlations of all lags (theorem SD.8(iv)). The first steps of a block, during which the buffer still remembers the previous block, cost a total of order $\log^2(2B)/L$, and the fluctuations of the block increments form a martingale whose maximum vanishes.

The error stays diffuse, which is what keeps the mean update close to $\mathfrak{d}_{\mathrm{iso}} D/\sqrt{\mathscr{R}(D)}$ along the trajectory. The operator norm of the mean update is at most a constant times the operator norm of the normalized error, plus a vanishing term, at every state, diffuse or not (lemma SE.7(iv)), so the argument is not circular. The centered increments of the error over a block have operator norm $O(L/B)$, and their partial sums vanish in operator norm by a matrix martingale inequality (lemma SE.23). A discrete Gronwall inequality then bounds $\max_{t\le BT_\star}\|\Delta_t\|$ by a constant times $\|\Delta_0\|$ plus a vanishing term, while the Frobenius norm stays of order one because the risk stays above $r_0/2$; this is (iii). Gronwall's inequality for eq. (28) and routine estimates give (i) and (ii). ◇

Proposition K.3 (The zero-ridge constant)**.** *Let $\beta \in [0,1)$ and $\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}} > 0$. Put $K_0 \overset{\mathrm{def}}{=} 1$, and for $\beta > 0$ let K_β be the unique $q > 0$ with $qF_\beta(q) = \min(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}})$.*

(i) For every $t > 0$,

$$0 < q(-t)\,L_\beta\big(q(-t)\big) \le \min\Big(K_\beta, \frac{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}{t}\Big),$$

so the integral eq. (SF.21) *converges and $0 < C_\beta(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}}) < \infty$. Moreover $C_\beta(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}}) = C_\beta(\gamma_{\mathrm{out}}, \gamma_{\mathrm{in}})$, and*

$$C_\beta(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}}) = \frac{1}{\sqrt{2}\,\pi\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}\int_0^\infty t^{-1/2}\,q(-t)\,L_\beta\big(q(-t)\big)\,\mathrm{d}t. \tag{30}$$

(ii) Let $\mathfrak{d}_{\mathrm{iso}}(\varrho)$ be the kernel eq. (SF.10) *at momentum β, for the measures eq.* (SF.1)*. Then, at this fixed β, $\mathfrak{d}_{\mathrm{iso}}(\varrho) \to C_\beta(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}})$ as $\varrho \downarrow 0$.*

Full statement and proof: proposition SF.5 of the supplement. Lean: proved: `Muon.Isotropic.prop_app_iso_zero_i`, `Muon.Isotropic.prop_app_iso_zero_ii`.

The proposition makes sense of the zero-ridge drift $C_\beta(\gamma_{\mathrm{in}}, \gamma_{\mathrm{out}})$ of eq. (SF.21), which is the drift $\mathfrak{d}_\beta$ of the main text's recursion eq. (8) and the constant of eq. (9): the integral converges, the constant is positive and symmetric in the two aspect ratios, and it is the limit of the drift kernel $\mathfrak{d}_{\mathrm{iso}}(\varrho)$ as $\varrho \downarrow 0$

at a fixed momentum. Like every zero-ridge statement of the chapter, (ii) concerns the deterministic kernel and not the algorithm at zero ridge. The form eq. (30) in the spectral parameter $t = u^2$ is the one used for the batch limits. At $\beta = 0$ the integrand is the participation level, $qJ(q) = \psi$, so C_0 is an average over the spectral scale of the effective rank of the update divided by B.

Proof sketch. The integrand is bounded near the origin and decays like $1/t$, which gives convergence. Since $J \le 1$ and the weights sum to one, $L_\beta \le 1$ and $qL_\beta(q) \le q$. Two bounds on the root follow from the scalar equation. The trace equations give $\sigma(-t) \le \gamma_{\text{out}}/t$ and $\tilde\sigma(-t) \le \gamma_{\text{in}}/t$, so $q(-t) = t\,\sigma(-t)\tilde\sigma(-t) \le \gamma_{\text{in}}\gamma_{\text{out}}/t$. And the participation level stays below $\min(\gamma_{\text{in}}, \gamma_{\text{out}})$, so for $\beta > 0$ the root stays below the root K_β at which r_β reaches that level, because r_β is increasing; at $\beta = 0$ the integrand is $\Phi(q) < 1$ directly. Together,

$$0 < q(-t)\, L_\beta\big(q(-t)\big) \le \min\Big(K_\beta, \frac{\gamma_{\text{in}}\gamma_{\text{out}}}{t}\Big),$$

so the integrand of eq. (SF.21) at $t = u^2$ is at most $\min(K_\beta, \gamma_{\text{in}}\gamma_{\text{out}}/u^2)$, which is integrable. The scalar equation is symmetric in $(\gamma_{\text{in}}, \gamma_{\text{out}})$, so the root and the constant are too, and the substitution $t = u^2$ gives eq. (30).

The ridge enters only as a shift of the spectral parameter, so the limit is a limit under the integral. By theorem SF.1(i) the kernel at ridge ϱ is the same integral with the root evaluated at $-(\varrho + u^2)$. The root is continuous in the spectral parameter (theorem I.1(iv)) and L_β is continuous, so the integrand converges pointwise as $\varrho \downarrow 0$. The bound above at $t = \varrho + u^2$ dominates it by the same function $\min(K_\beta, \gamma_{\text{in}}\gamma_{\text{out}}/u^2)$, and dominated convergence gives (ii). The limit is taken at a fixed β; at zero ridge the limits $\beta \downarrow 0$ and $\varrho \downarrow 0$ do not commute for $\gamma > 1$ (proposition SF.12). ◇

For a finite measure ν on $[0, \infty)$ write $S_\nu(-t) \stackrel{\text{def}}{=} \int \nu(\mathrm{d}x)/(x + t)$ for $t > 0$. At $\beta = 0$ the root solutions $\sigma(-t)$, $\tilde\sigma(-t)$ of the fixed-point system of fig. 2 are the transforms of a unique pair of measures $\mathfrak{m}_{\text{out}}, \mathfrak{m}_{\text{in}}$ on $[0, \infty)$ of masses $\gamma_{\text{out}}, \gamma_{\text{in}}$, the isotropic measure pair. For $r > 0$, M_r is the Marchenko–Pastur law with an atom $(1 - 1/r)_+$ at zero and density $\sqrt{(b_r - x)(x - a_r)}/(2\pi x)$ on (a_r, b_r), where $a_r = (1 - \sqrt{r})^2/r$ and $b_r = (1 + \sqrt{r})^2/r$.

Theorem K.1 (Small aspect ratios)**.** *Let $\gamma_{\text{in}}^{(j)}, \gamma_{\text{out}}^{(j)} \to 0$ with $r_j \stackrel{\text{def}}{=} \gamma_{\text{out}}^{(j)}/\gamma_{\text{in}}^{(j)} \to r \in (0, \infty)$, and let $(\mathfrak{m}_{\text{out}}^{(j)}, \mathfrak{m}_{\text{in}}^{(j)})$ be the isotropic measure pair at $(\gamma_{\text{in}}^{(j)}, \gamma_{\text{out}}^{(j)})$ of proposition SF.6. Let P_j be the image of the probability measure $\mathfrak{m}_{\text{out}}^{(j)}/\gamma_{\text{out}}^{(j)}$ under $\lambda \mapsto \lambda/\gamma_{\text{out}}^{(j)}$, and write $\mathsf{m}_j(t) \stackrel{\text{def}}{=} S_{\mathsf{P}_j}(-t)$.*

(i) For every $t > 0$,

$$\Big| t\,\mathsf{m}_j(t)^2 + \Big(t + \frac{1}{r_j} - 1\Big)\mathsf{m}_j(t) - 1 \Big| \le \frac{3\gamma_{\text{out}}^{(j)}}{r_j^2 t^2}. \tag{31}$$

(ii) $\mathsf{P}_j \to \mathsf{M}_r$ weakly, and $\int \sqrt{x}\,\mathsf{P}_j(\mathrm{d}x) \to J_r \stackrel{\text{def}}{=} \int \sqrt{x}\,\mathsf{M}_r(\mathrm{d}x) > 0$.

(iii) The constant satisfies

$$\begin{aligned} C\big(\gamma_{\text{in}}^{(j)}, \gamma_{\text{out}}^{(j)}\big) &= \frac{r_j}{\sqrt{2\gamma_{\text{out}}^{(j)}}} \int \sqrt{x}\,\mathsf{P}_j(\mathrm{d}x) \sim \frac{r}{\sqrt{2\gamma_{\text{out}}^{(j)}}}\, J_r, \\ J_r &= \int_{a_r}^{b_r} \frac{\sqrt{(b_r - x)(x - a_r)}}{2\pi\sqrt{x}}\, \mathrm{d}x. \end{aligned} \tag{32}$$

In particular $J_1 = 8/(3\pi)$, and $C(\gamma) \sim 8/(3\pi\sqrt{2\gamma})$ as $\gamma \downarrow 0$.

Full statement and proof: theorem SF.4 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_small_i`, `Muon.Isotropic.thm_app_iso_small_ii`, *+1 more.*

The theorem gives the zero-ridge drift without momentum at large batch, $N \ll B$. In the square case $C(\gamma) \sim \frac{8}{3\pi}(2\gamma)^{-1/2}$, that is $\mathfrak{d}_0 \approx \frac{8}{3\pi}\sqrt{B/(2N)}$, the large-batch limit of the approximate form of $\mathfrak{d}_0$ in eq. (9). The constant J_r is the mean square root of the Marchenko–Pastur law M_r of ratio $r = \gamma_{\text{out}}/\gamma_{\text{in}}$, with its atom at zero when $r > 1$, so the drift constant is a mean singular value. Part (i) is a quantitative form of the convergence that (ii) states.

Proof sketch. Without momentum the drift constant is a half-moment of a spectral measure. The isotropic measure pair exists and is unique (proposition SF.6): it is obtained as a limit of the expected spectral measures of the Gram matrices of one minibatch gradient, and the transforms identify the limit. The integrand of the drift is then $\psi(t) = \int \frac{\lambda}{\lambda+t}\,\mathfrak{m}_{\mathrm{out}}(\mathrm{d}\lambda)$, and the elementary integral $\int_0^\infty t^{-1/2}\frac{x}{x+t}\,\mathrm{d}t = \pi\sqrt{x}$ turns the form eq. (30) into a mean of square roots, that is of singular values (proposition SF.7):

$$C(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}) = \frac{1}{\sqrt{2}\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}\int \sqrt{\lambda}\;\mathfrak{m}_{\mathrm{out}}(\mathrm{d}\lambda) = \frac{r_j}{\sqrt{2\gamma_{\mathrm{out}}}}\int \sqrt{x}\;\mathsf{P}_j(\mathrm{d}x),$$

the second form after the rescaling $\lambda = \gamma_{\mathrm{out}}x$ that defines P_j.

At small aspect ratio the closure becomes the Marchenko–Pastur quadratic. The transform of P_j is $\mathsf{m}_j(t) = \sigma(-\gamma_{\mathrm{out}}t)$, and both transforms are bounded by their mass over the spectral parameter, so the argument v_j of Φ in the closure is at most $\gamma_{\mathrm{out}}/(r_j t)$, which is small. Without momentum the closure reads $\gamma_{\mathrm{out}}(1 - t\,\mathsf{m}_j) = \Phi(v_j)$, and Φ is linear up to a quadratic error, $0 \le v - \Phi(v) \le 3v^2$ (lemma SF.1(iii)). Inserting v_j and dividing by γ_{out} leaves the quadratic equation of eq. (31) with the residual $3\gamma_{\mathrm{out}}/(r_j^2t^2)$. This is (i).

The rescaled laws converge to M_r, and so do their half-moments. For each t, every limit point of $\mathsf{m}_j(t)$ is a positive root of $t\,m^2 + (t + \frac{1}{r} - 1)m - 1 = 0$, and that root is the transform of M_r (lemma SF.9). The first moments are exactly $1/r_j$, because the first moment of $\mathfrak{m}_{\mathrm{out}}$ is $\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}$ (lemma SF.3(ii)). Convergence of the transforms on the negative axis then gives weak convergence, and the bounded first moments give convergence of the half-moments (lemma SF.2(i)). This is (ii), and (iii) follows from the display. At $r = 1$, $J_1 = \frac{1}{2\pi}\int_0^4 \sqrt{4-x}\,\mathrm{d}x = \frac{8}{3\pi}$. ◇

The linearized closure replaces the participation count $r_\beta(q)$ by q. Its root is $Y_{\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}}(s) = 2\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}/\big(\gamma_{\mathrm{in}} + \gamma_{\mathrm{out}} + s + \sqrt{(\gamma_{\mathrm{in}} + \gamma_{\mathrm{out}} + s)^2 - 4\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}\big)$, the root below $\gamma_\wedge$ of $sY = (\gamma_{\mathrm{out}} - Y)(\gamma_{\mathrm{in}} - Y)$, and its drift is

$$\mathfrak{d}_{\mathrm{lin}}(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}};\varrho) \overset{\mathrm{def}}{=} \frac{\sqrt{2}}{\pi\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}\int_0^\infty Y_{\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}}(\varrho + v^2)\,\mathrm{d}v \qquad (\varrho \ge 0),$$

the drift integral eq. (SF.10) with $qL_\beta(q)$ replaced by the linear root. At $\varrho = 0$, $\mathfrak{d}_{\mathrm{iso}}(0;\beta,\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}) \overset{\mathrm{def}}{=} C_\beta(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}})$.

Theorem K.2 (Small aspect ratios with momentum)**.** *Let $\beta \in [0,1)$ and $\gamma_{\mathrm{in}},\gamma_{\mathrm{out}} > 0$, and assume $\varepsilon_\beta \overset{\mathrm{def}}{=} 6\gamma_\wedge(1-\beta^2) \le \frac{1}{2}$.*

(i) For every $\varrho \ge 0$ there is $\vartheta \in \big[1-\varepsilon_\beta, (1-\varepsilon_\beta)^{-1/2}\big]$ with

$$\mathfrak{d}_{\mathrm{iso}}(\varrho;\beta,\gamma_{\mathrm{in}},\gamma_{\mathrm{out}}) = \vartheta\;\mathfrak{d}_{\mathrm{lin}}(s_2\gamma_{\mathrm{in}}, s_2\gamma_{\mathrm{out}};\varrho).$$

(ii) With $r \overset{\mathrm{def}}{=} \gamma_{\mathrm{out}}/\gamma_{\mathrm{in}}$ and J_r of theorem K.1,

$$\mathfrak{d}_{\mathrm{lin}}(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}};0) = \frac{r\,J_r}{\sqrt{2\gamma_{\mathrm{out}}}}, \qquad \mathfrak{d}_{\mathrm{lin}}(\gamma,\gamma;0) = \frac{8}{3\pi\sqrt{2\gamma}}.$$

In particular, in the square case $C_\beta(\gamma) = 8\vartheta/(3\pi\sqrt{2s_2\gamma})$ with ϑ of (i) at $\varrho = 0$.

(iii) If $6\gamma_\wedge \le \frac{1}{2}$, then for every $\beta \in [0,1)$

$$\big(1 - 6\gamma_\wedge\big)^{3/2} \le \frac{C_\beta(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}})}{s_2^{-1/2}\,C_0(\gamma_{\mathrm{in}},\gamma_{\mathrm{out}})} \le \big(1 - 6\gamma_\wedge\big)^{-3/2}.$$

(iv) For every $\varrho \ge 0$,

$$(1-\varepsilon_\beta)\big(1 - 6s_2\gamma_\wedge\big)^{1/2} \le \frac{\mathfrak{d}_{\mathrm{iso}}(\varrho;\beta,\gamma_{\mathrm{in}},\gamma_{\mathrm{out}})}{\mathfrak{d}_{\mathrm{iso}}(\varrho;0,s_2\gamma_{\mathrm{in}},s_2\gamma_{\mathrm{out}})} \le (1-\varepsilon_\beta)^{-1/2}\big(1 - 6s_2\gamma_\wedge\big)^{-1}.$$

Full statement and proof: theorem SF.7 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_small_momentum_i`, `Muon.Isotropic.thm_app_iso_small_momentum_ii`, *+2 more.*

At large batch, momentum acts on the drift as the effective batch $B_{\text{eff}} = B/s_2 = B(1+\beta)/(1-\beta)$ of the main text: up to the factor ϑ, the drift is that of the linearized closure at the aspect ratios $(s_2\gamma_{\text{in}}, s_2\gamma_{\text{out}}) = (N_{\text{in}}/B_{\text{eff}}, N_{\text{out}}/B_{\text{eff}})$. In the square case $C_\beta(\gamma) \approx \frac{8}{3\pi}\sqrt{B_{\text{eff}}/(2N)}$, the large-batch form of $\mathfrak{d}_\beta$ in eq. (9). The small parameter is $\gamma_\wedge(1-\beta^2)$, of the order of N/B_{eff}, so the theorem also covers a fixed aspect ratio as $\beta \uparrow 1$.

Proof sketch. At small aspect ratio the root stays where the participation count is linear, with slope $s_2 = \sum_j w_j^2$. Expanding each term $\Phi(w_j^2 q)$ to second order gives, for all $q \geq 0$ (eq. (SF.49)),

$$s_2 q\Big(1 - 3\frac{s_4}{s_2}\,q\Big) \leq r_\beta(q) \leq s_2 q, \qquad 1 - 3s_3 q \leq L_\beta(q) \leq 1.$$

The level $r_\beta(q)$ stays below $\gamma_\wedge$, and $r_\beta(2\gamma_\wedge/s_2) \geq \gamma_\wedge$ once $\varepsilon_\beta \leq \frac{1}{2}$, so the root stays below $\bar{q} = 2\gamma_\wedge/s_2$. At such roots the two expansions are accurate to relative errors $6\gamma_\wedge s_4/s_2^2$ and $6\gamma_\wedge s_3/s_2$, both at most ε_β, because $s_4/s_2^2 = (1-\beta^2)/(1+\beta^2)$.

A larger participation count produces a smaller root, so the root is squeezed between two linear roots. The right side $(\gamma_{\text{out}} - r)(\gamma_{\text{in}} - r)$ of the scalar equation decreases in the count r, and comparing the two sides shows that raising the count lowers the root (lemma SF.5). For the linear count $r(q) = cq$ the root is $c^{-1}Y(s/c)$, where $Y(s)$ is the explicit root of $sY = (\gamma_{\text{out}} - Y)(\gamma_{\text{in}} - Y)$. The two expansions give

$$\frac{1}{s_2}\,Y\Big(\frac{s}{s_2}\Big) \leq q(-s) \leq \frac{1}{c}\,Y\Big(\frac{s}{c}\Big), \qquad c = s_2\Big(1 - 6\gamma_\wedge\frac{s_4}{s_2^2}\Big).$$

Inserting this sandwich and the bounds on L_β in the drift integral eq. (SF.10), and rescaling $u = \sqrt{s_2}\,v$, gives (i): the factor s_2 in the count is absorbed by multiplying both aspect ratios by s_2, because $Y_{t\gamma_{\text{in}}, t\gamma_{\text{out}}}(ts) = t\,Y_{\gamma_{\text{in}},\gamma_{\text{out}}}(s)$ (eq. (SF.50)).

The linear drift at zero ridge is a half-moment of the Marchenko–Pastur law, as in theorem K.1. The rescaled linear root $y(x) = Y(\gamma_{\text{out}}x)/\gamma_{\text{out}}$ solves $xy = (1-y)(\frac{1}{r}-y)$, which is the equation satisfied by $1 - x\,m_r(x)$ for the transform m_r of M_r, and the half-moment formula gives $\mathfrak{d}_{\text{lin}}(\gamma_{\text{in}}, \gamma_{\text{out}}; 0) = rJ_r/\sqrt{2\gamma_{\text{out}}}$. In the square case $y(v^2) = e^{-2\theta}$ at $v = 2\sinh\theta$, so $\int_0^\infty y(v^2)\,\mathrm{d}v = \int_0^\infty (e^{-\theta} + e^{-3\theta})\,\mathrm{d}\theta = \frac{4}{3}$, which gives $8/(3\pi\sqrt{2\gamma})$. This is (ii). Parts (iii) and (iv) divide (i) at momentum β by (i) at momentum zero, at the aspect ratios $(\gamma_{\text{in}}, \gamma_{\text{out}})$ and $(s_2\gamma_{\text{in}}, s_2\gamma_{\text{out}})$ respectively. ◇

For $n \geq 0$, $0 \leq z \leq 1$ and $\beta \in (0,1)$ put

$$k_n \stackrel{\text{def}}{=} \frac{1}{2n+1} + \frac{1}{2n+3}, \qquad \Omega(z) \stackrel{\text{def}}{=} \frac{4}{\pi^2}\sum_{n\geq 0} k_n^2\, z^{2n+1}, \qquad V_\beta \stackrel{\text{def}}{=} 1 + 2\sum_{\ell \geq 1}\Omega(\beta^\ell).$$

Lemma K.1 (The volatility factor)**.** $\Omega(1) = 1$, *and for* $\beta \in (0,1)$

$$V_\beta = \frac{4}{\pi^2}\sum_{n\geq 0} k_n^2\,\frac{1+\beta^{2n+1}}{1-\beta^{2n+1}}, \qquad 1 < V_\beta < \frac{1}{s_2} = \frac{1+\beta}{1-\beta},$$

$$\lim_{\beta\uparrow 1} s_2 V_\beta = \frac{1}{4} + \frac{1}{\pi^2} + \frac{7\zeta(3)}{2\pi^2} = 0.77759\ldots.$$

Full statement and proof: lemma SF.10 of the supplement. Lean: proved: `Muon.Isotropic.lem_app_iso_V_i`, `Muon.Isotropic.lem_app_iso_V_ii`, *+2 more.*

V_β is the factor by which momentum multiplies the zero-ridge volatility at large batch: $\mathfrak{v}_{\text{iso},\beta} \approx NV_\beta$ (theorem K.3), where $\Omega(\beta^\ell)$ is, at large batch, the correlation of two updates ℓ steps apart. The lemma places V_β strictly between 1, the value without correlations, and $1/s_2 = B_{\text{eff}}/B$, the value if successive updates were correlated like successive buffers, by β^ℓ. The main text writes $c_\beta = s_2 V_\beta$, so that $\mathfrak{v}_\beta \approx c_\beta N B_{\text{eff}}/B$ in eq. (9); the lemma gives $c_\beta < 1$ and the limit $0.7776\ldots$ as $\beta \uparrow 1$.

Proof sketch. The coefficients of Ω form a probability distribution on the odd powers. With $x = 2n{+}1$, the square $k_n^2 = x^{-2}{+}2/(x(x{+}2)){+}(x{+}2)^{-2}$ sums over odd x to $\frac{\pi^2}{8}{+}1{+}(\frac{\pi^2}{8}{-}1) = \frac{\pi^2}{4}$, by $\sum_{x\ \mathrm{odd}} x^{-2} = \pi^2/8$ and a telescoping sum, so $\Omega(1) = 1$. Summing the geometric series in the lag, $1 + 2\sum_{\ell\geq 1}\beta^{\ell x} = (1+\beta^x)/(1-\beta^x)$, shows that V_β is the average of $(1+\beta^x)/(1-\beta^x)$ over odd x under these weights. Each value lies in $(1, (1+\beta)/(1-\beta)]$, with equality on the right only at $x = 1$, which gives the two bounds.

The limit is an average of $1/x$. As $\beta \uparrow 1$, $s_2(1{+}\beta^x)/(1{-}\beta^x) \to 1/x$, with the bound 2, so dominated convergence gives $\lim s_2 V_\beta = \frac{4}{\pi^2}\sum_n k_n^2/(2n+1)$. Partial fractions and $\sum_{x\ \mathrm{odd}} x^{-3} = \frac{7}{8}\zeta(3)$ evaluate the sum as $\frac{7}{8}\zeta(3) + \frac{\pi^2}{16} + \frac{1}{4}$, and multiplying by $4/\pi^2$ gives $\frac{1}{4} + \frac{1}{\pi^2} + \frac{7\zeta(3)}{2\pi^2}$. ◇

In the square case $N_{\mathrm{in}} = N_{\mathrm{out}} = N$ every row has the same lag coefficients $\overline{C}_{1,\ell}(\varrho)$: the deterministic covariance, in units of one row, of two updates ℓ steps apart (eq. (SG.2) at $\mu_1 = \cdots = \mu_N = 1$). Their limits $\overline{C}_{1,\ell}(0)$ as $\varrho \downarrow 0$ exist (proposition SF.3(ii)).

Theorem K.3 (Small-aspect integrated volatility)**.** *Let $\beta \in (0,1)$, $N_{\mathrm{in}} = N_{\mathrm{out}} = N$ and $\gamma = N/B \leq (1+\beta^2)/96$. Then for every $\ell \geq 1$*

$$\left|\overline{C}_{1,\ell}(0) - \Omega(\beta^\ell)\right| \leq \frac{96\gamma}{1+\beta^2}\,\Omega(\beta^\ell),$$

and the zero-ridge volatility of definition K.1 satisfies

$$\left|\frac{\mathfrak{v}_{\mathrm{iso},\beta}}{N} - V_\beta\right| \leq \frac{96\gamma}{1+\beta^2}\,(V_\beta - 1), \qquad \left|\frac{\mathfrak{v}_{\mathrm{iso},\beta}}{N\,V_\beta} - 1\right| \leq 96\gamma.$$

Full statement and proof: theorem SF.8 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_small_volatility_i`, `Muon.Isotropic.thm_app_iso_small_volatility_ii`; foundation results: `GaussianOpNorm`.

At large batch, in the square case, the zero-ridge volatility with momentum is NV_β up to a relative error 96γ, uniformly in β when $\gamma \leq 1/96$. This is the large-batch form $\mathfrak{v}_\beta \approx c_\beta N B_{\mathrm{eff}}/B$ of eq. (9), with $c_\beta = s_2 V_\beta$ (lemma K.1). The correlation of two updates ℓ steps apart is $\Omega(\beta^\ell)$, not the correlation β^ℓ of the buffers. Since $V_\beta < 1/s_2$, momentum amplifies the volatility by less than the effective batch $B_{\mathrm{eff}} = B/s_2$ by which it acts on the drift (theorem K.2); remark K.1 draws the consequence.

Proof sketch. The volatility is a sum over lags of explicit double integrals. By proposition SF.3(ii), $\mathfrak{v}_{\mathrm{iso},\beta}/N = 1 + 2\sum_{\ell\geq 1}\overline{C}_{1,\ell}(0)$, where the term 1 is the same-time value at zero ridge and $\overline{C}_{1,\ell}(0)$ is the correlation of two updates ℓ steps apart. In the square case this correlation is a double integral over two spectral parameters of a kernel built from the two roots $q = q(-x^2)$ and $q' = q(-y^2)$ (lemma SF.11):

$$\overline{C}_{1,\ell}(0) = \frac{4}{\pi^2}\int_0^\infty\!\!\int_0^\infty \frac{b_\ell\, q\, q'}{\gamma^3\,\Delta_\ell}\,\mathrm{d}x\,\mathrm{d}y, \qquad \Delta_\ell = \Big(1 - \frac{qq' e_\ell}{\gamma}\Big)^2 - \frac{qq' b_\ell^2}{\gamma^2}.$$

Here $b_\ell = \beta^\ell\,(r_\beta(q) - r_\beta(\beta^{2\ell}q'))/(q - \beta^{2\ell}q')$ is β^ℓ times a divided difference of the participation count, close to $\beta^\ell s_2$ when both roots are small, and $e_\ell = O(\beta^{2\ell}s_4)$.

At small aspect ratio the integrand is a fixed function of the linear root, up to a relative error of order γ. Rescale both spectral parameters by $\sqrt{s_2\gamma}$. By the sandwich of theorem K.2, the rescaled roots $s_2 q/\gamma$ lie within a factor $(1-\delta)^{-1}$ of $y(\xi^2)$, where y is the root of $xy = (1-y)^2$ and $\delta = 12\gamma(1-\beta^2)/(1+\beta^2)$; the normalized $b_\ell/(\beta^\ell s_2)$ lies in $[1-\delta, 1]$, and $qq' e_\ell/\gamma \leq \delta$. With $r = \beta^\ell$ and $\mathsf{Y}_0 = y(\xi^2)y(\xi'^2) \leq 1$, the integrand is increasing in each of these three quantities, so it is squeezed around $\mathfrak{f} \stackrel{\mathrm{def}}{=} r\mathsf{Y}_0/(1-r^2\mathsf{Y}_0)$:

$$\mathfrak{f}\Big(1 - \frac{3\delta}{1-\beta^2}\Big) \leq \text{integrand} \leq \mathfrak{f}\Big(1 + \frac{8\delta}{1-\beta^2}\Big), \qquad \frac{8\delta}{1-\beta^2} = \frac{96\gamma}{1+\beta^2}.$$

The perturbations are measured against the gap $1 - r^2\mathsf{Y}_0 \geq 1 - \beta^2$ in the denominator, which is why δ is divided by $1-\beta^2$. The hypothesis $\gamma \leq (1+\beta^2)/96$ is exactly $4\delta \leq (1-\beta^2)/2$, which keeps the perturbed denominator above half the gap.

The limit kernel is a geometric series of products, and each product integrates to k_n. Expanding $\mathfrak{f} = \sum_{n\ge 0} r^{2n+1}\mathsf{Y}_0^{n+1}$, the double integral factorizes into squares of one-dimensional integrals. With $\xi = 2\sinh\theta$ the linear root is $y(\xi^2) = e^{-2\theta}$, so

$$\int_0^\infty y(\xi^2)^{n+1}\,\mathrm{d}\xi = \int_0^\infty e^{-2(n+1)\theta}\,2\cosh\theta\,\mathrm{d}\theta = \frac{1}{2n+1} + \frac{1}{2n+3} = k_n,$$

and $\frac{4}{\pi^2}\iint \mathfrak{f} = \Omega(\beta^\ell)$. This is the lag bound. Summing it over ℓ against $V_\beta = 1 + 2\sum_{\ell\ge 1}\Omega(\beta^\ell)$ gives the first volatility bound, and dividing by $V_\beta \ge 1$ gives the second. ◇

The threshold aspect ratio is $\gamma_\beta^\circ \stackrel{\text{def}}{=} 2 + 1/(1-\beta^2)$, and $\mathsf{K}_\beta^\pm$ are explicit constants that depend only on β.

Theorem K.4 (Zero ridge, large aspect ratios)**.** *Let $\beta \in (0,1)$ and $\gamma_{\text{in}}, \gamma_{\text{out}} > 0$, put $\iota \stackrel{\text{def}}{=} \gamma_\wedge/\gamma_\vee \in (0,1]$ and*

$$\mathsf{F}(\iota) \stackrel{\text{def}}{=} \frac{1}{\iota}\int_0^\iota \sqrt{(1-x)(\iota-x)}\,\mathrm{d}x.$$

(i) The zero-ridge constant satisfies

$$\Big|C_\beta(\gamma_{\text{in}},\gamma_{\text{out}}) - \frac{\mathsf{F}(\iota)}{\sqrt{\pi}}\Big| \le \frac{2.32}{(1-\beta)\sqrt{\gamma_{\text{in}}\gamma_{\text{out}}}},$$

where $\mathsf{F}(1) = \frac{1}{2}$ and, for $\iota < 1$,

$$\mathsf{F}(\iota) = \frac{1}{\iota}\Big[\frac{(1+\iota)\sqrt{\iota}}{4} + \frac{(1-\iota)^2}{8}\log\frac{1-\sqrt{\iota}}{1+\sqrt{\iota}}\Big].$$

In particular, in the square case,

$$\Big|C_\beta(\gamma) - \frac{1}{2\sqrt{\pi}}\Big| \le \frac{2.32}{(1-\beta)\gamma} \qquad (\gamma > 0).$$

(ii) Let $N_{\text{in}} = N_{\text{out}} = N$ and $\gamma = N/B \ge \gamma_\beta^\circ$. The zero-ridge volatility of definition K.1 satisfies, with $\mathsf{K}_\beta^\pm$ of eq. (SF.57)*,*

$$\frac{1-2\mathsf{K}_\beta^-}{\gamma} \le \frac{\mathfrak{v}_{\text{iso},\beta}}{N\gamma} - \frac{1}{2} \le \frac{1+2\mathsf{K}_\beta^+}{\gamma}.$$

In particular $\mathfrak{v}_{\text{iso},\beta} = \frac{N^2}{2B}\big(1 + O_\beta(B/N)\big)$, and $\mathfrak{v}_{\text{iso},\beta}/(N\gamma) \to \frac{1}{2}$ as $\gamma \to \infty$ at every fixed $\beta \in (0,1)$.

Full statement and proof: theorem SF.11 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_zero_ridge_i`*,* `Muon.Isotropic.thm_app_iso_zero_ridge_ii`*, +1 more; foundation results:* `GaussianOpNorm`*.*

The theorem gives the isotropic kernels of MUON with momentum at large aspect ratio $\gamma = N/B$. The drift does not decay: in the square case $C_\beta(\gamma) \to 1/(2\sqrt{\pi})$, with no logarithm and no dependence on β, whereas without momentum $C_0(\gamma) \sim 1/(\sqrt{\pi}\gamma)$ (theorem SF.5). The volatility grows linearly, $\mathfrak{v}_{\text{iso},\beta} = \frac{N^2}{2B}(1 + O_\beta(B/N))$. The two kernels give the limit $1/(2\sqrt{\pi})$ of fig. 4 and, inserted in eq. (11), the learning-rate ratio $\sqrt{\pi}N^2\mathfrak{d}_\beta/(\sqrt{B}\,\mathfrak{v}_\beta) = \sqrt{B}\,(1 + O_\beta(B/N))$. Both kernels are deterministic limits as $\varrho \downarrow 0$ at fixed β; nothing is claimed for the algorithm at zero ridge.

Proof sketch. With momentum, the update approaches full rank as $\varrho \downarrow 0$. In the scalar equation of the introduction to appendix K, each term $\Phi(w_j^2 q)$ of $r_\beta(q)$ lies in $(0,1)$ and is close to one when $w_j^2 q \gg 1$. Hence $r_\beta(q)$ counts the past batches in the buffer, of age j and weight w_j, that the normalized update whitens, and the scalar equation keeps this count below $\gamma_\wedge$. With momentum the count grows without bound in q: for $(1-\beta)^2 q \ge 1$ (lemma SF.6(i)),

$$r_\beta(q) = \frac{\log\big((1-\beta)^2 q\big)}{2\log(1/\beta)} + O\Big(\frac{1}{1-\beta}\Big).$$

As $\varrho \downarrow 0$ the root grows, so the participation level ψ tends to its ceiling $\gamma_\wedge$. We write $\psi(0) = \gamma_\wedge$, and the squared Frobenius norm $B\psi$ of the update tends to the full rank $\min\{N_{\mathrm{in}}, N_{\mathrm{out}}\}$ (proposition SF.3(i)). Without momentum there is one age, $r_0 = \Phi < 1$, and $\psi(\varrho) \to \min\{1, \gamma_\wedge\}$; then $B\psi$ tends to $\min\{B, N_{\mathrm{in}}, N_{\mathrm{out}}\}$, the rank of one gradient (proposition K.1).

The drift is, up to an error of order $1/((1-\beta)\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}})$, an explicit function of the level. The drift eq. (SF.10) is an integral of $qL_\beta(q)$ along the root. Let Q_m be the root at level m. The scalar equation inverts exactly as (lemma SF.8)

$$s = \frac{\mathsf{W}(m)^2}{Q_m}, \qquad \mathsf{W}(m) \stackrel{\mathrm{def}}{=} \sqrt{(\gamma_{\mathrm{in}} - m)(\gamma_{\mathrm{out}} - m)},$$

so the integral can be taken over the level $m \in (0, \psi(\varrho))$ instead of the spectral parameter. After one integration by parts the integrand contains $\Lambda_\beta(Q_m)$, where $\Lambda_\beta(q) \stackrel{\mathrm{def}}{=} \int_0^q x^{-1/2} L_\beta(x)\,\mathrm{d}x$. Every whitened age contributes the same amount to Λ_β, and the error comes from the few ages near the threshold $w_j^2 q \approx 1$ (lemma SF.4(iv), lemma SF.7):

$$\int_0^\infty y^{-1/2} J(y)\,\mathrm{d}y = \sqrt{2\pi}, \qquad \Lambda_\beta(Q_m) = \sqrt{2\pi}\, m + O\Big(\frac{1}{1-\beta}\Big).$$

This gives the master estimate (theorem SF.3),

$$\Big|\mathfrak{d}_{\mathrm{iso}}(\varrho) - \frac{1}{\sqrt{\pi}\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}} \int_0^{\psi(\varrho)} \mathsf{W}(m)\,\mathrm{d}m\Big| < \frac{2.32}{(1-\beta)\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}}.$$

At $\varrho = 0$ the drift is C_β and the level is $\gamma_\wedge$, and the substitution $m = \gamma_\vee x$ evaluates the main term exactly as $\mathsf{F}(\iota)/\sqrt{\pi}$; in the square case $\mathsf{F}(1) = \frac{1}{2}$. This is part (i). The drift stays of order one because the integral runs over all $\gamma_\wedge$ directions.

With momentum the volatility is a sum over lags, and each lag coefficient is an exact average over the level. Successive updates are correlated, and the zero-ridge volatility is (proposition SF.3(ii))

$$\mathfrak{v}_{\mathrm{iso},\beta} = N\Big(1 + 2\sum_{\ell \geq 1} \overline{C}_{1,\ell}(0)\Big),$$

where $\overline{C}_{1,\ell}(0)$ is the correlation of two updates ℓ steps apart, a double integral of the closed-form kernel of the lag-ℓ covariance (lemma SF.11). Part (ii) says that the lag sum is $\gamma/4$ up to $O_\beta(1)$ (theorem SF.9). To see this, write the double integral in the two roots y and $z_\ell \stackrel{\mathrm{def}}{=} \beta^{-2\ell} z$; multiplying a root by $\beta^{-2\ell}$ shifts every age of the buffer by ℓ places. Let $\mathsf{A}(y) \stackrel{\mathrm{def}}{=} \gamma - r_\beta(y)$ be the residual count, the directions not yet whitened at the root y, and $\mathsf{j}(y) \stackrel{\mathrm{def}}{=} \mathsf{A}(y) + 2y\, r'_\beta(y)$. Averaging over the intermediate level η between y and z gives the exact identity (proposition SF.10, lemma SF.16, and eq. (SF.58))

$$\sum_{\ell \geq 1} \overline{C}_{1,\ell}(0) = \frac{1}{\gamma} \int_0^\gamma \mathbb{E}_\star\big[\Psi(y,z)\big]\,\mathrm{d}\eta,$$

$$\Psi(y,z) = \mathbf{1}\{y < Q_\gamma\}\, \frac{\mathsf{j}(y)\,\mathsf{T}(z)}{\gamma^2 \widehat{\Delta}(y,z)}, \qquad \mathsf{T}(z) = \sum_{\ell \geq 1} \mathbf{1}\{z_\ell < Q_\gamma\}\, \mathsf{j}(z_\ell).$$

Here $\widehat{\Delta}$ is the determinant that appears in that closed-form kernel, and $\mathbb{E}_\star$ averages over how far y and z sit from the root at level η, on a logarithmic scale, under a fixed probability law with density proportional to $1/\sinh$ of half the total offset.

At level η the summed integrand is approximately half the residual count, $(\gamma - \eta)/2$, whose average over $\eta \in (0, \gamma)$ is $\gamma/4$. On the diagonal $\gamma^2 \widehat{\Delta}(y,y) = \mathsf{A}(y)\,\mathsf{j}(y)$ (eq. (SF.54)), so the first factor of Ψ is $1/\mathsf{A}(y)$. A shift by ℓ ages raises the count by about ℓ, so the lag ℓ contributes about $\mathsf{A}(z) - \ell$ to $\mathsf{T}(z)$ while ℓ is below the residual count, and nothing beyond it (lemma SF.14). Precisely, with $\mathsf{A}_\vee$ the larger of the two residual counts, E_z the error of the counting formula at z, and $\bar{\mathsf{e}}_\beta$ a bound on the slope $y\, r'_\beta(y)$ (eqs. (SF.59) and (SF.60) and lemma SF.13),

$$\frac{(\mathsf{A}(z) - 1)_+^2}{2} \leq \mathsf{T}(z) \leq \frac{(\mathsf{A}_\vee + E_z)^2}{2} + 2\bar{\mathsf{e}}_\beta(\mathsf{A}_\vee + E_z).$$

The correction in (ii) is of relative order B/N, with a constant that is bounded but not identified, because these errors do not vanish. The integrand differs from $(\gamma-\eta)/2$ by terms linear or quadratic in the logarithmic offset and in E_z, and these terms have bounded averages. The lag sum is therefore $\gamma/4+O_\beta(1)$, and the relative error is $O_\beta(1/\gamma)=O_\beta(B/N)$ (proof of theorem SF.9, lemma SF.17). The hypothesis $\gamma\ge\gamma^\circ_\beta$ keeps the residual count at least $\Phi(1)$ wherever $\Psi>0$, and hence $\widehat{\Delta}$ away from zero there (lemma SF.14(v), lemma SF.15(v)). $\diamond$

The fixed-ridge participation count and the constants of the counting formula are

$$\widehat{\mathfrak{p}}_\beta(\varrho)\stackrel{\text{def}}{=}\frac{\log\big((1-\beta)^2\gamma_{\text{in}}\gamma_{\text{out}}/\varrho\big)}{2\log(1/\beta)},\qquad \mathsf{c}_\beta\stackrel{\text{def}}{=}\frac{\sqrt{\pi/2}}{1-\beta},\qquad \mathsf{c}'_\beta\stackrel{\text{def}}{=}1+\frac{1}{1-\beta^2},$$
$$\mathsf{c}''_\beta\stackrel{\text{def}}{=}\frac{1}{\log(1/\beta)}+\mathsf{c}_\beta+\mathsf{c}'_\beta,$$

and $\mathsf{k}_\beta\stackrel{\text{def}}{=}\mathsf{e}^-_\beta(e^2-1)e^{-3}/(\pi^2\mathsf{e}^+_\beta)$, with $\mathsf{e}^-_\beta\stackrel{\text{def}}{=}\beta^2\Phi'(1)$ and $\mathsf{e}^+_\beta\stackrel{\text{def}}{=}\mathsf{c}_\beta+1/(1-\beta^2)$.

Theorem K.5 (Fixed ridge, large aspect ratios)**.** *Let $\beta\in(0,1)$, $\varrho>0$ and $\gamma_{\text{in}},\gamma_{\text{out}}>0$, write $\widehat{\mathfrak{p}}=\widehat{\mathfrak{p}}_\beta(\varrho)$, and assume*

$$\widehat{\mathfrak{p}}\ge\frac{2\mathsf{c}'_\beta}{\log(1/\beta)\,\gamma_\wedge},\qquad \widehat{\mathfrak{p}}+\mathsf{c}'_\beta\le\frac{\gamma_\wedge}{2}.$$

Put $e_\beta\stackrel{\text{def}}{=}\max\big(\mathsf{c}'_\beta,\ \mathsf{c}_\beta+2(\widehat{\mathfrak{p}}+\mathsf{c}'_\beta)/(\log(1/\beta)\gamma_\wedge)\big)$. Then

$$\big|\psi(\varrho)-\widehat{\mathfrak{p}}\big|\le e_\beta,\qquad \Big|\sqrt{\gamma_{\text{in}}\gamma_{\text{out}}}\,\mathfrak{d}_{\text{iso}}(\varrho)-\frac{\widehat{\mathfrak{p}}}{\sqrt{\pi}}\Big|\le\frac{1}{\sqrt{\pi}}\Big(e_\beta+\frac{(\widehat{\mathfrak{p}}+\mathsf{c}'_\beta)^2}{\gamma_\wedge}\Big)+\frac{2.32}{1-\beta}.$$

If moreover $N_{\text{in}}=N_{\text{out}}=N$ and $\gamma=N/B$, then

$$\mathfrak{v}_{\text{iso}}(\varrho)\ge B\,\psi(\varrho)+\mathsf{k}_\beta B\Big(1-\frac{\psi(\varrho)}{\gamma}\Big)^2\big(\psi(\varrho)-\mathsf{c}''_\beta-\mathsf{c}'_\beta-1\big)_+^2.$$

Full statement and proof: theorem SF.13 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_fixed_ridge_i`, `Muon.Isotropic.thm_app_iso_fixed_ridge_ii`.

The theorem describes what a fixed positive ridge does at small batch with momentum. The main text uses the regularized polar map as a model of K Newton–Schulz steps, with an effective ridge of order a^{-2K}, and cites this theorem for the effect of a small ridge. A past batch of age j enters the buffer with weight $w_j=(1-\beta)\beta^j$, and the ridge whitens it only if its weight is large enough; the number of ages that pass is about $\widehat{\mathfrak{p}}$. The theorem says that the participation level $\psi(\varrho)$ is this count up to a bounded error, that the drift is about $\widehat{\mathfrak{p}}/\sqrt{\pi\gamma_{\text{in}}\gamma_{\text{out}}}$, which decays like $\log\gamma/\gamma$ in the square case instead of reaching the zero-ridge value $1/(2\sqrt{\pi})$, and that the volatility is at least of order $B\psi(\varrho)^2$. The two hypotheses ask that the count be past the first age and well below its ceiling $\gamma_\wedge$.

Proof sketch. The participation level counts the ages whose weight passes the threshold set by the root. The terms $\Phi(w_j^2q)$ of $r_\beta(q)$ are within $\sqrt{\pi/2}\,(w_j^2q)^{-1/2}$ of one for the ages with $w_j^2q\ge1$, and at most w_j^2q for the others; both corrections decay geometrically with the distance of the age to the threshold. Hence, for $(1-\beta)^2q\ge1$ (lemma SF.6(i)),

$$\mathsf{n}_\beta(q)-\mathsf{c}_\beta<r_\beta(q)\le\mathsf{n}_\beta(q)+\mathsf{c}'_\beta,\qquad \mathsf{n}_\beta(q)=\frac{\log\big((1-\beta)^2q\big)}{2\log(1/\beta)},$$

where $\mathsf{n}_\beta(q)$ is, up to one, the number of ages with $w_j^2q\ge1$.

At a fixed ridge the root is about $\gamma_{\text{in}}\gamma_{\text{out}}/\varrho$, and the count at that root is $\widehat{\mathfrak{p}}$. Write $H=\psi(\varrho)$ and $\mathsf{W}(m)=\sqrt{(\gamma_{\text{in}}-m)(\gamma_{\text{out}}-m)}$. The scalar equation at $s=\varrho$ says that the root is $\mathsf{W}(H)^2/\varrho$, so

$$\mathsf{n}_\beta\big(q(-\varrho)\big)=\widehat{\mathfrak{p}}+\frac{\log\big(\mathsf{W}(H)/\sqrt{\gamma_{\text{in}}\gamma_{\text{out}}}\big)}{\log(1/\beta)}\le\widehat{\mathfrak{p}}.$$

The upper count then gives $H\le\widehat{\mathfrak{p}}+\mathsf{c}'_\beta\le\gamma_\wedge/2$. Below half the ceiling, $\mathsf{W}(H)\ge\sqrt{\gamma_{\text{in}}\gamma_{\text{out}}}\,(1-H/\gamma_\wedge)$, so the logarithm costs at most $2H/(\gamma_\wedge\log(1/\beta))$, and the lower count gives $H\ge\widehat{\mathfrak{p}}-e_\beta$.

The case $(1-\beta)^2 q(-\varrho) < 1$, in which not even the newest batch passes the threshold, would force a ridge that the first hypothesis excludes.

The drift is a fixed function of the level, so each whitened age contributes the same amount. By the master estimate (theorem SF.3, explained in the note on theorem K.4),

$$\Big|\mathfrak{d}_{\mathrm{iso}}(\varrho) - \frac{1}{\sqrt{\pi}\,\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}\int_0^H \mathsf{W}(m)\,\mathrm{d}m\Big| < \frac{2.32}{(1-\beta)\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}}.$$

For $m \le H$ the integrand lies between $\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}(1 - m/\gamma_\wedge)$ and $\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}$, so the main term times $\sqrt{\gamma_{\mathrm{in}}\gamma_{\mathrm{out}}}$ is $H/\sqrt{\pi}$ up to $H^2/(\sqrt{\pi}\gamma_\wedge)$. With the bound on H this gives the drift bound.

The volatility exceeds its same-time value by the correlations of updates that whiten common ages. By proposition SF.3(ii), $\mathfrak{v}_{\mathrm{iso}}(\varrho) = B\psi(\varrho) + 2N\sum_{\ell\ge1}\overline{C}_{1,\ell}(\varrho)$. The lag lower bound (lemma SF.12) bounds the determinant of the lag kernel by one, writes the double integral in the two roots, and keeps only pairs of roots past the threshold and within a factor e^2 of each other, where the count is accurate. A shift of ℓ ages raises the count by about ℓ, so the lag ℓ is seen only at levels that stay below $\psi(\varrho)$ after the shift, a range of length about $\psi(\varrho) - \ell$:

$$\overline{C}_{1,\ell}(\varrho) \ge \frac{\mathsf{k}_\beta}{\gamma^3}\big(\gamma - \psi(\varrho)\big)^2\big(\psi(\varrho) - \ell - \mathsf{c}''_\beta - \mathsf{c}'_\beta\big)_+.$$

Summing over ℓ, with $\sum_{\ell\ge1}(X-\ell)_+ \ge (X-1)_+^2/2$, gives the volatility bound. At leading order the integrated volatility is $B\psi(\varrho)^2$ (corollary SF.2(iii)). ◇

For $0 < \varrho < 1$ the participation count of the ridge is $\mathfrak{p}_\beta(\varrho) \overset{\mathrm{def}}{=} \log(1/\varrho)/(2\log(1/\beta))$, up to one the number of ages $j \ge 0$ with $\beta^{2j} \ge \varrho$, and the crossover function is

$$\mathcal{P}(x) \overset{\mathrm{def}}{=} \tfrac12 - \tfrac12(1-x)^2 + (1-x)^2\log(1-x) \quad (0 \le x < 1), \qquad \mathcal{P}(1) \overset{\mathrm{def}}{=} \tfrac12.$$

Theorem K.6 (Crossover)**.** *Let $\beta \in (0,1)$ be fixed, let $N_{\mathrm{in}} = N_{\mathrm{out}} = N$ and $\gamma = N/B \to \infty$, and let $\varrho_\gamma \in (0,1)$ satisfy $\mathfrak{p}_\beta(\varrho_\gamma)/\gamma \to c \in (0,\infty)$. Then $\psi(\varrho_\gamma)/\gamma \to c \wedge 1$, the same-time total satisfies $B\psi(\varrho_\gamma)/N \to c \wedge 1$,*

$$\mathfrak{d}_{\mathrm{iso}}(\varrho_\gamma) \longrightarrow \Xi(c) \overset{\mathrm{def}}{=} \frac{(c\wedge1)\big(1 - \frac12(c\wedge1)\big)}{\sqrt{\pi}},$$

and, with $\mathcal{P}$ of eq. (SF.63)*,*

$$\frac{\mathfrak{v}_{\mathrm{iso}}(\varrho_\gamma)}{N\gamma} \longrightarrow \mathcal{P}(c\wedge1).$$

Full statement and proof: theorem SF.14 of the supplement. Lean: proved: `Muon.Isotropic.thm_app_iso_crossover`.

The theorem connects the fixed-ridge regime with the zero-ridge regime at small batch, in the square case. The ridge lets about $\mathfrak{p}_\beta(\varrho)$ past batches be whitened, while there are only $\gamma = N/B$ directions to fill. When the ridge shrinks with γ so that $\mathfrak{p}_\beta(\varrho_\gamma) \approx c\gamma$, the update whitens the fraction $c \wedge 1$ of the directions, and the drift and the volatility have explicit limits in this fraction. As $c \to 0$ they reduce to the fixed-ridge laws, $\Xi(c) \sim c/\sqrt{\pi}$ and $\mathcal{P}(c) \sim c^2$, and for $c \ge 1$, that is $\varrho \lesssim \beta^{2\gamma}$, they equal the zero-ridge limits $1/(2\sqrt{\pi})$ and $\frac12$ of theorem K.4. The main text uses these laws at the ridge calibrated to Newton–Schulz, where they predict the NS-5 speedups of fig. 6.

Proof sketch. The whitened fraction $\psi(\varrho_\gamma)/\gamma$ tends to $c\wedge1$, by the counting argument of theorem K.5. Write $H = \psi(\varrho_\gamma)$ and $\mathfrak{p} = \mathfrak{p}_\beta(\varrho_\gamma)$. The root at the ridge is $(\gamma - H)^2/\varrho_\gamma$, and the count of ages that pass its threshold, n_β of that theorem's note, brackets H (lemma SF.6(i)):

$$\mathfrak{p} + \frac{\log\big((1-\beta)(\gamma-H)\big)}{\log(1/\beta)} - \mathsf{c}_\beta < H \le \mathfrak{p} + \frac{\log\big((1-\beta)\gamma\big)}{\log(1/\beta)} + \mathsf{c}'_\beta.$$

The corrections are logarithmic in γ, hence negligible against $\mathfrak{p} \approx c\gamma$, unless $\gamma - H$ is small. If $c < 1$, the upper bound keeps H below $(1-\delta)\gamma$, and then the lower bound gives $H \ge c\gamma - o(\gamma)$. If $c \ge 1$ and $H \le (1-\delta)\gamma$ along a subsequence, the lower bound gives $H > (1-\delta/2)\gamma$, a contradiction;

so $H/\gamma \to 1$. When the ridge would admit more ages than there are directions, all directions are whitened, as at zero ridge.

In the square case the drift is a quadratic function of the whitened fraction, up to $O(1/\gamma)$. The master estimate (theorem SF.3(ii)) with $\mathsf{W}(m) = \gamma - m$ reads

$$\Big|\mathfrak{d}_{\mathrm{iso}}(\varrho) - \Xi_0\Big(\frac{\psi(\varrho)}{\gamma}\Big)\Big| < \frac{2.32}{(1-\beta)\gamma}, \qquad \Xi_0(x) = \frac{x - \frac{1}{2}x^2}{\sqrt{\pi}},$$

because $\gamma^{-2}\int_0^{\psi}(\gamma - m)\,\mathrm{d}m = \psi/\gamma - \frac{1}{2}(\psi/\gamma)^2$. Continuity of Ξ_0 gives the drift limit $\Xi(c) = \Xi_0(c \wedge 1)$. The order-one quantity is the drift itself, so $\gamma\,\mathfrak{d}_{\mathrm{iso}}(\varrho_\gamma) \to \infty$.

The lag sum at a positive ridge is the zero-ridge computation with the levels cut at ψ instead of γ (theorem SF.12). The lag coefficients have the same exact representation in the participation levels as at zero ridge (proposition SF.10; see the note on theorem K.4), with the levels restricted to $(0, \psi)$ and an extra factor $\mathsf{u}_\varrho(y) = (1 - \varrho y/\mathsf{A}(y)^2)^{-1/2} \geq 1$ at each root, where $\mathsf{A}(y) = \gamma - r_\beta(y)$ is the residual count. At level η there are $X = \psi - \eta$ shifts left before the cut and $x = \gamma - \eta$ residual directions. The lag $\ell < X$ contributes about $x - \ell$. Dividing the sum of these contributions by the residual count leaves $\mathsf{H}(\eta) = X - X^2/(2x)$. Its average is exactly

$$\frac{1}{\gamma}\int_0^{\psi}\Big(\psi - \eta - \frac{(\psi-\eta)^2}{2(\gamma-\eta)}\Big)\mathrm{d}\eta = \frac{\gamma}{2}\,\mathcal{P}\Big(\frac{\psi}{\gamma}\Big),$$

and at $\psi = \gamma$ it is the value $\gamma/4$ of the zero-ridge case (theorem SF.9). The errors of the counting formula cost $O_\beta((1+\psi)/\gamma)$ in the lag sum. So does the ridge factor, which is close to one except for roots within a bounded logarithmic distance of the top root $q(-\varrho)$; the shifted roots $\beta^{-2\ell}z$, equally spaced in logarithmic scale with step $2\log(1/\beta)$, meet that window about once per step, about ψ times in all (lemma SF.18). Hence $\mathfrak{v}_{\mathrm{iso}}(\varrho)/N = \psi/\gamma + \gamma\,\mathcal{P}(\psi/\gamma) + O_\beta((1+\psi)/\gamma)$, and dividing by γ and using the continuity of $\mathcal{P}$ gives the volatility limit $\mathcal{P}(c \wedge 1)$. ◇

Remark K.1 (Momentum as a change of batch and of step)**.** *In the oversampled square case $N < B$, with $\gamma = N/B$ small, the small-aspect theorems describe momentum at the level of the scalar equation eq. (28), read with its formal zero-ridge coefficients. For every step h, with $h' \overset{\mathrm{def}}{=} h/\sqrt{s_2} = h\sqrt{(1+\beta)/(1-\beta)}$,*

$$-2h\,C_\beta\sqrt{\mathcal{R}} + \frac{h^2}{2B}\,\mathfrak{v}_{\mathrm{iso},\beta} = -2h'\big(\sqrt{s_2}\,C_\beta\big)\sqrt{\mathcal{R}} + \frac{h'^2}{2B}\big(s_2\,\mathfrak{v}_{\mathrm{iso},\beta}\big).$$

On the drift, $\sqrt{s_2}\,C_\beta(\gamma) = C_0(\gamma)(1 + O(\gamma))$ by theorem K.2(iii): momentum acts as the effective batch $B/s_2 = B(1+\beta)/(1-\beta)$. On the volatility the factor is V_β, not $1/s_2$: $\mathfrak{v}_{\mathrm{iso},0} = N$ and $\mathfrak{v}_{\mathrm{iso},\beta} = NV_\beta(1+O(\gamma))$ (theorem K.3), so that $s_2\,\mathfrak{v}_{\mathrm{iso},\beta} = s_2V_\beta\,\mathfrak{v}_{\mathrm{iso},0}(1+O(\gamma))$ with $s_2V_\beta \in (s_2, 1)$ (lemma K.1). Momentum at step h is therefore no momentum at step h' with the volatility multiplied by $s_2V_\beta < 1$. The step-rescaling reading is only approximate in the oversampled isotropic case: momentum carries less noise than the rescaled step, by the factor s_2V_β, which is $0.826\ldots$ at $\beta = \frac{1}{2}$ and tends to $0.7776\ldots$ as $\beta \uparrow 1$. The main text writes $B_{\mathrm{eff}} \overset{\mathrm{def}}{=} B/s_2$ for the effective batch and $c_\beta \overset{\mathrm{def}}{=} s_2V_\beta$ for this factor, so that $\mathfrak{v}_{\mathrm{iso},\beta} = c_\beta N B_{\mathrm{eff}}/B\,\big(1+O(\gamma)\big)$. The factor $1/s_2$ for the volatility of corollary L.1(ii) belongs to a different regime, the resolved rows of a power-law spectrum with $\gamma \to \infty$; that regime and the oversampled isotropic one are disjoint, and neither statement specializes to the other.

Full statement and proof: remark SF.1 of the supplement. Lean: not formalized (remark).

At large batch, momentum at step h is no momentum at the larger step $h' = h\sqrt{B_{\mathrm{eff}}/B}$, with the drift C_0 of no momentum and the volatility of no momentum multiplied by $c_\beta = s_2V_\beta < 1$, up to factors $1 + O(\gamma)$. The displayed identity is algebra; its content is the two large-batch laws $\sqrt{s_2}\,C_\beta = C_0(1 + O(\gamma))$ (theorem K.2(iii)) and $\mathfrak{v}_{\mathrm{iso},\beta} = NV_\beta(1 + O(\gamma))$ with $\mathfrak{v}_{\mathrm{iso},0} = N$ for $N < B$ (theorem K.3 and proposition K.1). This is the reading behind B_{eff} and c_β in fig. 4 and in the large-batch discussion of the speedup in section 3. It concerns the scalar equation with its formal zero-ridge coefficients, and the factor $1/s_2$ of the power-law regime (corollary L.1(ii)) belongs to a disjoint regime.

L Power-law spectra and the large-batch regime

This chapter computes the coefficients $\mathfrak{d}_i$ and $\mathfrak{v}_i$ of the row recursion eq. (14) in the half-anisotropic setting $\Sigma_{\text{in}} = \mathrm{Id}_N$, $N_{\text{out}} = N_{\text{in}} = N$, at zero normalized ridge (except corollary L.2, which concerns the expected update at a positive ridge). At each finite dimension these are deterministic functions of the spectrum $\mu_1 \geq \cdots \geq \mu_N$, of B and of β, obtained as limits $\varrho \downarrow 0$ of the coefficients of the drift and volatility chapters; no result concerns the algorithm at zero ridge. Two regimes are treated with one set of tools (appendix G.5). In the thin-batch regime the first eight results locate the crossover between resolved and unresolved rows and give both coefficients on either side of it, which is table 1. In the oversampled regime the last six results concern the drift alone, for a general spectrum (the last one specializes to power laws), and give the square-root law of theorem 4.1 and the modes that escape it.

The thin-batch results share the following setup. Assumption S4 fixes $\alpha > 0$ and $\beta \in (0, 1)$ and lets $N, B \to \infty$ with $\gamma = N/B \to \infty$, with $\mu_i = c_N i^{-\alpha}$ for amplitudes $c_N > 0$. The weights are $w_j = (1-\beta)\beta^j$, $s_p = \sum_j w_j^p$, so $s_2 = (1-\beta)/(1+\beta)$, and $c_0 = \mathbb{E}\,|g| = \sqrt{2/\pi}$ for a standard Gaussian g. With $J(x) = \mathbb{E}[g^2/(1+xg^2)]$ put

$$F_\beta(q) = \sum_j w_j^2 J(w_j^2 q), \qquad L_\beta(q) = \sum_j w_j J(w_j^2 q),$$

$$r_\beta(q) = qF_\beta(q), \quad Q_m = r_\beta^{-1}(m), \quad E(m) = Q_m r_\beta'(Q_m),$$

and the participation $m_N(\Lambda) = B^{-1}\sum_i \Lambda\mu_i/(1+\Lambda\mu_i)$ with inverse $\Lambda_N : (0,\gamma) \to (0,\infty)$. The constants are

$$a_N = \sqrt{c_N} N^{-\alpha/2}, \quad K_\alpha = \frac{\pi}{\alpha \sin(\pi/\alpha)}, \quad i_{\text{c}} = \frac{B}{K_\alpha},$$
$$b_N = \sqrt{c_N}\big(\gamma i_{\text{c}}^\alpha\big)^{-1/2}, \quad d_\alpha = \frac{c_0 \pi}{\alpha \cos(\pi/(2\alpha))},$$

the last four used only for $\alpha > 1$. The zero-ridge drift $\mathfrak{d}_i(0)$, row lag coefficients $\overline{C}_{i,\ell}(0)$ and integrated volatility $\mathfrak{v}_i(0) = \sum_{\ell\in\mathbb{Z}} \overline{C}_{i,\ell}(0)$ (with $\overline{C}_{i,-\ell} = \overline{C}_{i,\ell}$) are the limits as $\varrho \downarrow 0$ of the drift coefficient $\mathfrak{d}_i(\varrho) = \mathfrak{d}(\mu_i, 1)$ and of the row coefficients of the volatility chapter, at each fixed dimension (definition SG.2).

The oversampled results share the following objects. The effective batch is $B_{\text{eff}} = B/s_2$, the small parameter is $\epsilon_{\text{os}} = 3(1-\beta^2)N_{\text{in}}/B$, and $\hat{\nu} = \hat{\nu}_{\text{out}} = {N_{\text{out}}}^{-1}\sum_k \delta_{\mu_k}$ is the empirical spectral distribution of Σ_{out}, with shape ratio $r = N_{\text{out}}/N_{\text{in}}$. For a probability measure ν on $[0,\infty)$ let $m_\nu(c) = \int x\,\nu(\mathrm{d}x)/(c+x)$, and let $b_0(\hat{s})$ be the unique root in $(0,1]$ of $b + r\,m_\nu(\hat{s}/b) = 1$. Then

$$\mathfrak{d}^{\text{os}}_{\nu,r}(\mu;\hat{\varrho}) = \frac{\sqrt{2}}{\pi}\int_0^\infty \frac{b_0(\hat{s})\,\mu}{\hat{s} + b_0(\hat{s})\,\mu}\,\mathrm{d}v \quad (\hat{s} = \hat{\varrho} + v^2),$$

$$\chi_\nu(\mu) = \int_0^\infty \varpi(y)\sqrt{1 - m_\nu(\mu y)}\,\mathrm{d}y, \qquad \varpi(y) = \frac{2}{\pi}\frac{\sqrt{y}}{(1+y)^2},$$

and $\mathfrak{d}^{\text{os}}_\nu(\mu) = \mathfrak{d}^{\text{os}}_{\nu,1}(\mu;0)$. The zero-ridge coefficient $\mathfrak{d}_i(0)$ is the limit of $\mathfrak{d}_i(\varrho)$ as $\varrho \downarrow 0$, which exists for every spectrum (theorem SH.1(ii)).

For participations $m_1, m_2 > 0$, with $q_k = Q_{m_k}$, an integer lag $\ell \geq 0$ and a standard Gaussian g, the functions referred to in the definition as those of eq. (SG.20) are

$$\mathsf{D}_\ell(m_1,m_2) = \sum_{a\geq 0} \mathbb{E}\,\frac{\beta^\ell\sqrt{q_1 q_2}\,w_a^2 g^2}{(1+q_1 w_a^2 g^2)(1+\beta^{2\ell} q_2 w_a^2 g^2)},$$

$$\mathsf{M}_\ell(m_1,m_2) = \sum_{a\geq 0} \mathbb{E}\,\frac{\beta^{2\ell} q_1 q_2\,w_a^4 g^4}{(1+q_1 w_a^2 g^2)(1+\beta^{2\ell} q_2 w_a^2 g^2)},$$

and for a negative lag $-\ell$ they are those of the lag ℓ at (m_2, m_1).

Definition L.1 (Profiles)**.** *Let* $S_\alpha(V) \stackrel{\text{def}}{=} \int_0^1 V(V+t^\alpha)^{-1}\mathrm{d}t$ *for* $V>0$*, a strictly increasing bijection onto* $(0,1)$*, and* $V_\alpha \stackrel{\text{def}}{=} S_\alpha^{-1}$*. For* $0<x\le 1$ *and* $0<p<1$ *put*

$$f_x(p) \stackrel{\text{def}}{=} \frac{V_\alpha(p)}{V_\alpha(p)+x^\alpha}, \qquad U_\alpha(p) \stackrel{\text{def}}{=} \frac{V_\alpha(p)}{p(1-p)}.$$

The bulk profiles are

$$\begin{aligned} \Psi_\alpha(x) &\stackrel{\text{def}}{=} c_0 \int_0^1 \sqrt{\frac{1-p}{p\,V_\alpha(p)}}\; f_x(p)\,\mathrm{d}p, \\ C_{\alpha,x}(\tau) &\stackrel{\text{def}}{=} \int_0^{1-\tau} \frac{f_x(p+\tau)-f_x(p)}{\tau}\sqrt{\frac{U_\alpha(p)}{U_\alpha(p+\tau)}}\,\mathrm{d}p \quad (0<\tau<1), \\ \Gamma_\alpha(x) &\stackrel{\text{def}}{=} 2\int_0^1 C_{\alpha,x}(\tau)\,\mathrm{d}\tau. \end{aligned} \tag{33}$$

For $\alpha>1$*, the crossover profiles are, for* $y, X>0$*,*

$$\begin{aligned} \Phi_{\alpha,\beta}(y) \stackrel{\text{def}}{=} \frac{1}{\pi}\int_0^\infty &\sqrt{q}\,L_\beta(q)\,r_\beta(q)^{-\frac{\alpha+1}{2}} \\ &\times \frac{y\,r_\beta(q)^\alpha}{1+y\,r_\beta(q)^\alpha}\Big(1+(\alpha-1)\frac{q\,r_\beta'(q)}{r_\beta(q)}\Big)\frac{\mathrm{d}q}{q}, \end{aligned} \tag{34}$$

$$\begin{aligned} G_{\alpha,\beta}(X) \stackrel{\text{def}}{=} \frac{1}{\pi^2}\sum_{\ell\in\mathbb{Z}}\int_0^\infty\int_0^\infty &\frac{\sqrt{\zeta_X(u)\big(1-\zeta_X(u)\big)\zeta_X(v)\big(1-\zeta_X(v)\big)}}{\sqrt{uv}} \\ &\times \frac{\mathsf{D}_\ell(u,v)\,\vartheta_\infty(u)\vartheta_\infty(v)}{\Delta_\ell^\infty(u,v)}\,\mathrm{d}u\,\mathrm{d}v, \end{aligned} \tag{35}$$

where D_ℓ *and* M_ℓ *are the functions of eq.* (SG.20)*, which depend on* $(m_1,m_2)=(u,v)$ *only through* (Q_u,Q_v)*, and*

$$\begin{aligned} \zeta_X(u) &\stackrel{\text{def}}{=} \frac{u^\alpha}{X^\alpha+u^\alpha}, \qquad \vartheta_\infty(u) \stackrel{\text{def}}{=} \frac{1}{E(u)}+\frac{\alpha-1}{u}, \\ t_\alpha(u,v) &\stackrel{\text{def}}{=} \frac{v^\alpha u-u^\alpha v}{v^\alpha-u^\alpha}, \qquad \Delta_\ell^\infty \stackrel{\text{def}}{=} 1-\frac{t_\alpha(u,v)\,\mathsf{M}_\ell(u,v)}{uv}, \end{aligned}$$

with $t_\alpha(u,u) \stackrel{\text{def}}{=} (\alpha-1)u/\alpha$*, and with the convention of eq.* (SG.20) *for negative lags.*

Full statement and proof: definition SG.3 of the supplement. Lean: proved: `Muon.PowerLaw.profU`*,* `Muon.PowerLaw.profF`*, +10 more.*

These are the profiles of table 1. The function S_α is the continuum limit of the participation of a power law: a participation fraction $p=m/\gamma$ corresponds to the level $V_\alpha(p)$, and $f_x(p)$ is the share of the row at position $x=i/N$ that is whitened at that level (appendix G.5). The drift profile Ψ_α integrates this share against the weight $c_0\sqrt{(1-p)/(pV_\alpha)}$ that the averaged momentum weights produce, and $C_{\alpha,x}(\tau)$ is the limiting correlation of two updates of row x that are $\tau\gamma$ steps apart, so that Γ_α is the integrated volatility per unit of γ. Neither depends on β. For $\alpha>1$ the crossover profiles $\Phi_{\alpha,\beta}$ and $G_{\alpha,\beta}$ describe the rows near the crossover index $i_\mathrm{c}=B/K_\alpha$, and they keep the full momentum weight functions.

Proposition L.1 (The profiles)**.**

(i) Ψ_α and Γ_α are finite, positive and continuous on $(0,1]$, Ψ_α is bounded below by a positive constant there, and as $x \downarrow 0$

$$\Psi_\alpha(x) \sim \begin{cases} \dfrac{c_0}{\sqrt{1-\alpha}} \log(x^{-\alpha}), & \alpha < 1, \\ \dfrac{2}{3} c_0 \big[\log(1/x)\big]^{3/2}, & \alpha = 1, \\ d_\alpha \sqrt{K_\alpha}\, x^{-(\alpha-1)/2}, & \alpha > 1, \end{cases}$$
$$\Gamma_\alpha(x) \sim \begin{cases} \dfrac{x^\alpha \big[\log(x^{-\alpha})\big]^2}{1-\alpha}, & \alpha < 1, \\ \dfrac{4}{9} x \big[\log(1/x)\big]^3, & \alpha = 1, \\ \dfrac{4K_\alpha^2}{\alpha - 1}\, x, & \alpha > 1. \end{cases} \tag{36}$$

(ii) For $\alpha > 1$, $\Phi_{\alpha,\beta}$ and $G_{\alpha,\beta}$ are finite, positive and continuous on $(0,\infty)$, and

$$\Phi_{\alpha,\beta}(y) \sim \sqrt{y/s_2}\ (y \to \infty), \quad \Phi_{\alpha,\beta}(y) \sim d_\alpha\, y^{\frac{\alpha-1}{2\alpha}}\ (y \downarrow 0),$$
$$G_{\alpha,\beta}(X) \to \frac{1}{s_2}\ (X \downarrow 0), \quad G_{\alpha,\beta}(X) \sim \frac{4K_\alpha}{\alpha - 1} X\ (X \to \infty). \tag{37}$$

Full statement and proof: proposition SG.5 of the supplement. Lean: proved: `Muon.PowerLaw.prop_app_pl_profiles_i`, `Muon.PowerLaw.prop_app_pl_profiles_ii`.

Part (i) is the small-x table to the right of table 1, with its constants: as $x \downarrow 0$ the unresolved drift profile grows only logarithmically in $1/x$ for $\alpha \le 1$ and like $x^{-(\alpha-1)/2}$ for $\alpha > 1$, while the unresolved volatility per unit of γ vanishes like x^α times logarithms for $\alpha \le 1$ and linearly in x for $\alpha > 1$. Part (ii) gives the two ends of the crossover profiles for $\alpha > 1$; they are the resolved and unresolved laws of theorems L.1 and L.2, which the profiles join. All statements concern explicit functions and involve no dimension.

Proof sketch. The drift profile is dominated, as $x \downarrow 0$, by small participation fractions, where the inverse V_α has the power laws of appendix G.5. For $\alpha < 1$, $V_\alpha(p) \sim (1-\alpha)p$ turns the integrand, without the factor c_0, into $(1-\alpha)^{-1/2}(p + x^\alpha/(1-\alpha))^{-1}$ on a fixed small interval, whose integral is $\log(x^{-\alpha})(1 + o(1))$; at $\alpha = 1$ the extra logarithm of $V_1(p) \sim p/\log(1/p)$ gives the power $3/2$. For $\alpha > 1$ the substitution $p = K_\alpha x u$ isolates a convergent integral (lemma SG.9),

$$\Psi_\alpha(x) \sim c_0 \sqrt{K_\alpha}\, x^{-(\alpha-1)/2} \int_0^\infty \frac{u^{(\alpha-1)/2}}{1+u^\alpha}\, \mathrm{d}u = d_\alpha \sqrt{K_\alpha}\, x^{-(\alpha-1)/2}.$$

The volatility profile is a positive double integral, and for $\alpha \le 1$ it is asymptotically the square of a single integral. With $A_x(p) = \sqrt{f_x(p)(1-f_x(p))(1-p)/p}$, a partial-fraction identity in the variable of the spectrum gives exactly (lemma SG.21)

$$\Gamma_\alpha(x) = 2 \int_{0<p<q<1} \frac{A_x(p) A_x(q)}{\Delta_{\mathrm{c}}(p,q)}\, \mathrm{d}p\, \mathrm{d}q, \qquad \Big(\int_0^1 A_x\Big)^2 = \frac{x^\alpha \Psi_\alpha(x)^2}{c_0^2},$$

where $\Delta_{\mathrm{c}} \le 1$ is the continuum determinant of the lag kernel. Without the determinant the double integral is the square on the right. For $\alpha \le 1$ the integral of A_x is spread over a range of scales of p whose logarithmic length tends to infinity, so most of the double integral comes from pairs with $q > Rp$ and $p, q < \eta$, where $1 - \Delta_{\mathrm{c}} \le \eta + 1/R$; the diagonal strip $q \le Rp$ and the upper end $p \to 1$ are of lower order. This gives $\Gamma_\alpha(x) \sim x^\alpha \Psi_\alpha(x)^2/c_0^2$ and the volatility equivalents of (i). For $\alpha > 1$ the mass of A_x sits at $p \asymp x$, the determinant matters, and the substitution $p = K_\alpha x r$ turns $\Gamma_\alpha(x)/(K_\alpha x)$ into a double integral over the lag and the participation, with value $4K_\alpha/(\alpha-1)$ (lemma SG.23).

The ends of the crossover profiles in (ii) come from the two mechanisms of appendix G.5. As $y \to \infty$ only small participations count in $\Phi_{\alpha,\beta}(y)$ and the head Lorentzian gives $\sqrt{y/s_2}$; as $y \downarrow 0$ the substitution $m = y^{-1/\alpha} w$ moves it to large participation, where averaging gives the integral above

(lemma SG.13). For $G_{\alpha,\beta}(X)$ the rescaling $m = Xr$ sends every lag term to $\beta^{|\ell|}$ as $X \downarrow 0$, with sum $1/s_2$, and as $X \to \infty$ the concentration of the lag kernel on $m_2 = m_1 + \ell$ (lemma SG.22) gives the double integral $4K_\alpha/(\alpha-1)$ (lemma SG.23). Finiteness, positivity and continuity follow from majorants that are locally uniform in the argument. $\diamond$

Theorem L.1 (Drift crossover)**.** *Assume Assumption S4 and let $i = i_N \to \infty$ be a row sequence.*

(i) Let $0 < \alpha \le 1$ and put

$$\begin{aligned} d &\stackrel{\text{def}}{=} \frac{c_0}{\sqrt{1-\alpha}}, \quad i_{\mathrm{c},\mathfrak{d}} \stackrel{\text{def}}{=} \Big[\frac{BN^{\alpha-1}}{s_2 d^2 \log^2\gamma}\Big]^{1/\alpha}, \quad a_{\mathfrak{d}} \stackrel{\text{def}}{=} d\, a_N \log\gamma \quad (\alpha<1); \\ d &\stackrel{\text{def}}{=} \frac{2}{3}c_0, \quad i_{\mathrm{c},\mathfrak{d}} \stackrel{\text{def}}{=} \frac{B}{s_2 d^2 \log^3\gamma}, \quad a_{\mathfrak{d}} \stackrel{\text{def}}{=} d\, a_N (\log\gamma)^{3/2} \quad (\alpha=1). \end{aligned} \tag{38}$$

If $i/i_{\mathrm{c},\mathfrak{d}} \to x \in (0,\infty)$, then $\sqrt{2}\,\mathfrak{d}_i(0)/a_{\mathfrak{d}} \to 1 + x^{-\alpha/2}$. If $i/i_{\mathrm{c},\mathfrak{d}} \to 0$, then $\sqrt{2}\,\mathfrak{d}_i(0) \sim \sqrt{\mu_i/(\gamma s_2)}$. If $i/i_{\mathrm{c},\mathfrak{d}} \to \infty$, then $\sqrt{2}\,\mathfrak{d}_i(0) \sim a_N \Psi_\alpha(i/N)$, and if moreover $i/N \to 0$,

$$\sqrt{2}\,\mathfrak{d}_i(0) \sim d\, a_N \log\big((N/i)^\alpha\big) \ (\alpha<1), \qquad \sqrt{2}\,\mathfrak{d}_i(0) \sim d\, a_N \big[\log(N/i)\big]^{3/2} \ (\alpha=1).$$

(ii) Let $\alpha > 1$. If $i/i_{\mathrm{c}} \to x \in (0,\infty)$, then $\sqrt{2}\,\mathfrak{d}_i(0)/b_N \to \Phi_{\alpha,\beta}(x^{-\alpha})$. If $i/i_{\mathrm{c}} \to 0$, then $\sqrt{2}\,\mathfrak{d}_i(0) \sim \sqrt{\mu_i/(\gamma s_2)}$. If $i/i_{\mathrm{c}} \to \infty$, then $\sqrt{2}\,\mathfrak{d}_i(0) \sim a_N \Psi_\alpha(i/N)$, and if moreover $i/N \to 0$,

$$\sqrt{2}\,\mathfrak{d}_i(0) \sim d_\alpha \sqrt{c_N K_\alpha}\, N^{-1/2}\, i^{-(\alpha-1)/2}.$$

Full statement and proof: theorem SG.2 of the supplement. Lean: proved: `Muon.PowerLaw.thm_app_pl_drift_i`, `Muon.PowerLaw.thm_app_pl_drift_ii`.

This theorem gives the drift column of table 1 for $\beta > 0$, with the crossover between its two rows. Rows below the crossover index are resolved: $\sqrt{2}\,\mathfrak{d}_i(0) \sim \sqrt{\mu_i/(\gamma s_2)}$, which is $\mathfrak{d}_i \sim \sqrt{\mu_i B_{\text{eff}}/(2N)}$, the square-root law of the large-batch section. Rows above it are not whitened individually, and their drift $\sqrt{2}\,\mathfrak{d}_i(0) \sim a_N \Psi_\alpha(i/N)$, with $a_N = \sqrt{c_N} N^{-\alpha/2}$, is a profile of i/N that does not depend on β; for $i \ll N$ it decays only logarithmically ($\alpha \le 1$) or like $i^{-(\alpha-1)/2}$ ($\alpha > 1$), far more slowly than $\sqrt{\mu_i}$. For $\alpha \le 1$ the two laws add at the crossover, so that $\sqrt{2}\,\mathfrak{d}_i(0)/a_{\mathfrak{d}} \to 1 + x^{-\alpha/2}$; for $\alpha > 1$ the crossover is described by the profile $\Phi_{\alpha,\beta}$, which keeps the momentum. This sketch also proves the drift half of the uniform approximation behind corollary L.1.

Proof sketch. The drift is the sum of a head and a tail, each with an explicit limit (appendix G.5). Changing variables from u to the participation m gives (lemma SG.8)

$$\sqrt{2}\,\mathfrak{d}_i(0) = \frac{1}{\pi\gamma}\int_0^\gamma \mathcal{A}(m)\,\mathcal{B}_i(m)\,E(m)\vartheta_N(m)\,\mathrm{d}m,$$

$$\mathcal{B}_i(\gamma p) = a_N \sqrt{\frac{1-p}{p\,V_N(p)}}\,\frac{V_N(p)}{V_N(p)+x^\alpha}, \quad x = \frac{i}{N},$$

where $\vartheta_N = -\frac{\mathrm{d}}{\mathrm{d}m}\log s_N$ comes from the Jacobian of $u^2 = s_N(m)$, and $E\vartheta_N$ is within $C(1/m + 1/(\gamma-m))$ of one for $m \ge M$. On $m > M$ the averaging of appendix G.5 replaces $\mathcal{A}$ by $\sqrt{2\pi}$, the factor $E\vartheta_N$ is replaced by one, and V_N by V_α, all with relative error $C/M + o(1)$ uniformly in the row; the tail is therefore $a_N \Psi_\alpha^{[M/\gamma]}(x)$, the profile with the integral restricted to $p \ge M/\gamma$ (lemma SG.10). The head $m \le M$ is at most $C_M\sqrt{\mu_i/\gamma}$ for every row and equals $\sqrt{\mu_i/(\gamma s_2)}(1+o(1))$ on resolved rows (lemma SG.11). For bulk rows, $i/N \to x_0 > 0$, the head is $o(a_N)$ and $\sqrt{2}\,\mathfrak{d}_i(0) \sim a_N \Psi_\alpha(x_0)$.

For $\alpha \le 1$ the crossover is the row at which head and tail are equal. With $c_N = 1$, write the head in units of a_N as

$$W \stackrel{\text{def}}{=} \frac{\sqrt{\mu_i/(\gamma s_2)}}{a_N} = s_2^{-1/2} e^{(T-\log\gamma)/2}, \qquad T = \log(x^{-\alpha}) \ (\alpha<1), \quad T = \log(1/x) \ (\alpha=1),$$

and recall $\Psi_\alpha(x) \sim dT^\nu$ with $\nu = 1$ for $\alpha < 1$ and $\nu = 3/2$ for $\alpha = 1$ (proposition L.1). The cutoff $p \geq M/\gamma$ removes a part of the profile that is negligible as long as $W \lesssim (\log\gamma)^\nu$, and when $W \gg (\log\gamma)^\nu$ both profiles are $o(W)$ (lemma SG.12). In every case, after $M \to \infty$,

$$\sqrt{2}\,\mathfrak{d}_i(0) = a_N\big(W + \Psi_\alpha(x)\big)\big(1 + o(1)\big),$$

which is the drift approximant of eq. (SG.29). The index $i_{\mathrm{c},\mathfrak{d}}$ is where $W = d(\log\gamma)^\nu$: at $i/i_{\mathrm{c},\mathfrak{d}} \to x_0$ one has $W = d(\log\gamma)^\nu x_0^{-\alpha/2}(1 + o(1))$ and $T = \log\gamma + O(\log\log\gamma)$, so both terms are multiples of $a_\mathfrak{d} = d\,a_N(\log\gamma)^\nu$, which gives $1 + x_0^{-\alpha/2}$. On the rows $i \asymp i_\#$, where $i_\# = (c_N\kappa)^{1/\alpha}$ is the no-momentum threshold of proposition L.2, the drift exceeds the no-momentum drift scale $(\gamma\kappa)^{-1/2}$ by a factor of order $\log\gamma$, which comes from the growth $\log Q_m \sim 2|\log\beta|\,m$ of the momentum participation (remark SG.2); this is why $i_{\mathrm{c},\mathfrak{d}}$ lies below the no-momentum threshold $i_\#$ by a factor of order $(\log\gamma)^{-2/\alpha}$ (proposition L.2). The main-text display eq. (15), which also writes $i_\#$, is for $\beta > 0$ and $\alpha \leq 1$ of the order of $i_{\mathrm{c},\mathfrak{d}}$ itself; in this chapter $i_\#$ always means the no-momentum threshold.

For $\alpha > 1$ the crossover sits at participation of order one, where the weights cannot be averaged, so its profile keeps the momentum. At $i \asymp B$ put $y = (i_\mathrm{c}/i)^\alpha$. The scaling $c_N\Lambda_N(m) \sim (Bm/K_\alpha)^\alpha$ of appendix G.5 gives $\Lambda_N(m)\mu_i \to ym^\alpha$ and $m\Lambda_N'(m)/\Lambda_N(m) \to \alpha$, so that on compact sets of m the drift integrand, divided by b_N, converges to that of (eq. (SG.46))

$$\Phi_{\alpha,\beta}(y) = \frac{1}{\pi}\int_0^\infty \mathcal{A}(m)E(m)\vartheta_\infty(m)\,m^{-\frac{\alpha+1}{2}}\,\frac{ym^\alpha}{1+ym^\alpha}\,\mathrm{d}m,$$

while the head bounds and the tail bound $CK_\alpha^{-\alpha/2}M^{-(\alpha-1)/2}$ make small and large m negligible (lemma SG.13). On the two windows the head and the tail are estimated directly, as for $\alpha \leq 1$ (lemma SG.13), and the identities $b_N\sqrt{y} = \sqrt{\mu_i/\gamma}$ and $b_N d_\alpha y^{(\alpha-1)/(2\alpha)} = a_N d_\alpha\sqrt{K_\alpha}(i/N)^{-(\alpha-1)/2}$ show that the results match the two endpoints of $\Phi_{\alpha,\beta}$ (proposition L.1(ii)). ◇

Theorem L.2 (Volatility crossover and bulk profile)**.** *Assume Assumption S4 and let $i = i_N$ be a row sequence.*

(i) (Bulk.) If $i/N \to x \in (0,1]$ and $\ell = \ell_N$ are integers with $\ell/\gamma \to \tau \in (0,1)$, then

$$\overline{C}_{i,\ell}(0) \longrightarrow C_{\alpha,x}(\tau), \qquad \frac{\mathfrak{v}_i(0)}{\gamma} \longrightarrow \Gamma_\alpha(x),$$

locally uniformly in $x \in (0,1]$ and $\tau \in (0,1)$.

(ii) Let $0 < \alpha \leq 1$, let $i \to \infty$, and put

$$i_{\mathrm{c},\mathfrak{v}} \overset{\mathrm{def}}{=} c_0^{2/\alpha}\,i_{\mathrm{c},\mathfrak{d}} = \begin{cases} N\big[(1-\alpha)/(s_2\gamma\log^2\gamma)\big]^{1/\alpha}, & \alpha < 1, \\ 9B/(4s_2\log^3\gamma), & \alpha = 1. \end{cases} \tag{39}$$

If $i/i_{\mathrm{c},\mathfrak{v}} \to y \in (0,\infty)$, then $s_2\,\mathfrak{v}_i(0) \to 1 + 2c_0y^{\alpha/2} + y^\alpha$, uniformly for y in compact subsets of $(0,\infty)$. If $i/i_{\mathrm{c},\mathfrak{v}} \to 0$, then $\mathfrak{v}_i(0) \to 1/s_2$. If $i/i_{\mathrm{c},\mathfrak{v}} \to \infty$, then $\mathfrak{v}_i(0) \sim \gamma\Gamma_\alpha(i/N)$, and if moreover $i/N \to 0$,

$$\mathfrak{v}_i(0) \sim \frac{\gamma(i/N)^\alpha\big[\log((N/i)^\alpha)\big]^2}{1-\alpha} \quad (\alpha < 1), \qquad \mathfrak{v}_i(0) \sim \frac{4}{9}\,\frac{i}{B}\,\big[\log(N/i)\big]^3 \quad (\alpha = 1).$$

(iii) Let $\alpha > 1$ and $i \to \infty$. If $i/i_\mathrm{c} \to X_0 \in (0,\infty)$, then $\mathfrak{v}_i(0) \to G_{\alpha,\beta}(X_0)$. If $i/i_\mathrm{c} \to 0$, then $\mathfrak{v}_i(0) \to 1/s_2$. If $i/i_\mathrm{c} \to \infty$, then $\mathfrak{v}_i(0) \sim \gamma\Gamma_\alpha(i/N)$, and if moreover $i/N \to 0$,

$$\mathfrak{v}_i(0) \sim \frac{4K_\alpha^2}{\alpha-1}\,\frac{i}{B}.$$

Full statement and proof: theorem SG.3 of the supplement. Lean: proved: `Muon.PowerLaw.thm_app_pl_volatility_i_part`, `Muon.PowerLaw.thm_app_pl_volatility_iii_part`, +3 more; foundation results: `GaussianOpNorm`.

This theorem gives the volatility column of table 1 for $\beta > 0$. On resolved rows $\mathfrak{v}_i(0) \to 1/s_2 = B_{\text{eff}}/B$: the row is fully whitened at every step, and (as the proof shows for $\alpha > 1$) momentum correlates the updates ℓ steps apart by $\beta^{|\ell|}$. On unresolved rows $\mathfrak{v}_i(0) \sim \gamma\Gamma_\alpha(i/N)$, which grows with $\gamma = N/B$, and part (i) identifies the correlation of updates $\tau\gamma$ steps apart in a bulk row as $C_{\alpha,x}(\tau)$. For $\alpha \le 1$ the crossover profile $1 + 2c_0 y^{\alpha/2} + y^\alpha$ is not a sum of the two endpoint laws: it is the second moment of a sum of two correlated terms. For $\alpha > 1$ the crossover profile is $G_{\alpha,\beta}$. For $\alpha \le 1$ the crossover index differs from the drift index by the fixed factor $c_0^{2/\alpha}$. This sketch also proves the volatility half of the uniform approximation behind corollary L.1.

Proof sketch. The zero-ridge volatility is a positive double integral over two participations, summed over the lags (proposition SG.4):

$$\mathfrak{v}_i(0) = \frac{1}{\pi^2} \sum_{\ell \in \mathbb{Z}} \int_0^\gamma \int_0^\gamma W_i(m_1) W_i(m_2) \, \frac{\mathsf{D}_\ell(m_1, m_2)}{\Delta_\ell^N(m_1, m_2)} \, \mathrm{d}m_1 \, \mathrm{d}m_2.$$

It is obtained from the closed-form lag kernel of appendix G.3 by writing the regularized filter as an integral of resolvents and changing variables to the participation. Here W_i is the weight of the row, D_ℓ correlates two buffers ℓ steps apart through the past batches they share, and the determinant $\Delta_\ell^N \in (0, 1]$ of the kernel is close to one when the smaller participation is much smaller than the larger one and than γ (lemma SG.14).

In the bulk, the lag ℓ couples the participation m to $m + \ell$, because shifting the buffer by ℓ steps moves every past batch by ℓ places. Exactly, $\mathsf{D}_\ell = (r_\beta(q_1 e^t) - r_\beta(q_1))/(2\sinh(t/2))$ with $t = \log(\beta^{2\ell} q_2/q_1)$, and for large u, $r_\beta(e^u)$ is within $Ce^{-u/2}$ of a function that grows by exactly one when u grows by $2|\log\beta|$. The kernel $\mathsf{D}_\ell/(\pi^2 E(m_1)E(m_2))$ therefore concentrates on the shell $m_2 = m_1 + \ell + O(1)$ and carries mass one per unit of m_1, by the identity (lemma SG.22)

$$\int_{\mathbb{R}} \frac{t}{2\sinh(t/2)} \, \mathrm{d}t = \pi^2.$$

In the coordinates $m = \gamma p$ and $\ell = \tau\gamma$ the row factors converge to the function $A_x(p)$ and the determinant to its continuum limit $\Delta_{\mathrm{c}}(p, p + \tau)$, both of the sketch for proposition L.1, so $\overline{C}_{i,\ell}(0) \to C_{\alpha,x}(\tau)$ (the identity of lemma SG.21). The lag sum is then a Riemann sum with mesh $1/\gamma$; lags beyond $(1+\eta)\gamma$ contribute $O(\gamma\beta^{\ell-\gamma})$ each, because the later buffer then shares no whitened batch, and $\mathfrak{v}_i(0)/\gamma \to 2\int_0^1 C_{\alpha,x} = \Gamma_\alpha(x)$.

For $\alpha \le 1$ the mesoscopic rows are governed by the coefficient with every determinant replaced by one, which is a perfect square. With one Gaussian g shared by the two buffers and $\mathcal{Z}_g(q) = \sqrt{q}\sum_a w_a \, g/(1 + w_a^2 q g^2)$, the lag sum factorizes as $\sum_\ell \mathsf{D}_\ell(m_1, m_2) = \mathbb{E}[\mathcal{Z}_g(Q_{m_1})\mathcal{Z}_g(Q_{m_2})]$ (lemma SG.15). The flat coefficient is therefore $\mathbb{E}\,\mathcal{V}_i(g)^2$ with $\mathcal{V}_i(g) = \pi^{-1}\int_0^\gamma W_i \, \mathcal{Z}_g(Q_m) \, \mathrm{d}m$, and it lies below $\mathfrak{v}_i(0)$. The head of $\mathcal{V}_i$ is linear in g, and in the tail $\mathcal{Z}_g/E$ averages to $\pi \operatorname{sign}(g)$, as $\mathcal{A}$ does for the drift, so (lemma SG.18)

$$\left\| \mathcal{V}_i(g) - \frac{g}{\sqrt{s_2}} - b \operatorname{sign}(g) \right\|_{L^2(g)} = o(1 + b), \qquad b = \frac{\sqrt{\gamma}\, x^{\alpha/2} \Psi_\alpha(x)}{c_0}.$$

With $\mathbb{E}\, g^2 = \mathbb{E}\operatorname{sign}(g)^2 = 1$ and $\mathbb{E}[g \operatorname{sign}(g)] = c_0$ the flat coefficient is $1/s_2 + (2c_0/\sqrt{s_2})\, b + b^2$ up to $o((1+b)^2)$. The exact coefficient exceeds it by $o((1+b)^2)$: off the diagonal the determinant is close to one, the strip $m_1 \asymp m_2$ carries little mass, and the top of the participation range contributes $O(\gamma x^\alpha)$ (lemmas SG.17, SG.19 and SG.20). At $i/i_{\mathrm{c},\mathfrak{v}} \to y$ one has $b \to y^{\alpha/2}/\sqrt{s_2}$, which gives $s_2 \mathfrak{v}_i(0) \to 1 + 2c_0 y^{\alpha/2} + y^\alpha$.

For $\alpha > 1$ the rows near i_{c} have a limit profile in which nothing is averaged. With $X = i/i_{\mathrm{c}}$ and m of order one, the participation scaling of appendix G.5 gives $f_i(m) \to \zeta_X(m)$, the normalized sum T_N becomes the integral $t_\alpha(m_1, m_2)$, and $\Delta_\ell^N \to \Delta_\ell^\infty$, so $\mathfrak{v}_i(0) \to G_{\alpha,\beta}(X)$ (lemma SG.23). Along rows with $X \to 0$ or $X \to \infty$, the rescalings that give the ends of $G_{\alpha,\beta}$ (proposition L.1(ii)), carried out with the dimension errors, give the resolved limit $1/s_2$ and the mesoscopic law $4K_\alpha X/(\alpha - 1) = 4K_\alpha^2 i/((\alpha - 1)B)$. ◇

Corollary L.1 (Crossover windows and joint endpoints)**.** *Assume Assumption S4, let $i = i_N \to \infty$ be a row sequence, and let $i_\star \overset{\text{def}}{=} i_{\mathrm{c},\mathfrak{v}}$ for $\alpha \le 1$ and $i_\star \overset{\text{def}}{=} i_{\mathrm{c}}$ for $\alpha > 1$.*

(i) If $\alpha \le 1$ and $i/i_\star \to x \in (0,\infty)$, then

$$\frac{\sqrt{2}\,\mathfrak{d}_i(0)}{a_{\mathfrak{d}}} \to 1 + x^{-\alpha/2}, \qquad s_2\,\mathfrak{v}_i(0) \to 1 + 2x^{\alpha/2} + c_0^{-2}x^{\alpha}, \qquad \frac{\mu_i}{c_N i_\star^{-\alpha}} \to x^{-\alpha}.$$

(ii) The joint endpoint equivalents are

	$i/i_\star \to 0$	$i/i_\star \to \infty$
$\sqrt{2}\,\mathfrak{d}_i(0)$	$\sqrt{\mu_i/(\gamma s_2)}$	$a_N \Psi_\alpha(i/N)$
$\mathfrak{v}_i(0)$	$1/s_2$	$\gamma\,\Gamma_\alpha(i/N)$
μ_i	$c_N i^{-\alpha}$ *(exact)*	$c_N i^{-\alpha}$ *(exact)*

(40)

(iii) One has $i_\star/N \to 0$. There are row sequences $i \to \infty$ in both windows $i/i_\star \to 0$ and $i/i_\star \to \infty$ if and only if

$$\frac{BN^{\alpha-1}}{\log^2\gamma} \to \infty \ \ (\alpha<1), \qquad \frac{B}{\log^3\gamma} \to \infty \ \ (\alpha=1), \qquad B \to \infty \ \ (\alpha>1), \tag{41}$$

and then $1 \ll i_\star \ll N$.

(iv) For $\alpha > 1$, equating the two mesoscopic endpoint equivalents gives the indices $i_{\mathrm{c}}/(s_2 d_\alpha^2)$ for the drift and $B(\alpha-1)/(4s_2K_\alpha^2)$ for the volatility, whose ratio is $(\alpha-1)d_\alpha^2/(4K_\alpha)$; all indices of the corollary are independent of c_N.

Full statement and proof: corollary SG.1 of the supplement. Lean: proved: `Muon.PowerLaw.cor_app_pl_endpoints_i`, `Muon.PowerLaw.cor_app_pl_endpoints_ii`, +2 more; foundation results: `GaussianOpNorm`.

This corollary is the MUON part of table 1 for $\beta > 0$: it reads the drift and volatility crossovers in one coordinate $i/i_\star$. Below $i_\star$ the row keeps the resolved kernels of the large-batch section, $\sqrt{2}\,\mathfrak{d}_i \sim \sqrt{\mu_i/(\gamma s_2)}$ and $\mathfrak{v}_i \to 1/s_2 = B_{\mathrm{eff}}/B$; above it both kernels follow the bulk profiles of i/N, while the covariance $\mu_i = c_N i^{-\alpha}$ is exact throughout. Part (iii) says when both windows contain rows: always for $\alpha > 1$, and for $\alpha < 1$ only if the batch exceeds $N^{1-\alpha}$ up to logarithms; the crossover always lies far below N. Part (iv) records that for $\alpha > 1$ the points where the endpoint laws are equal differ from i_{c} only by constants, and that no index depends on the amplitude c_N.

Proof sketch. Both theorems are equivalences along row sequences, and they come from one approximation that is uniform over all rows (theorem SG.1). For $\alpha \le 1$ the approximants are (eq. (SG.29))

$$\sqrt{2}\,\widehat{\mathfrak{d}}_i = \sqrt{\frac{\mu_i}{\gamma s_2}} + a_N \Psi_\alpha\Big(\frac{i}{N}\Big), \qquad \widehat{\mathfrak{v}}_i = \frac{1}{s_2} + \frac{2c_0}{\sqrt{s_2}}\sqrt{\gamma\Gamma_\alpha\Big(\frac{i}{N}\Big)} + \gamma\Gamma_\alpha\Big(\frac{i}{N}\Big),$$

and for $\alpha > 1$ they are $b_N \Phi_{\alpha,\beta}((i_{\mathrm{c}}/i)^\alpha)$ and $G_{\alpha,\beta}(i/i_{\mathrm{c}})$, each multiplied by a factor that tends to one when $i/N \to 0$ and that turns the mesoscopic law into the bulk profile when i/N stays positive (eq. (SG.30)). For every sequence $n_N \to \infty$,

$$\sup_{n_N \le i \le N} \max\Big\{\Big|\frac{\mathfrak{d}_i(0)}{\widehat{\mathfrak{d}}_i} - 1\Big|,\ \Big|\frac{\mathfrak{v}_i(0)}{\widehat{\mathfrak{v}}_i} - 1\Big|\Big\} \longrightarrow 0.$$

The drift half is the head-plus-tail argument of theorem L.1. The volatility half is the argument of theorem L.2: the quadratic $\widehat{\mathfrak{v}}_i$ is the flat square of that sketch, with b replaced by $\sqrt{\gamma\Gamma_\alpha(i/N)}$, and the two agree on mesoscopic rows because $\Gamma_\alpha(x) \sim x^\alpha \Psi_\alpha(x)^2/c_0^2$ (lemma SG.21). A failure of uniformity would produce a row sequence along which a ratio stays away from one, so the uniform statement follows from the sequential one.

Parts (i) and (ii) are the two theorems in the coordinate $i/i_\star$. For $\alpha \le 1$ the volatility index is $i_{\mathrm{c},\mathfrak{v}} = c_0^{2/\alpha} i_\star$, so $i/i_\star \to x$ means $i/i_{\mathrm{c},\mathfrak{v}} \to y = c_0^{-2/\alpha}x$, and then $2c_0 y^{\alpha/2} = 2x^{\alpha/2}$ and $y^\alpha = c_0^{-2}x^\alpha$, which is (i). In the table the left column is the resolved case of both theorems, the two indices differing by a fixed factor. In the right column the drift is the unresolved case of theorem L.1, and the volatility follows from the uniform approximation, since $\widehat{\mathfrak{v}}_i \sim \gamma\Gamma_\alpha(i/N)$ beyond $i_\star$: for

$\alpha \le 1$ because $\gamma\Gamma_\alpha(i/N) \to \infty$ there, and for $\alpha > 1$ by the end $G_{\alpha,\beta}(X) \sim 4K_\alpha X/(\alpha-1)$ of the crossover profile.

Parts (iii) and (iv) are computations with the indices. The definitions give

$$\frac{i_\star}{N} = \left(\gamma s_2 d^2 \log^2\gamma\right)^{-1/\alpha} (\alpha < 1), \quad \frac{i_\star}{N} = \left(\gamma s_2 d^2 \log^3\gamma\right)^{-1} (\alpha = 1), \quad \frac{i_\star}{N} = \frac{1}{K_\alpha \gamma} (\alpha > 1),$$

which tend to zero. A row sequence $i \to \infty$ with $i/i_\star \to 0$ exists exactly when $i_\star \to \infty$, and $i = \lfloor\sqrt{i_\star}\rfloor$ realizes it; $i = N$ realizes the other window. The condition $i_\star \to \infty$ is eq. (41). For (iv), equating $\sqrt{\mu_i/(\gamma s_2)}$ with $d_\alpha\sqrt{c_N K_\alpha} N^{-1/2} i^{-(\alpha-1)/2}$ gives $i = i_{\mathrm{c}}/(s_2 d_\alpha^2)$, and equating $1/s_2$ with $4K_\alpha^2 i/((\alpha-1)B)$ gives $i = B(\alpha-1)/(4s_2K_\alpha^2)$; c_N cancels in both. ◇

Proposition L.2 (Reference scales)**.** *Assume Assumption S4. For $\gamma > 1$ let $\kappa \overset{\text{def}}{=} \Lambda_N(1)$ be the unique solution of*

$$\sum_{i=1}^{N} \frac{\kappa\mu_i}{1+\kappa\mu_i} = B,$$

and put $i_\# \overset{\text{def}}{=} (c_N\kappa)^{1/\alpha}$. Then

$$c_N\kappa \sim \begin{cases} (1-\alpha)BN^{\alpha-1}, & \alpha < 1, \\ B/\log\gamma, & \alpha = 1, \\ (B/K_\alpha)^\alpha, & \alpha > 1, \end{cases} \qquad \frac{i_{\mathrm{c},\mathfrak{d}}}{i_\#} \sim \begin{cases} \left(\pi/(2s_2\log^2\gamma)\right)^{1/\alpha}, & \alpha < 1, \\ 9\pi/(8s_2\log^2\gamma), & \alpha = 1, \end{cases} \quad \frac{i_{\mathrm{c}}}{i_\#} \to 1 \ (\alpha > 1), \tag{42}$$

and $i_{\mathrm{c},\mathfrak{v}}/i_\# \sim (s_2\log^2\gamma)^{-1/\alpha}$ for $\alpha < 1$ and $\sim 9/(4s_2\log^2\gamma)$ for $\alpha = 1$.

Full statement and proof: proposition SG.6 of the supplement. Lean: proved: `Muon.PowerLaw.kappaOf`, `Muon.PowerLaw.iSharp`, +3 more.

The scalar κ is the level at which the soft count of output directions equals the batch size: $\sum_i \kappa\mu_i/(1+\kappa\mu_i) = B$, the rank of one gradient. It is the resolution scalar of the no-momentum row of table 1, and $i_\# = (c_N\kappa)^{1/\alpha}$ is the row at which $\kappa\mu_i = 1$. The proposition compares this threshold with the crossover indices of the momentum coefficients. For $\alpha > 1$ the momentum crossover sits at the threshold, $i_{\mathrm{c}} \sim i_\#$. For $\alpha \le 1$ it sits below it by a factor of order $(\log\gamma)^{-2/\alpha}$. Two indices therefore carry the name $i_\#$ in the main text. The threshold of this proposition governs the no-momentum row of the table. The display eq. (15), which cites this proposition, has the form of the momentum crossover index, $i_{\mathrm{c},\mathfrak{d}}$ for $\alpha \le 1$ and i_{c} for $\alpha > 1$, up to constant factors; for $\alpha \le 1$ it is thus smaller than the threshold $(c_N\kappa)^{1/\alpha}$ by the logarithmic factor above, and for $\alpha > 1$ the two agree up to constants. In this chapter $i_\#$ always denotes $(c_N\kappa)^{1/\alpha}$.

Proof sketch. The scalar κ is the participation coordinate at $m = 1$, and the continuum law of appendix G.5 evaluates it. By definition $\kappa = \Lambda_N(1)$, so $V \overset{\text{def}}{=} c_N\kappa N^{-\alpha}$ solves $S_N(V) = 1/\gamma$. The Riemann sum S_N is within $1/N$ of S_α, hence

$$S_\alpha(V) \in \frac{1}{\gamma}\left[1, 1+\frac{1}{B}\right],$$

so $V \to 0$ and the small-V inversion of S_α applies (lemmas SG.5 and SG.6). It gives $c_N\kappa = N^\alpha V \sim N^\alpha(1-\alpha)/\gamma$, $N/(\gamma\log\gamma)$ and $N^\alpha(K_\alpha\gamma)^{-\alpha}$ in the three cases, which is the first display.

The ratios are then read off from the definitions of the crossover indices. With $d^2 = c_0^2/(1-\alpha) = 2/(\pi(1-\alpha))$ for $\alpha < 1$ and $d^2 = \frac{4}{9}c_0^2 = 8/(9\pi)$ for $\alpha = 1$,

$$\frac{i_{\mathrm{c},\mathfrak{d}}^\alpha}{c_N\kappa} \sim \frac{BN^{\alpha-1}}{s_2 d^2\log^2\gamma\cdot(1-\alpha)BN^{\alpha-1}} = \frac{\pi}{2s_2\log^2\gamma} \quad (\alpha < 1), \qquad \frac{i_{\mathrm{c},\mathfrak{d}}}{i_\#} \sim \frac{9\pi}{8s_2\log^2\gamma} \quad (\alpha = 1),$$

while $i_{\mathrm{c}} = B/K_\alpha \sim i_\#$ for $\alpha > 1$, and $i_{\mathrm{c},\mathfrak{v}} = c_0^{2/\alpha} i_{\mathrm{c},\mathfrak{d}}$ gives the volatility ratios. The logarithm for $\alpha \le 1$ is not an effective-batch effect: on the rows $i \asymp i_\#$ the momentum drift exceeds the no-momentum scale $(\gamma\kappa)^{-1/2}$ by a factor of order $\log\gamma$, because the momentum participation grows like

$\log Q_m \sim 2|\log\beta|\, m$, and a fixed change of the batch size cannot remove a logarithm (remark SG.2). $\diamond$

In the remark, $m(s) \in (0,\gamma)$ is the participation of the branch at the spectral parameter $-s$, the root of $s = (\gamma - m)m/(Q_m \Lambda_N(m))$ (proposition SG.1), and $\mathfrak{v}_i^{(0)}(\varrho) = \Lambda_N(m(\varrho))\mu_i/(1+\Lambda_N(m(\varrho))\mu_i)$ is the zero-lag volatility coefficient of row i at the ridge ϱ (eq. (SG.7)).

Remark L.1 (No momentum)**.** *At $\beta = 0$ the weights are $w_0 = 1$, so $F_\beta = L_\beta = J$ and $r_0(q) = qJ(q)$ increases from 0 to 1, not to ∞; the participation is confined to $m < 1$, and at zero ridge $m(s) \to 1$ when $\gamma > 1$. The computation of proposition SG.1(ii), which uses only the equations of the branch, then gives $\Lambda_N(m(s)) \to \kappa$, the root of proposition L.2, instead of infinity, and by eq.* (SG.7)

$$\lim_{\varrho\downarrow 0} \mathfrak{v}_i^{(0)}(\varrho) = \frac{\kappa\mu_i}{1+\kappa\mu_i} = \frac{1}{1+(i/i_\#)^\alpha}, \qquad \sum_i \lim_{\varrho\downarrow 0} \mathfrak{v}_i^{(0)}(\varrho) = B.$$

All lag coefficients with $\ell \geq 1$ vanish (theorem SD.7(iv)), so $\mathfrak{v}_i = \mathfrak{v}_i^{(0)}$, and the volatility crossover at $\beta = 0$ is this exact profile, with critical index $i_\#$ and the equivalents eq. (42). The drift at $\beta = 0$ is the single integral eq. (SG.7) *with $L_\beta/F_\beta = 1$; its two endpoint laws, on either side of $i_\#$, are proposition L.3 below. These no-momentum profiles are obtained by setting $\beta = 0$ before the ridge and dimension limits. For every fixed $\beta > 0$, r_β is unbounded, $\mathfrak{v}_i^{(0)}(0) = 1$ for every row (proposition SG.2), and the profiles of theorem SG.1 hold; neither family is a limit of the other under an exchange of $\beta \downarrow 0$ with the ridge and dimension limits.*

Full statement and proof: remark SG.1 of the supplement. Lean: not formalized (remark).

This remark explains the no-momentum row of table 1 and why it differs from the rows with momentum. Without momentum the buffer is a single gradient, whose participation cannot exceed one, so at zero ridge the update whitens only the directions above the level $1/\kappa$: the zero-lag volatility of row i is the soft indicator $1/(1+(i/i_\#)^\alpha)$, the coefficients sum to B, and there are no lag terms. With any fixed momentum the participation is unbounded, every row has zero-lag coefficient one at zero ridge, and the crossover moves to the indices of theorems L.1 and L.2. The two families are not limits of each other, because the limit $\beta \downarrow 0$ does not commute with the ridge and dimension limits.

The constant of the proposition is (eq. (SG.39))

$$A_\alpha = c_0 \quad (0 < \alpha \leq 1), \qquad A_\alpha = \frac{2\alpha}{\pi(\alpha+1)} \int_0^\infty q^{-3/2} \big(qJ(q)\big)^{\frac{\alpha+1}{2}}\, \mathrm{d}q \quad (\alpha > 1).$$

Proposition L.3 (No-momentum endpoints)**.** *Let $\beta = 0$, let $\alpha > 0$ be fixed, and let $N, B \to \infty$ with $\gamma = N/B \to \infty$, $\Sigma_{\text{in}} = \mathrm{Id}_N$, $N_{\text{out}} = N_{\text{in}} = N$ and $\mu_i = i^{-\alpha}$. For $\gamma > 1$ let κ and $i_\# = \kappa^{1/\alpha}$ be as in proposition L.2. At each such dimension the limits $\mathfrak{d}_i(0) \overset{\text{def}}{=} \lim_{\varrho\downarrow 0} \mathfrak{d}_i(\varrho)$ and $\mathfrak{v}_i(0) \overset{\text{def}}{=} \lim_{\varrho\downarrow 0} \mathfrak{v}_i(\varrho)$ exist, and:*

(i) for every i,

$$\mathfrak{v}_i(0) = \frac{\kappa\mu_i}{1+\kappa\mu_i} = \frac{1}{1+(i/i_\#)^\alpha}, \qquad \sum_{i=1}^N \mathfrak{v}_i(0) = B;$$

(ii) for every row sequence $1 \leq i = i_N \leq N$,

$$\frac{i}{i_\#} \to 0 \implies \mathfrak{d}_i(0) \sim \sqrt{\frac{\mu_i}{2\gamma}}, \qquad \frac{i}{i_\#} \to \infty \implies \mathfrak{d}_i(0) \sim A_\alpha\, \mu_i \sqrt{\frac{\kappa}{2\gamma}}.$$

Full statement and proof: proposition SG.8 of the supplement. Lean: proved: `Muon.PowerLaw.aAlpha`, `Muon.PowerLaw.mOf0`, +6 more; foundation results: `GaussianOpNorm`.

This proposition gives the no-momentum row of table 1 and the statement of the main text that unresolved modes at $\beta = 0$ have the drift of SGD. The volatility is exact at every dimension: row i carries the share $\kappa\mu_i/(1+\kappa\mu_i)$ of the B directions that one gradient spans. The drift has two endpoint laws on either side of the threshold $i_\#$ of proposition L.2. For $i \ll i_\#$, $\mathfrak{d}_i(0) \sim \sqrt{\mu_i/(2\gamma)}$

is the square-root law with $B_{\mathrm{eff}} = B$. For $i \gg i_{\#}$, $\mathfrak{d}_i(0) \sim A_\alpha \mu_i \sqrt{\kappa/(2\gamma)}$ is linear in μ_i, with the constant $A_\alpha = c_0$ for $\alpha \le 1$ and the integral eq. (SG.39) for $\alpha > 1$.

Proof sketch. The participation coordinate of appendix G.5 applies at $\beta = 0$, with the participation confined to $(0, 1)$. The computation of the branch uses only its equations, so with $m(s) \in (0, 1)$ defined by $s_N(m(s)) = s$ and $L_\beta/F_\beta = 1$,

$$\mathfrak{v}_i^{(0)}(\varrho) = f_i\big(m(\varrho)\big), \qquad \mathfrak{d}_i(\varrho) = \frac{\sqrt{2}}{\pi\gamma} \int_0^\infty f_i\big(m(\varrho + u^2)\big)\, \mathrm{d}u.$$

Now $r_0(q) = qJ(q)$ increases only to one, so $Q_m \to \infty$ as $m \uparrow 1$ while $\Lambda_N(m)$ stays bounded, and $s_N(m) \to 0$; hence $m(s) \to 1$ and $\Lambda_N(m(s)) \uparrow \Lambda_N(1) = \kappa$ as $s \downarrow 0$. Without momentum successive buffers are independent, so all lag coefficients vanish (theorem SD.7(iv)) and $\mathfrak{v}_i(\varrho) = \mathfrak{v}_i^{(0)}(\varrho) \to \kappa\mu_i/(1 + \kappa\mu_i)$, whose sum over i is $Bm_N(\kappa) = B$; this is (i). The drift integrand is nonincreasing in u with total variation at most one, and an integration by parts gives $|\mathfrak{d}_i(\varrho) - \mathfrak{d}_i(0)| \le \sqrt{2}\sqrt{\varrho}/(\pi\gamma)$.

On resolved rows the drift is the head Lorentzian. If $i/i_{\#} \to 0$ then $\kappa\mu_i \to \infty$, and for a fixed small δ the range $m \le \delta$ covers all $u \ge \sqrt{s_N(\delta)} = o(\sqrt{\gamma\mu_i})$. There $\tilde{s}_1 = \gamma(1 + O(\delta) + o(1))$ and

$$\int_{\sqrt{s_N(\delta)}}^\infty \frac{\tilde{s}_1 \mu_i}{u^2 + \tilde{s}_1 \mu_i}\, \mathrm{d}u = \frac{\pi}{2} \sqrt{\gamma\mu_i}\, \big(1 + O(\delta) + o(1)\big),$$

which after the factor $\sqrt{2}/(\pi\gamma)$ and $\delta \downarrow 0$ gives $\sqrt{\mu_i/(2\gamma)}$.

On unresolved rows the share is linear in μ_i, and the drift is a fixed integral over the participation. If $i/i_{\#} \to \infty$ then $\kappa\mu_i \to 0$, and since $\Lambda_N \le \kappa$ one has $f_i = \Lambda_N \mu_i (1 + O(\kappa\mu_i))$ uniformly. Changing variables to m gives

$$\frac{\mathfrak{d}_i(0)}{\mu_i \sqrt{\kappa/(2\gamma)}} = \big(1 + o(1)\big) \frac{1}{\pi} \int_0^1 \sqrt{\frac{\Lambda_N(m)}{\kappa}} \sqrt{1 - \frac{m}{\gamma}} \sqrt{\frac{m}{Q_m}}\, \vartheta_N(m)\, \mathrm{d}m.$$

For fixed m, $\Lambda_N(m)/\kappa \to m^\nu$ and $m\Lambda_N'/\Lambda_N \to \nu$, with $\nu = 1$ for $\alpha \le 1$ and $\nu = \alpha$ for $\alpha > 1$, by the continuum laws of appendix G.5. Dominated convergence then gives an integral in $q = Q_m$ that equals $\pi^{-1} \int_0^\infty q^{-1/2} J(q)\, \mathrm{d}q = \sqrt{2\pi}/\pi = c_0$ when $\nu = 1$; for $\nu = \alpha$ an integration by parts turns it into $\frac{2\alpha}{\pi(\alpha+1)} \int_0^\infty q^{-3/2} (qJ(q))^{(\alpha+1)/2}\, \mathrm{d}q = A_\alpha$. This is (ii). ◇

In the corollary, $Q_\varrho(M_t)$ is the regularized update and $[\varrho_-, \varrho_+] \subset (0, \infty)$ the ridge range of Assumption S3, and ϑ_ϱ is the ratio of the drift coefficient to its oversampled profile, $\vartheta_\varrho(\mu) = \sqrt{s_2\gamma_{\mathrm{in}}}\, \mathfrak{d}_\varrho(\mu)/\mathfrak{d}^{\mathrm{os}}_{\hat{\nu},r}(\mu; \varrho/(s_2\gamma_{\mathrm{in}}))$ for $\mu > 0$ and $\vartheta_\varrho(0) = 1$ (theorem SH.1(iv)), applied to Σ_{out} by functional calculus.

Corollary L.2 (Expected regularized update)**.** *Assume Assumption S3 with $\Sigma_{\mathrm{in}} = \mathrm{Id}_{N_{\mathrm{in}}}$, and assume that $\epsilon_{\mathrm{os}} = 3(1-\beta^2)N_{\mathrm{in}}/B \le \bar{\epsilon}$ along the sequence of dimensions for a fixed $\bar{\epsilon} < 1$. For $\varrho \in [\varrho_-, \varrho_+]$ let ϑ_ϱ be the function of theorem SH.1(iv) at the current dimension. Then*

$$\sup_{\varrho \in [\varrho_-, \varrho_+]} \Big\| \mathbb{E}\, Q_\varrho(M_t) - \sqrt{\frac{B_{\mathrm{eff}}}{N_{\mathrm{in}}}}\, \vartheta_\varrho(\Sigma_{\mathrm{out}})\, \mathfrak{d}^{\mathrm{os}}_{\hat{\nu}_{\mathrm{out}},r}\Big(\Sigma_{\mathrm{out}}; \frac{\varrho}{s_2\gamma_{\mathrm{in}}}\Big) \frac{\Delta_t}{\sqrt{\mathscr{R}(t)}} \Big\|_{\mathrm{F}} \longrightarrow 0, \tag{43}$$

and $\operatorname{spec} \vartheta_\varrho(\Sigma_{\mathrm{out}}) \subset \big[1 - \epsilon_{\mathrm{os}}, (1 - \epsilon_{\mathrm{os}})^{-1}\big]$ *at every dimension and for every ϱ.*

Full statement and proof: corollary SH.1 of the supplement. Lean: proved: `Muon.Oversampled.OsUpdateHyp`, `Muon.Oversampled.osUpdateTarget`, *+2 more.*

This corollary is the precise form of the main-text display $W_{t+1} - W_t \approx -\eta\sqrt{B_{\mathrm{eff}}/(2N)}\, \sqrt{\mathcal{K}}\, (W_t - W^*)/\sqrt{\mathcal{R}(t)}$ of section 4.1. When $\epsilon_{\mathrm{os}} = 3(1 - \beta^2)N_{\mathrm{in}}/B$ stays below a fixed $\bar{\epsilon} < 1$, the expected regularized update is, up to an error of vanishing Frobenius norm, a function of Σ_{out} applied to the normalized error $\Delta_t/\sqrt{\mathscr{R}(t)}$: the factor $\sqrt{B_{\mathrm{eff}}/N_{\mathrm{in}}}$ times the oversampled profile $\mathfrak{d}^{\mathrm{os}}$ evaluated at Σ_{out}, times an operator $\vartheta_\varrho(\Sigma_{\mathrm{out}})$ that is within $\epsilon_{\mathrm{os}}/(1 - \epsilon_{\mathrm{os}})$ of the identity at every dimension. On packed modes and at zero rescaled ridge the profile is within constants of $\sqrt{\mu/2}$ (corollary L.3),

which is the square-root preconditioner of the main text. The statement holds at ridges in a fixed compact subset of $(0,\infty)$, in the proportional regime; it says nothing about the algorithm at zero ridge.

Proof sketch. For an isotropic input side the deterministic equivalent of the expected update is an explicit function of Σ_{out} applied to the error. The drift theorem gives $\mathbb{E}\, Q_\varrho(M_t)$ up to a vanishing Frobenius error as an integral, over the spectral parameters $s = \varrho + u^2$, of a candidate built from the deterministic resolvents (theorem SC.2, appendix G.2). With $\Sigma_{\text{in}} = \mathrm{Id}$ these resolvents are $(s\,\mathrm{Id} + \tilde{s}_1\Sigma_{\text{out}})^{-1}$ and $(s+s_1)^{-1}\,\mathrm{Id}$, so the candidate is the scalar kernel $\mathsf{k}(s;\cdot)$ evaluated at Σ_{out} and applied to $\Delta_t/\sqrt{2\mathscr{R}(t)}$, and the integral is exactly (eq. (SH.13))

$$\overline{\widetilde{G}}_\varrho = \mathfrak{d}_\varrho(\Sigma_{\text{out}})\,\frac{\Delta_t}{\sqrt{\mathscr{R}(t)}}.$$

The oversampled law rewrites $\mathfrak{d}_\varrho(\Sigma_{\text{out}})$ at every finite dimension, with a relative error that depends on nothing but ϵ_{os}. By the sandwich of appendix G.5, applied eigenvalue by eigenvalue (theorem SH.1(iv)),

$$\mathfrak{d}_\varrho(\Sigma_{\text{out}}) = \sqrt{\frac{B_{\text{eff}}}{N_{\text{in}}}}\;\vartheta_\varrho(\Sigma_{\text{out}})\,\mathfrak{d}^{\text{os}}\Big(\Sigma_{\text{out}};\frac{\varrho}{s_2\gamma_{\text{in}}}\Big), \qquad \operatorname{spec}\vartheta_\varrho(\Sigma_{\text{out}}) \subset \big[1-\epsilon_{\text{os}},(1-\epsilon_{\text{os}})^{-1}\big].$$

Inserting this into the previous display gives the corollary. The bound on ϑ_ϱ holds at every dimension, uniformly in the spectrum, the eigenvalue and the ridge, so the dimension limit of the drift theorem is taken with a factor that is already controlled and no exchange of limits occurs. Because the error is relative to the profile, the law is accurate also on modes where $\mathfrak{d}^{\text{os}}(\mu) \ll \sqrt{\mu}$.

The profile sees the ridge only through the rescaled ridge $\hat{\varrho} = \varrho/(s_2\gamma_{\text{in}})$, and this decides when the square-root law is visible. The profile is at most $\min\{\sqrt{\mu/2},\ \mu/\sqrt{2\hat{\varrho}}\}$ (lemma SH.2), so at a fixed normalized ridge and $\gamma_{\text{in}} \downarrow 0$ it becomes linear in μ; the square-root law concerns bounded $\hat{\varrho}$, in particular the zero-ridge coefficient of corollary L.3. ◇

Proposition L.4 (Head modes). *(i) For every probability measure ν, every $\mu > 0$ and every $\delta > 0$,*

$$\frac{\delta}{2(1+\sqrt{\delta})^2}\,\nu\big([\delta\mu,\infty)\big) \le 1-\chi_\nu(\mu) \le \delta + \nu\big([\delta\mu,\infty)\big).$$

(ii) For every sequence of probability measures ν_N and numbers $\mu_N > 0$, one has $\mathfrak{d}^{\text{os}}_{\nu_N}(\mu_N)/\sqrt{\mu_N/2} \to 1$ if and only if $\nu_N([\delta\mu_N,\infty)) \to 0$ for every $\delta > 0$.

Full statement and proof: proposition SH.1 of the supplement. Lean: proved: `Muon.Oversampled.prop_app_os_head_i`, `Muon.Oversampled.prop_app_os_head_ii`.

This proposition supports the statement of section 4.1 that for modes high in the spectrum the constant of the square-root law tends to one. By the representation of appendix G.5, $\mathfrak{d}^{\text{os}}_\nu(\mu)/\sqrt{\mu/2}$ is the efficiency $\chi_\nu(\mu)$, and part (i) bounds its defect by the mass of the spectrum above a fraction δ of the mode. Part (ii) says that the efficiency tends to one exactly for head modes, which lie above all but a vanishing fraction of the spectrum.

Proof sketch. The defect $1-\chi_\nu(\mu)$ is within a factor two of an exact average. The average of the soft count against ϖ has a closed form, by partial fractions in $y = w^2$ (lemma SH.9):

$$\int_0^\infty \varpi(y)\,\frac{x}{\mu y + x}\,\mathrm{d}y = \frac{x}{(\sqrt{\mu}+\sqrt{x})^2}, \quad m^\varpi_\nu(\mu) \overset{\text{def}}{=} \int_0^\infty \varpi(y)\,m_\nu(\mu y)\,\mathrm{d}y = \int \frac{x\,\nu(\mathrm{d}x)}{(\sqrt{\mu}+\sqrt{x})^2}.$$

The elementary bounds $m/2 \le 1-\sqrt{1-m} \le m$ on $[0,1]$, integrated against ϖ, give $\frac{1}{2}m^\varpi_\nu(\mu) \le 1-\chi_\nu(\mu) \le m^\varpi_\nu(\mu)$ (lemma SH.10).

The averaged count is then bounded by splitting the spectrum at $\delta\mu$. Its integrand $x/(\sqrt{\mu}+\sqrt{x})^2$ is nondecreasing in x, smaller than δ below $\delta\mu$, at most one above, and at least $\delta/(1+\sqrt{\delta})^2$ on $[\delta\mu,\infty)$, which gives (i). Part (ii) follows by letting $N \to \infty$ and then $\delta \downarrow 0$ in the upper bound, and from the lower bound at each fixed δ for the converse. ◇

A mode $\mu_i > 0$ is (K,δ)-packed, for $K \ge 1$ and $\delta \in (0,1]$, if $\#\{k \le N : \mu_k \le K\mu_i\} \ge \delta N$, as in definition 4.1 (definition SH.2).

Corollary L.3 (Packed modes)**.** *Assume $N_{\mathrm{out}} = N_{\mathrm{in}} = N$ and $\epsilon_{\mathrm{os}} < 1$, and let $\mathfrak{d}_i(0)$ be the zero-ridge coefficient of theorem SH.1(ii). For every mode with $\mu_i > 0$,*

$$\mathfrak{d}_i(0) \le \frac{1}{1-\epsilon_{\mathrm{os}}} \sqrt{\frac{B_{\mathrm{eff}}}{2N}}\, \sqrt{\mu_i},$$

and if μ_i is (K,δ)-packed, then

$$\mathfrak{d}_i(0) \ge (1-\epsilon_{\mathrm{os}})\, C_* \sqrt{\frac{\delta}{1+K}} \sqrt{\frac{B_{\mathrm{eff}}}{2N}}\, \sqrt{\mu_i}, \qquad C_* = \frac{1}{2} + \frac{2\log 2}{\pi}.$$

Full statement and proof: corollary SH.2 of the supplement. Lean: proved: `Muon.Oversampled.cor_app_os_packed_i`, `Muon.Oversampled.cor_app_os_packed_ii`.

This corollary is theorem 4.1 of the main text, whose constant 0.94 is C_* rounded. When the effective batch exceeds $3(1+\beta)^2 N$, the drift of every mode is at most the square-root law $\sqrt{B_{\mathrm{eff}}/(2N)}\sqrt{\mu_i}$, up to the factor $(1-\epsilon_{\mathrm{os}})^{-1}$, and it is within a constant of that law whenever a fixed fraction δ of the spectrum lies below a fixed multiple $K\mu_i$ of the mode. Both bounds concern the deterministic zero-ridge coefficient.

Proof sketch. The oversampled law reduces the coefficient to the profile. At zero ridge (theorem SH.1(iii), appendix G.5),

$$\mathfrak{d}_i(0) = \vartheta_i \sqrt{\frac{B_{\mathrm{eff}}}{N}}\, \mathfrak{d}^{\mathrm{os}}_{\hat\nu}(\mu_i) = \vartheta_i \sqrt{\frac{B_{\mathrm{eff}}}{N}} \sqrt{\frac{\mu_i}{2}}\, \chi_{\hat\nu}(\mu_i), \qquad \vartheta_i \in \big[1-\epsilon_{\mathrm{os}}, (1-\epsilon_{\mathrm{os}})^{-1}\big],$$

and $\chi \le 1$ gives the upper bound.

The efficiency is at least C_* times the square root of the soft fraction of the spectrum below the mode. Since $1-m_\nu$ is nondecreasing, and since $\mu y/(\mu y + x) \ge y\,\mu/(\mu + x)$ for $y \le 1$,

$$1 - m_\nu(\mu y) \ge 1 - m_\nu(\mu) \quad (y \ge 1), \qquad 1 - m_\nu(\mu y) \ge y\,\big(1 - m_\nu(\mu)\big) \quad (y \le 1).$$

Inserted in the definition of χ_ν, these give $\chi_\nu(\mu) \ge \sqrt{1-m_\nu(\mu)} \cdot \frac{2}{\pi}\big[(\log 2 - \frac{1}{2}) + (\frac{\pi}{4} + \frac{1}{2})\big] = C_*\sqrt{1-m_\nu(\mu)}$, the two brackets being $\int_0^1 y(1+y)^{-2}\mathrm{d}y$ and $\int_1^\infty \sqrt{y}(1+y)^{-2}\mathrm{d}y$ (theorem SH.2(i)). Packing bounds the soft fraction below: $\mu/(\mu+x) \ge 1/(1+K)$ for $x \le K\mu$, so $1 - m_{\hat\nu}(\mu_i) \ge \hat\nu([0, K\mu_i])/(1+K) \ge \delta/(1+K)$ (theorem SH.2(ii)), which gives the lower bound. ◇

Remark L.2 (Where packing holds)**.** *The count $\#\{k : \mu_k \le K\mu\}$ is nondecreasing in μ, so if the smallest positive eigenvalue is (K,δ)-packed, then every mode is. If all eigenvalues lie in $[\mu_{\min}, \kappa\mu_{\min}]$ with $\mu_{\min} > 0$, every mode is $(\kappa, 1)$-packed, since $\mu_k \le \kappa\mu_{\min} \le \kappa\mu_i$ for all k. If the spectrum consists of clusters of equal eigenvalues, each holding at least pN of them, every mode is $(1,p)$-packed, since its own cluster lies in $[0,\mu_i]$. For the geometric spectrum $\mu_k = q^k$ with fixed $q \in (0,1)$, the smallest eigenvalue μ_N has only $1 + \lfloor \log K/\log(1/q) \rfloor$ eigenvalues in $[0, K\mu_N]$, a vanishing fraction as $N \to \infty$, so it is not (K,δ)-packed for any fixed K, δ once N is large. For the power law $\mu_k = c_N k^{-\alpha}$, $1 \le k \le N$, every mode is $(2, 1-2^{-1/\alpha})$-packed, because $\mu_k \le 2\mu_i$ for every $k \ge 2^{-1/\alpha} i$, and there are at least $N - 2^{-1/\alpha} i \ge (1-2^{-1/\alpha})N$ such k. Conversely, packing is necessary for a lower bound of this form: by proposition L.5(vi), along a sequence of spectra the efficiency $\chi_{\nu_N}(\mu_N)$ stays bounded below only if, for some K, the fraction of the spectrum in $[0, K\mu_N]$ stays bounded below, and by proposition L.5(ii),(iii) a mode far below most of the spectrum receives drift of gradient-descent order up to a logarithm.*

Full statement and proof: remark SH.3 of the supplement. Lean: proved: `Muon.Oversampled.rem_app_os_packed_i`, `Muon.Oversampled.rem_app_os_packed_ii`.

This remark supports the examples after definition 4.1 in the main text. Packing is monotone in the mode, so it is enough to check it at the smallest eigenvalue; bounded condition numbers and spectra made of large clusters are packed with explicit constants, and every mode of a power law is $(2, 1-2^{-1/\alpha})$-packed, because at least that fraction of the indices lies beyond $2^{-1/\alpha} i$. A geometric spectrum is not packed at its smallest eigenvalue, since only boundedly many eigenvalues lie within a factor K of it. Packing is also necessary for a lower bound of the form of corollary L.3: by

proposition L.5(vi) the efficiency stays bounded below only if a fixed fraction of the spectrum lies below a fixed multiple $K\mu$ of the mode.

For a probability measure ν, $\ell \geq 0$, $\mu > 0$ and $\Gamma \geq 0$ put

$$\mathcal{I}_\nu(\ell) = \int_{(\ell,\infty)} \frac{\nu(\mathrm{d}x)}{x}, \qquad \mathcal{I}_\nu = \mathcal{I}_\nu(0), \qquad \mathcal{E}(\mu,\Gamma) = \frac{\sqrt{2}}{\pi}\,\mu \int_0^\infty \frac{\min\{\sqrt{c},\, c\sqrt{\Gamma}\}}{(c+\mu)^2}\,\mathrm{d}c,$$

and let $C_* = \frac{1}{2} + \frac{2\log 2}{\pi}$ (lemma SH.12 and theorem SH.2).

Proposition L.5 (Tail modes)**.** *Let ν be a probability measure and $\mu > 0$.*

(i) (Split bound.) For every $\ell \geq 0$ with $\mathcal{I}_\nu(\ell) < \infty$,

$$\mathfrak{d}_\nu^{\mathrm{os}}(\mu) \leq \sqrt{\nu([0,\ell])}\,\sqrt{\frac{\mu}{2}} + \mathcal{E}\big(\mu, \mathcal{I}_\nu(\ell)\big).$$

(ii) (Gradient-descent-like upper bound.) If $\nu(\{0\}) = 0$ and $\mathcal{I}_\nu < \infty$, then

$$\mathfrak{d}_\nu^{\mathrm{os}}(\mu) \leq \frac{\sqrt{2}}{\pi}\sqrt{\mathcal{I}_\nu}\;\mu\Big[\log\Big(1 + \frac{1}{\mu\mathcal{I}_\nu}\Big) + 2\Big].$$

(iii) (Asymptotics.) If $\nu(\{0\}) = 0$ and $\mathcal{I}_\nu < \infty$, then, for fixed ν,

$$\mathfrak{d}_\nu^{\mathrm{os}}(\mu) \sim \frac{\sqrt{2}}{\pi}\sqrt{\mathcal{I}_\nu}\;\mu\log\frac{1}{\mu} \qquad (\mu \downarrow 0).$$

(iv) (Counts.) For every $K > 0$,

$$C_*\sqrt{\frac{\nu([0,K\mu])}{1+K}} \leq \chi_\nu(\mu) \leq \sqrt{\nu([0,K\mu])} + \frac{2}{\pi\sqrt{K}}\big[\log(1+K) + 2\big].$$

(v) (Universal lower bound.) If $\nu([0,X]) = 1$ for some $X > 0$, then

$$\mathfrak{d}_\nu^{\mathrm{os}}(\mu) \geq \frac{1}{\pi}\,\frac{\mu}{\sqrt{X}}\Big[\log\Big(1 + \frac{X}{\mu}\Big) - \frac{X}{X+\mu}\Big] \geq \frac{1}{2\pi}\,\frac{\mu}{\sqrt{X}}\,\log\frac{X}{\mu}.$$

(vi) (Sequences.) For probability measures ν_N and numbers $\mu_N > 0$ put $\chi_N \overset{\mathrm{def}}{=} \chi_{\nu_N}(\mu_N)$. Then $\liminf_N \chi_N > 0$ if and only if $\liminf_N \nu_N([0,K\mu_N]) > 0$ for some $K > 0$; and $\chi_N \to 0$ if and only if $\nu_N([0,K\mu_N]) \to 0$ for every $K > 0$.

Full statement and proof: proposition SH.2 of the supplement. Lean: proved: `Muon.Oversampled.prop_app_os_tail_i`, `Muon.Oversampled.prop_app_os_tail_ii`, *+5 more.*

This proposition supports the last sentence of section 4.1: a mode that is not packed, such as an isolated small eigenvalue, can be as slow as under gradient descent, up to a logarithm. If the spectrum has a finite mean inverse $\mathcal{I}_\nu$ and no atom at zero, the profile of a small mode is of order $\mu\log(1/\mu)$ instead of $\sqrt{\mu}$, by (ii) and (iii); by (v) it is never smaller than that for a bounded spectrum. Part (iv) brackets the efficiency by the fraction of the spectrum in $[0, K\mu]$, and (vi) turns this into the criterion along sequences of spectra.

Proof sketch. All bounds come from the first form of the representation of appendix G.5,

$$\mathfrak{d}_\nu^{\mathrm{os}}(\mu) = \frac{\sqrt{2}}{\pi}\,\mu\int_0^\infty \frac{\sqrt{\Upsilon_\nu(c)}}{(c+\mu)^2}\,\mathrm{d}c, \qquad \Upsilon_\nu(c) = c\big(1 - m_\nu(c)\big) = c^2\int\frac{\nu(\mathrm{d}x)}{c+x},$$

in which the mode enters only through the kernel $\mu(c+\mu)^{-2}$, which concentrates near $c = \mu$. Splitting the spectrum at ℓ and using $c/(c+x) \leq c/x$ above ℓ gives $1 - m_\nu(c) \leq \nu([0,\ell]) + c\,\mathcal{I}_\nu(\ell)$, hence $\sqrt{\Upsilon_\nu(c)} \leq \sqrt{c\,\nu([0,\ell])} + \min\{\sqrt{c}, c\sqrt{\mathcal{I}_\nu(\ell)}\}$, and the two terms give the two terms of (i). At $\ell = 0$ the first term vanishes, and splitting the remaining integral at $c = 1/\mathcal{I}_\nu$, where the minimum changes branch, gives the logarithm of (ii) (lemma SH.12).

The matching lower bounds use that $\Upsilon_\nu(c)/c^2$ is nonincreasing in c. Below a fixed c_1 this gives $\sqrt{\Upsilon_\nu(c)} \ge c\sqrt{\Upsilon_\nu(c_1)}/c_1$, and the integral of $c(c+\mu)^{-2}$ over $(0, c_1)$ is at least $\log(1 + c_1/\mu) - 1$; letting $\mu \downarrow 0$ and then $c_1 \downarrow 0$, so that $\Upsilon_\nu(c_1)/c_1^2 \uparrow \mathcal{I}_\nu$, gives (iii). If the spectrum lies in $[0, X]$, then $\Upsilon_\nu(c) \ge c^2/(2X)$ for $c \le X$, and the elementary inequality $\log(1+k) - k/(1+k) \ge \frac{1}{2}\log k$ gives (v) (lemma SH.13). Part (iv) combines the square-root law of corollary L.3 (theorem SH.2) with (i) at $\ell = K\mu$, where $\mathcal{I}_\nu(K\mu) \le 1/(K\mu)$, and (vi) follows from (iv) by choosing K large. ⋄

For a probability measure ν with $\nu([0, \mu_+]) = 1$ and a mode $0 < \mu \le \mu_+/2$, the acceleration of corollary SH.4 compares the profile with the gradient-descent rate μ/μ_+, and $\varkappa$ is the condition number:

$$\mathcal{A} = \frac{\mathfrak{d}_\nu^{\mathrm{os}}(\mu)/\mathfrak{d}_\nu^{\mathrm{os}}(\mu_+)}{\mu/\mu_+}, \qquad \varkappa = \frac{\mu_+}{\mu}, \qquad C_* = \frac{1}{2} + \frac{2\log 2}{\pi}.$$

Corollary L.4 (Power-law spectra)**.** *Let $\alpha > 0$, $N \ge 1$ and $c_N > 0$, let $\mu_k = c_N k^{-\alpha}$ for $1 \le k \le N$, and put $\nu_N \overset{\mathrm{def}}{=} N^{-1}\sum_{k=1}^N \delta_{\mu_k}$. For $0 < x \le 1$ put*

$$\Xi_\alpha(x) \overset{\mathrm{def}}{=} \int_0^\infty \varpi(y)\Big(\int_0^1 \frac{y\,u^\alpha}{y\,u^\alpha + x^\alpha}\,\mathrm{d}u\Big)^{1/2}\mathrm{d}y.$$

(i) For every $1 \le i \le N$,

$$0 \le \chi_{\nu_N}(\mu_i) - \Xi_\alpha(i/N) \le N^{-1/2}.$$

(ii) If $i \le \varepsilon N$ with $\varepsilon \in (0, 1]$, then $1 - 2\varepsilon^{\alpha/(1+\alpha)} \le \chi_{\nu_N}(\mu_i) \le 1$. Consequently, for every sequence $\varepsilon_N \to 0$,

$$\max_{i \le \varepsilon_N N} \Big|\frac{\mathfrak{d}_{\nu_N}^{\mathrm{os}}(\mu_i)}{\sqrt{\mu_i/2}} - 1\Big| \longrightarrow 0.$$

(iii) The function Ξ_α is nonincreasing on $(0, 1]$, and

$$1 - \Xi_\alpha(x) \le 2x^{\alpha/(1+\alpha)}, \qquad \Xi_\alpha(x) \ge \frac{C_*}{\sqrt{2(1+\alpha)}}.$$

(iv) For every $1 \le i \le N$,

$$\frac{C_*}{\sqrt{2(1+\alpha)}}\sqrt{\frac{\mu_i}{2}} \le \mathfrak{d}_{\nu_N}^{\mathrm{os}}(\mu_i) \le \sqrt{\frac{\mu_i}{2}}.$$

(v) If $N^\alpha \ge 2$, the acceleration of corollary SH.4 for $\nu = \nu_N$, $\mu_+ = \mu_1$ and $\mu = \mu_N$, for which $\varkappa = N^\alpha$, satisfies

$$\frac{C_*}{\sqrt{2(1+\alpha)}}\,N^{\alpha/2} \le \mathcal{A} \le \frac{\sqrt{2}}{C_*}\,N^{\alpha/2}.$$

Full statement and proof: corollary SH.5 of the supplement. Lean: proved: `Muon.Oversampled.plMu`, `Muon.Oversampled.plSpec`, +6 more.

This corollary is the large-batch paragraph of section 4.2. For a power law every mode keeps the square-root law up to the constant $C_*/\sqrt{2(1+\alpha)}$, which depends only on α, by (iv); the modes $i \ll N$ have efficiency tending to one, by (ii), which with the factor $\vartheta_i \to 1$ of the oversampled law as $\epsilon_{\mathrm{os}} \to 0$ gives $\mathfrak{d}_i \sim \sqrt{\mu_i B_{\mathrm{eff}}/(2N)}$ (theorem SH.1). By (v) the slowest mode contracts faster than under gradient descent, relative to the top mode, by a factor of order $N^{\alpha/2}$, the square root of the condition number. The efficiency itself has the limit profile $\Xi_\alpha(i/N)$ with error at most $N^{-1/2}$, by (i).

Proof sketch. After rescaling the spectrum, the soft fraction is a Riemann sum of a monotone function. The efficiency is unchanged by $\mu_k \mapsto a\mu_k$ (lemma SH.14), so one may take $\mu_k = (k/N)^{-\alpha}$. Then

$$1 - m_{\nu_N}(c) = \frac{1}{N}\sum_{k=1}^N \phi_c(k/N), \qquad \phi_c(u) = \frac{c\,u^\alpha}{c\,u^\alpha + 1},$$

and since ϕ_c is nondecreasing with values in $[0,1)$ the sum exceeds $\int_0^1 \phi_c$ by at most $1/N$. With $|\sqrt{a}-\sqrt{b}| \le \sqrt{|a-b|}$ and $\int \varpi = 1$ this gives (i), and $\Xi_\alpha(x)$ is the efficiency of the law of $U^{-\alpha}$, U uniform on $(0,1)$, at the level $x^{-\alpha}$.

The remaining parts are the head and packing bounds applied to this spectrum. The fraction of the spectrum above $\delta\mu_i$ is at most $\varepsilon\delta^{-1/\alpha}$ when $i \le \varepsilon N$, and proposition L.4(i) with $\delta = \varepsilon^{\alpha/(1+\alpha)}$ gives $1-\chi_{\nu_N}(\mu_i) \le 2\varepsilon^{\alpha/(1+\alpha)}$, which is (ii) and the first bound of (iii). For the limit law $1-m(x^{-\alpha}) = \int_0^1 u^\alpha(u^\alpha+x^\alpha)^{-1}\mathrm{d}u \ge 1/(2(1+\alpha))$, and the bound $\chi \ge C_*\sqrt{1-m}$ of theorem SH.2(i) (sketch of corollary L.3) gives the lower bound of (iii) and, with (i), of (iv). Part (v) is corollary SH.4(i), which brackets the acceleration by $\chi_\nu(\mu)\sqrt{\varkappa}$ and $(\sqrt{2}/C_*)\chi_\nu(\mu)\sqrt{\varkappa}$, at the slowest mode, where $\varkappa = N^\alpha$ and $\chi_{\nu_N}(\mu_N)$ lies between $C_*/\sqrt{2(1+\alpha)}$ and one by (iv). ◇

M The phase map and time to accuracy

In this chapter the kernels are fixed numbers. The deterministic row model (definition SI.1) runs the row recursion eq. (14) in continuous time at a constant step η, with weights $\mu_i > 0$, drifts $\mathfrak{d}_i > 0$, volatilities $\mathfrak{v}_i \ge 0$ and initial row energies $\Omega_i(0)$, $i \le N$; its risk is $\mathcal{R} = \frac{1}{2}\sum_i \mu_i\Omega_i$, $F_0 = \mathcal{R}(0)$, and $\mathcal{S}$ is as in eq. (16). The coefficients are modelled on the zero-ridge kernels of the main text and are not proved to be those of the algorithm along the trajectory (remark SI.1), so every statement is about the model. The matched step $2\sqrt{\epsilon}/\mathcal{S}$ makes the floor $\mathcal{R}(\infty)$ equal to ϵ, and $t_a = \inf\{t : \mathcal{R}(t) \le a\}$. The power-law arrays (definition SI.3) have $\mu_i = i^{-\alpha}$, $\Omega_i(0) = i^{-\varsigma}$, $p = \alpha+\varsigma > 1$, $B \ge N$, $\gamma = N/B$, $H_N(a) = \sum_{i\le N} i^{-a}$, $\bar\mu = H_N(\alpha)/N$, and

$$\mathfrak{d}_i^{\text{s-SVD}} = \sqrt{\mu_i B/N}, \quad \mathfrak{v}_i^{\text{s-SVD}} = 1; \qquad \mathfrak{d}_i^{\text{s-SGD}} = \mu_i\sqrt{B/(\pi\bar\mu)}, \quad \mathfrak{v}_i^{\text{s-SGD}} = N\big[1+\tfrac{2}{\pi}\big(\tfrac{\mu_i}{\bar\mu}-1\big)\big],$$

the large-batch MUON kernels without momentum (SIGNSVD; the adopted drift is $\sqrt{2}$ times the main text's $\sqrt{\mu_i B/(2N)}$, and with the main text's value every ratio of hitting times below is halved and no exponent changes, remark SI.1) and the kernels of table 1. For $j \in \{\text{s-SVD}, \text{s-SGD}\}$, $\mathcal{S}_j$ is the noise constant, $T_0^{\text{s-SVD}} = \sqrt{\gamma}$, $T_0^{\text{s-SGD}} = \sqrt{\pi\bar\mu/B}$, $q_{\text{s-SVD}} = \alpha/2$, $q_{\text{s-SGD}} = \alpha$, $r_j = (p-1)/(2q_j)$, and $D_N = \pi H_N(\alpha)/H_N(\alpha/2)$. The chapter solves the model when all drifts are equal, then gives the power-law hitting times and their ratio in windows of dimension and accuracy, and finally at every accuracy.

In the setting of proposition SI.2 the index set is finite and every row has the same drift, $\mathfrak{d}_i = \mathfrak{d} > 0$. The total volatility is $\mathfrak{v}_{\text{tot}} = \sum_i \mu_i\mathfrak{v}_i > 0$, the volatility $\mathfrak{v}$ of eq. (8), and $\mathcal{S} = \mathfrak{v}_{\text{tot}}/(2\mathfrak{d})$.

Theorem M.1 (Matched floors)**.** *In the setting of proposition SI.2, let $\epsilon > 0$, $T > 1$ and $F_0 > T\epsilon$, and take the matched step $\eta = 2\sqrt{\epsilon}/\mathcal{S} = 4\mathfrak{d}\sqrt{\epsilon}/\mathfrak{v}_{\text{tot}}$. Put $X \overset{\text{def}}{=} \sqrt{F_0/\epsilon} > \sqrt{T}$. Then the floor is ϵ and*

$$t_{T\epsilon} = \frac{\mathfrak{v}_{\text{tot}}}{4\mathfrak{d}^2}\Big[X - \sqrt{T} + \log\frac{X-1}{\sqrt{T}-1}\Big]. \tag{44}$$

If $x \overset{\text{def}}{=} 1/X \le \frac{1}{2}$, then

$$\Big|\frac{t_{T\epsilon}}{\frac{\mathfrak{v}_{\text{tot}}}{4\mathfrak{d}^2}\sqrt{F_0/\epsilon}} - 1\Big| \le x\Big[\sqrt{T} + |\log(\sqrt{T}-1)| + \log\frac{1}{x} + 2x\Big]. \tag{45}$$

For two such models with the same F_0, ϵ and T,

$$\frac{t^{(1)}_{T\epsilon}}{t^{(2)}_{T\epsilon}} = \frac{\mathfrak{v}^{(1)}_{\text{tot}}/(\mathfrak{d}^{(1)})^2}{\mathfrak{v}^{(2)}_{\text{tot}}/(\mathfrak{d}^{(2)})^2}.$$

Full statement and proof: theorem SI.4 of the supplement. Lean: proved: `Muon.Phase.thm_app_phase_matched`.

This theorem is the constant-learning-rate part of the main text's statement, after eq. (12), that the MUON speedup is also the ratio of the times to a common floor. With a common drift the model closes on the risk, as the recursion eq. (8) in continuous time, and at the matched step the time to

$T\epsilon$ is explicit. Its leading term is $\sqrt{F_0/\epsilon}$ times $\mathfrak{v}_{\rm tot}/(4\mathfrak{d}^2)$, half the inverse of the rate $2\mathfrak{d}^2/\mathfrak{v}$, so the ratio of the times of two models is the inverse ratio of their rates, exactly and at every accuracy.

Proof sketch. With a common drift the risk solves a scalar equation, which integrates in closed form: summing the row equations with weights $\mu_i/2$ turns the drift term into $-2\eta\mathfrak{d}\sqrt{\mathcal{R}}$, so exactly (proposition SI.2)

$$\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}t} = -2\eta\mathfrak{d}\sqrt{\mathcal{R}} + \frac{\eta^2\mathfrak{v}_{\rm tot}}{2}.$$

In $\mathsf{u} = \sqrt{\mathcal{R}}$ this reads $\mathsf{u}' = -a_\circ(\mathsf{u} - b_\circ)/\mathsf{u}$, with $a_\circ = \eta\mathfrak{d}$ and $b_\circ = \eta\mathcal{S}/2$. The difference $\mathsf{u} - b_\circ$ keeps its sign, so for $\sqrt{F_0} > b_\circ$ the risk decreases to $b_\circ^2$ without reaching it, and separating variables gives (eq. (SI.12))

$$a_\circ t = \sqrt{F_0} - \sqrt{\mathcal{R}(t)} + b_\circ \log\frac{\sqrt{F_0} - b_\circ}{\sqrt{\mathcal{R}(t)} - b_\circ}.$$

At the matched step the closed form evaluated at the level $T\epsilon$ is the time. There $b_\circ = \sqrt{\epsilon}$ and $a_\circ = 4\mathfrak{d}^2\sqrt{\epsilon}/\mathfrak{v}_{\rm tot}$, and since $F_0 > T\epsilon > b_\circ^2$, setting $\sqrt{\mathcal{R}} = \sqrt{T\epsilon}$ gives eq. (44). With $X - 1 = X(1 - x)$ the exact relative error is

$$\frac{t_{T\epsilon}}{\frac{\mathfrak{v}_{\rm tot}}{4\mathfrak{d}^2}X} - 1 = x\Big[-\sqrt{T} + \log\frac{1}{x} + \log(1-x) - \log(\sqrt{T} - 1)\Big],$$

and $|\log(1-x)| \le 2x$ for $x \le \frac{1}{2}$ gives eq. (45). The bracket in eq. (44) depends only on F_0, ϵ and T, so it cancels in a ratio. ◇

Let $\mathcal{N}_B = \mathbb{E}\left[(\sum_{b\le B} x_b^2 y_b^2)^{1/2}\right]$ for independent standard Gaussian x_b, y_b. The isotropic arrays eq. (SI.15) have N rows with $\mu_i = 1$, a common initial energy, and

$$\mathfrak{d}_i^{\text{s-SVD}} = C(\gamma), \quad \sum_i \mathfrak{v}_i^{\text{s-SVD}} = m; \qquad \mathfrak{d}_i^{\text{s-SGD}} = \mathcal{N}_B/\sqrt{\pi}, \quad \sum_i \mathfrak{v}_i^{\text{s-SGD}} = N^2,$$

where $C(\gamma) = C_0(\gamma)$ is the zero-ridge drift constant without momentum of appendix K. Put $\rho(\gamma) = \pi\, C(\gamma)^2\, \gamma \max(1, \gamma)$.

Theorem M.2 (Isotropic comparison)**.** *Let $N, B \ge 1$, $\gamma = N/B$, $m = \min(B, N)$, and take the isotropic arrays eq.* (SI.15) *with a common initial risk $F_0 = \mathcal{R}^{\text{s-SVD}}(0) = \mathcal{R}^{\text{s-SGD}}(0)$. Choose $\eta^\star_{\text{s-SVD}}$ and $\eta^\star_{\text{s-SGD}}$ by eq.* (16)*, so that both models have floor ϵ, and fix $T > 1$ with $F_0 > T\epsilon$.*

(i) For every such ϵ, $t^{\text{s-SGD}}_{T\epsilon}/t^{\text{s-SVD}}_{T\epsilon} = \rho(\gamma)\, B/\mathcal{N}_B^2$. As $\epsilon/F_0 \to 0$, uniformly in N, B and $C(\gamma)$,

$$t^{\text{s-SVD}}_{T\epsilon} \sim \frac{m}{4C(\gamma)^2}\sqrt{\frac{F_0}{\epsilon}}, \qquad t^{\text{s-SGD}}_{T\epsilon} \sim \frac{\pi N^2}{4\mathcal{N}_B^2}\sqrt{\frac{F_0}{\epsilon}}.$$

(ii) If in addition $B \to \infty$, with N and γ allowed to vary, then

$$t^{\text{s-SGD}}_{T\epsilon} \sim \frac{\pi N^2}{4B}\sqrt{\frac{F_0}{\epsilon}}, \qquad \frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}}\Big/\rho(\gamma) \longrightarrow 1,$$

and the ratio tends to $\rho(\gamma_0)$ if $\gamma \to \gamma_0 \in (0, \infty)$.

(iii) For every $\gamma > 0$,

$$\rho(1) \le \rho(\gamma) < \frac{32}{9\pi}, \qquad \rho(1) \le \frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}} \le \frac{\pi^2}{4}\cdot\frac{32}{9\pi},$$

with $\rho(\gamma) = \rho(1)$ only at $\gamma = 1$; and $\rho(\gamma) \to \frac{32}{9\pi}$ as $\gamma \downarrow 0$ and $\rho(\gamma) \to 1$ as $\gamma \to \infty$.

Full statement and proof: theorem SI.6 of the supplement. Lean: proved: `Muon.Phase.thm_app_phase_iso`.

At $\beta = 0$ this is the precise form of the statement, after eq. (12), that the MUON speedup is the ratio of the times to a common floor. The identity in (i) is exact: the ratio is $\rho(\gamma)$ times $B/\mathcal{N}_B^2 \in [1, \pi^2/4]$, which tends to one at large batch. By (iii), ρ has its minimum $\rho(1) = \frac{1}{4}(1 + \log(1+\sqrt{2})/\sqrt{2})^2 \approx$

0.6587 at the square aspect ratio, where SIGNSGD is faster, tends to 1 at small batch ($\gamma \to \infty$), and has supremum $32/(9\pi) \approx 1.13$ at large batch ($\gamma \downarrow 0$), the values of eq. (12) at $\beta = 0$. With momentum the same algebra applies to the pair $(C_\beta(\gamma), \mathfrak{v}_{\mathrm{iso},\beta})$.

Proof sketch. Each isotropic array has a common drift, so by theorem M.1 and $N^2/(mB) = \gamma \max(1,\gamma)$ (proposition SI.3),

$$\frac{t_{T\epsilon}^{\text{s-SGD}}}{t_{T\epsilon}^{\text{s-SVD}}} = \frac{N^2/(\mathcal{N}_B^2/\pi)}{m/C(\gamma)^2} = \rho(\gamma)\,\frac{B}{\mathcal{N}_B^2}.$$

The Gaussian bounds $\max\{\frac{2}{\pi}\sqrt{B}, \sqrt{B}-4/\sqrt{B}\} \le \mathcal{N}_B \le \sqrt{B}$ (proposition N.1) give $1 \le B/\mathcal{N}_B^2 \le \pi^2/4$, and $B/\mathcal{N}_B^2 \le (1-4/B)^{-2}$ for $B > 4$. The relative error in theorem M.1 depends only on ϵ/F_0 and T, which gives (i) and (ii).

The shape of ρ comes from an integral for $C(\gamma)$ over the participation level. Let $\Phi(v) = \mathbb{E}[vZ^2/(1+vZ^2)]$ with Z standard Gaussian, and V its inverse. The square-case scalar equation of the isotropic analysis gives the spectral parameter as a function of the level y, $t = (\gamma - y)^2/V(y)$. Changing variables from t to y in the half moment that defines $C(\gamma)$ and integrating by parts gives (lemma SI.4)

$$C(\gamma) = \frac{\sqrt{2}\,I(\gamma)}{\pi\gamma^2}, \qquad I(\gamma) = \int_0^{\min(\gamma,1)} (\gamma - y)\,w(y)\,\mathrm{d}y, \qquad w = V^{-1/2},$$

where $w_\circ(y) = \sqrt{y}\,w(y)$ decreases strictly from 1. For $\gamma \ge 1$, $I(\gamma) = \gamma\mathcal{A} - \mathcal{D}$ with $\mathcal{A} = \int_0^1 w = \sqrt{\pi/2}$ and $\mathcal{D} = \int_0^1 yw$, so $\rho(\gamma) = (1 - \mathcal{D}/(\mathcal{A}\gamma))^2$ increases to 1. For $\gamma < 1$, $\rho = \frac{2}{\pi}\gamma^{-3}I^2$, whose derivative has the sign of $2\gamma I' - 3I$; this is negative, because the constant $w_\circ(\gamma/3)$ integrates to zero against $(3y-\gamma)y^{-1/2}$ and $w_\circ$ is decreasing:

$$2\gamma I'(\gamma) - 3I(\gamma) = \int_0^\gamma (3y-\gamma)\,y^{-1/2}\big(w_\circ(y) - w_\circ(\gamma/3)\big)\,\mathrm{d}y < 0.$$

As $\gamma \downarrow 0$, $\gamma^{-3/2}I(\gamma) \to \int_0^1 (1-y)y^{-1/2}\mathrm{d}y = \frac{4}{3}$, so $\rho \to \frac{32}{9\pi}$ (theorem SI.5); the closed form of $C(1)$ (theorem SF.6) gives $\rho(1)$. ◇

The low-exponent window eq. (SI.33) is the set of (N, B, ϵ) with $N \ge 4$, $B \ge N$, $0 < \epsilon < \epsilon_0$ and, for both j,

$$k_{N,j}\,C_j\,I_{r_j}\big(H_j(\epsilon)\big) \le \sqrt{\epsilon}, \qquad \text{and} \qquad H_j(\epsilon) \le (N/4)^{q_j} \ \text{ if } r_j \le 1,$$

where ϵ_0 is a fixed constant depending only on α, ς and T, C_j is the constant of the upper power-law bound on the noiseless risk of model j (lemma SI.7), depending only on α and ς, $H_j(\epsilon) = (C_j/\sqrt{\epsilon})^{1/r_j} - 1$, $I_r(x) = \int_0^x (1+w)^{-r}\,\mathrm{d}w$, and $k_{N,\text{s-SVD}} = 2H_N(\alpha)/H_N(\alpha/2)$ and $k_{N,\text{s-SGD}} = 2(1-\frac{2}{\pi})H_N(\alpha)/N + 4H_N(2\alpha)/(\pi H_N(\alpha))$ are the heights of the rescaled noise kernels (eq. (SI.29)).

Theorem M.3 (Time to $T\epsilon$ in the power-law phases)**.** *Fix $\alpha > 0$ and ς with $\alpha + \varsigma > 1$ and $\varsigma \notin \{1, \alpha+1\}$, and fix $T \ge 4$. Consider the arrays of definition SI.3 at their matched steps along a sequence of (N, B, ϵ) with $N \to \infty$, $\epsilon \downarrow 0$ and integers $B \ge N$, such that*

- *$\epsilon N^{\alpha+\varsigma-1} \to \infty$ if $\varsigma < 1$;*
- *$\epsilon N \to \infty$ if $\varsigma > 1$ and $\alpha > 1$;*
- *(N, B, ϵ) lies in the low-exponent window eq.* (SI.33) *for all large indices if $\varsigma > 1$ and $\alpha \le 1$.*

Then there are constants $0 < c_- \le c_+$, depending only on α, ς and T, such that for all large indices of the sequence, with $x \asymp y$ meaning $c_- \le x/y \le c_+$,

$$t_{T\epsilon}^{\text{s-SVD}} \asymp \gamma\,H_N(\alpha/2)\,\epsilon^{-\alpha/[2(\alpha+\varsigma-1)]} \ (\varsigma < 1), \quad t_{T\epsilon}^{\text{s-SVD}} \asymp \sqrt{\gamma}\,\mathcal{S}_{\text{s-SVD}}\,\epsilon^{-1/2} \ (\varsigma > 1), \tag{46}$$

$$t_{T\epsilon}^{\text{s-SGD}} \asymp \frac{N^2\pi\bar{\mu}}{B}\,\epsilon^{-\alpha/(\alpha+\varsigma-1)} \ (\varsigma < \alpha+1), \qquad t_{T\epsilon}^{\text{s-SGD}} \asymp \sqrt{\frac{\pi\bar{\mu}}{B}}\,\mathcal{S}_{\text{s-SGD}}\,\epsilon^{-1/2} \ (\varsigma > \alpha+1). \tag{47}$$

Moreover $H_N(\alpha/2) \asymp N^{1-\alpha/2}$ for $\alpha < 2$, $H_N(\alpha/2) \asymp \log N$ for $\alpha = 2$, and $H_N(\alpha/2) \asymp 1$ for $\alpha > 2$, with constants depending only on α.

Full statement and proof: theorem SI.7 of the supplement. Lean: proved: `Muon.Phase.thm_app_phase_map`, `Muon.Phase.thm_app_phase_map_moreover`.

This theorem supplies the exponents of theorem 5.1: the exponent of $1/\epsilon$ in each display is e_{Muon} or $e_{\text{s-SGD}}$ of eq. (18). Like every statement of this chapter, it concerns the model with the adopted arrays at the matched step, not the algorithm. Because the MUON drift scales like $\sqrt{\mu_i}$ and the SIGNSGD drift like μ_i, the time saturates at $\epsilon^{-1/2}$ beyond $\varsigma = 1$ for SIGNSVD but only beyond $\varsigma = \alpha + 1$ for SIGNSGD. The prefactors carry all the dependence on N and B, and their ratio is D_N of eq. (17). The limit is joint in N and ϵ. For $\varsigma < 1$ the window keeps ϵ above the accuracy scale ϵ_N. At fixed N the exponent is $\frac{1}{2}$ in every phase (corollary SI.2), and theorem M.4 treats $\epsilon/\epsilon_N \to 0$.

Proof sketch. The risk of the model solves a scalar equation once time is measured in the right units. In the row equations the drift of every row is divided by $\sqrt{\mathcal{R}}$, so the natural time is the accumulated effective step $s = \phi(t) = \int_0^t \eta\, \mathcal{R}(t')^{-1/2}\, \mathrm{d}t'$ (eq. (SI.8)). In this time each row solves a linear equation driven by $\sqrt{Q}$, and summing the rows with weights $\mu_i/2$ gives exactly

$$Q(s) = F(s) + \int_0^s k(s-v)\sqrt{Q(v)}\,\mathrm{d}v,$$
$$F(s) = \tfrac{1}{2}\sum_i \mu_i \mathcal{Q}_i(0) e^{-2\mathfrak{d}_i s}, \qquad k(s) = \tfrac{\eta}{2}\sum_i \mu_i \mathfrak{v}_i e^{-2\mathfrak{d}_i s},$$

where $Q(s)$ is the risk at time s (theorems SI.1 and SI.2). The forcing F is the risk the model would have without noise, and the kernel k feeds the volatility back in. The equation has a unique solution, and its limit is $\eta^2\mathcal{S}^2/4$, which equals ϵ at the matched step.

After a change of time scale, the time to $T\epsilon$ factors into a prefactor that carries N and B and a dimensionless integral. The time scale T_0^j of definition SI.3 makes the rescaled decay rates $2\mathfrak{d}_i T_0^j$ equal to $2i^{-q_j}$. The rescaled risk $Y_j(u) = Q_j(T_0^j u)$ then solves (proposition SI.4)

$$Y_j = f_{N,q_j} + \sqrt{\epsilon}\, g_j * \sqrt{Y_j}, \qquad f_{N,q}(u) = \frac{1}{2}\sum_{i=1}^{N} i^{-(\alpha+\varsigma)} e^{-2ui^{-q}},$$

where $f_{N,q}$ is the rescaled noiseless risk and g_j is a kernel of total mass one. Converting back to step counts, the first time $\hat{u}_j$ at which Y_j reaches $T\epsilon$ gives exactly (eqs. (SI.26) and (SI.30))

$$t^j_{T\epsilon} = \frac{\mathcal{S}_j T_0^j}{2\sqrt{\epsilon}} \int_0^{\hat{u}_j} \sqrt{Y_j(u)}\,\mathrm{d}u, \qquad \mathcal{S}_{\text{s-SVD}} T_0^{\text{s-SVD}} = \frac{\gamma}{2} H_N(\alpha/2), \qquad \mathcal{S}_{\text{s-SGD}} T_0^{\text{s-SGD}} = \frac{\pi N^2 \bar{\mu}}{2B}.$$

The phase is decided by whether the square root of the noiseless risk is integrable. The forcing is squeezed between two power laws (lemma SI.7), the lower one valid for $N \geq 4$ and $u \leq (N/4)^q$:

$$A_p\,(1+u)^{-2r_q} \leq f_{N,q}(u) \leq C^2_{p,q}\,(1+u)^{-2r_q}, \qquad r_q = \frac{p-1}{2q}.$$

Thus $r_j = r_{q_j}$. If $r_j > 1$, then $\sqrt{f_{N,q_j}}$ is integrable uniformly in N (eq. (SI.43)), and the integral in the time formula is of order one. If $r_j < 1$, the integral keeps growing until the first passage. At that time $f_{N,q}(\hat{u}) \leq T\epsilon$. The lower power law then bounds the first passage from below,

$$1 + \hat{u} \geq \left(\frac{A_p}{T\epsilon}\right)^{1/(2r)},$$

and the integral is of order $\epsilon^{-(1-r)/(2r)}$ (lemma SI.10). With the prefactor this gives

$$t^j_{T\epsilon} \asymp \mathcal{S}_j T_0^j\, \epsilon^{-1/(2r_j)} \ \ (r_j < 1), \qquad t^j_{T\epsilon} \asymp \mathcal{S}_j T_0^j\, \epsilon^{-1/2} \ \ (r_j > 1),$$
$$\frac{1}{2r_{\text{s-SVD}}} = \frac{\alpha}{2(\alpha+\varsigma-1)}, \quad \frac{1}{2r_{\text{s-SGD}}} = \frac{\alpha}{\alpha+\varsigma-1}.$$

The boundary $r_j = 1$ is the line $\varsigma = 1$ for SIGNSVD and $\varsigma = \alpha + 1$ for SIGNSGD, because $q_{\text{s-SVD}}$ is half of $q_{\text{s-SGD}}$.

The noise term is controlled by comparison with a function of the form $(h+\sqrt{\epsilon})^2$, and this comparison is where the windows come from. Any function that dominates the right side of the rescaled equation

when inserted into it lies above Y_j (lemma SI.3). Take for h a power-law envelope of $\sqrt{f_{N,q_j}}$ that decreases to $\sqrt{\epsilon}$ at a horizon H. Then $(h+\sqrt{\epsilon})^2$ dominates as soon as the kernel does not amplify the envelope, that is, $g_j * h \le 2h$ or $g_j * h \le \sqrt{\epsilon}$ on $[0, H]$ (lemma SI.9). This places the first passage before H, which supplies the upper bound in the previous paragraph. It also bounds the noise part of the integral by a fixed multiple of $\epsilon^{(r_0-1)/(2r_0)}$, where $r_0 \le r_j$ is the exponent of the envelope and $r_0 > 1$ exactly when $r_j > 1$ (eq. (SI.51)). For $r_j < 1$ this bound is of the order of the main term, and the theorems give orders and not equivalents for these models; for $r_j > 1$ it tends to zero.

The kernel condition holds in each range of exponents, and checking it is what fixes the windows. For $\varsigma < 1$ (theorem SI.8) the SIGNSGD kernel is decreasing with mass one, so $g_j * h \le h/(1-r_j)$ (lemma SI.13), which is at most $2h$ because here $r_{\text{s-SGD}} = r_{\text{s-SVD}}/2 < \frac{1}{2}$. The SIGNSVD kernel, normalized by $H_N(\alpha/2)$, has small mass up to times of order $N^{\alpha/2}$ when $\alpha < 2$ (lemma SI.15), a logarithmic substitute at $\alpha = 2$ (lemma SI.17) and an integrable tail when $\alpha > 2$ (lemma SI.14). The horizon is then $H_j \asymp \epsilon^{-1/(2r_j)}$, and since $2r_jq_j = \alpha + \varsigma - 1$ the window $\epsilon N^{\alpha+\varsigma-1} \ge E_*$, with E_* a fixed threshold, keeps H_j below $(N/4)^{q_j}$, where the lower power law holds. For $\alpha > 1$ and $\varsigma \ge 1$ (theorem SI.9) the kernels have integrable power tails (lemma SI.14), except the SIGNSVD kernel for $1 < \alpha \le 2$. For $1 < \alpha < 2$ the normalization $H_N(\alpha/2) \ge N^{1-\alpha/2}$ makes the late part of the convolution small up to the horizon $\sqrt{N}$, which the window $\epsilon N \ge E_*$ guarantees (lemma SI.16). At $\alpha = 2$, lemma SI.18 adds a correction term to the comparison function, needs no relation between N and ϵ, and bounds the noise part by a fixed constant that tends to zero as $N \to \infty$ and $\epsilon \downarrow 0$. For $\alpha \le 1$ and $\varsigma \ge 1$ (theorem SI.10) no tail is usable, but the kernels are decreasing with maxima $k_{N,j} = g_j(0) \to 0$ (proposition SI.5), so

$$(g_j * h_j)(u) \le k_{N,j}\, C_j\, I_{r_j}(H_j) \le \sqrt{\epsilon} \qquad (0 \le u \le H_j),$$

where $C_j = C_{p,q_j}$ and the second inequality is the low-exponent window eq. (SI.33). Elementary harmonic-sum bounds give the orders of $H_N(\alpha/2)$ in the last assertion (lemma SI.8). ◇

The windows of theorems SI.8 to SI.10 are the sets of (N, B, ϵ) with $B \ge N$ and $0 < \epsilon < \epsilon_0$ in which, respectively, $N \ge N_*$ and $\epsilon N^{p-1} \ge E_*$ ($\varsigma < 1$); $N \ge N_*$ and $\epsilon N \ge E_*$ ($\alpha > 1$, $\varsigma \ge 1$); and (N, B, ϵ) lies in the low-exponent window ($\alpha \le 1$, $\varsigma \ge 1$). Here N_*, E_* and ϵ_0 are fixed constants depending only on α, ς and T. With $f_{\infty,q}(u) = \frac{1}{2}\sum_{i\ge1} i^{-p} e^{-2ui^{-q}}$, put $J(p,q) = \int_0^\infty \sqrt{f_{\infty,q}(u)}\,\mathrm{d}u$, which is finite when $r_q = (p-1)/(2q) > 1$.

Corollary M.1 (Ratio of the hitting times)**.** *In the windows of theorem SI.8 ($\varsigma < 1$), theorem SI.9 ($\alpha > 1$, $\varsigma \ge 1$) and theorem SI.10 ($\alpha \le 1$, $\varsigma \ge 1$), with constants depending only on α, ς and T,*

$$\frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}} \asymp D_N \cdot \begin{cases} \epsilon^{-\alpha/[2(p-1)]}, & \varsigma < 1, \\ \epsilon^{-1/2}/\log(1/\epsilon), & \varsigma = 1, \\ \epsilon^{-(\alpha+1-\varsigma)/[2(p-1)]}, & 1 < \varsigma < \alpha+1, \\ \log(1/\epsilon), & \varsigma = \alpha+1, \\ 1, & \varsigma > \alpha+1. \end{cases} \tag{48}$$

Along every sequence with $N \to \infty$ and $\epsilon \downarrow 0$ in the window of theorem SI.9 or theorem SI.10,

$$\frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}} \sim \frac{D_N \log(1/\epsilon)}{2\sqrt{8\alpha}\, J(p,\alpha/2)} \quad (\varsigma = \alpha+1), \qquad \frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}} \sim D_N \frac{J(p,\alpha)}{J(p,\alpha/2)} \quad (\varsigma > \alpha+1).$$

Full statement and proof: corollary SI.3 of the supplement. Lean: proved: `Muon.Phase.cor_app_phase_ratio_i`, `Muon.Phase.cor_app_phase_ratio_ii`, *+3 more.*

Dividing the two hitting-time orders gives the ratio inside the windows, which is the windowed form of eq. (17). Off the lines $\varsigma = 1$ and $\varsigma = \alpha + 1$ the exponent of $1/\epsilon$ is κ of eq. (18), and on those lines a logarithm appears. The prefactor is D_N exactly. In the saturated range $\varsigma > \alpha + 1$ the second display gives the constant $D_N J(p,\alpha)/J(p,\alpha/2)$, a quotient of the clock integrals of the two noiseless risks. For $\alpha > 2$, $D_N \to \pi\zeta(\alpha)/\zeta(\alpha/2)$, so the ratio has a finite positive limit, which is Phase E of fig. 7.

Proof sketch. In each window both models have the order of theorems SI.8 to SI.10 at once. By appendix G.6, each time is $\mathcal{S}_j T_0^j/(2\sqrt{\epsilon})$ times a clock integral of order $\epsilon^{-(1-r_j)/(2r_j)}$, $\log(1/\epsilon)$ or 1 as $r_j < 1$, $r_j = 1$, $r_j > 1$, so

$$t^j_{T\epsilon} \asymp \mathcal{S}_j T_0^j\, \Pi_{r_j}(\epsilon), \qquad \Pi_r(\epsilon) = \epsilon^{-1/(2r)},\ \ \epsilon^{-1/2}\log\tfrac{1}{\epsilon},\ \ \epsilon^{-1/2} \quad (r<1,\ r=1,\ r>1).$$

The prefactors divide to D_N (eq. (SI.27)); the rest is arithmetic with $r_{\text{s-SVD}} = (p-1)/\alpha$, $r_{\text{s-SGD}} = (p-1)/(2\alpha)$. For example, for $1 < \varsigma < \alpha + 1$ only SIGNSGD has $r_j < 1$, and the quotient is $\epsilon^{-\alpha/(p-1)+1/2} = \epsilon^{-(\alpha+1-\varsigma)/(2(p-1))}$; at $\varsigma = 1$, where $p - 1 = \alpha$, it is $\epsilon^{-1}/(\epsilon^{-1/2}\log(1/\epsilon))$.

In the windows of theorems SI.9 and SI.10 the critical and saturated times have equivalents, and their quotient gives the constants. For $r_j > 1$ the noise part of the clock integral, the cost $O(N^{-q_j(r_j-1)})$ of truncating the forcing at N, and the part of $J(p, q_j)$ beyond the first passage all vanish, so the clock integral tends to $J(p, q_j)$. For $r_j = 1$, $\sqrt{f_{\infty,q}(u)} = a_q/u + O(u^{-1-1/q})$ with $a_q = (8q)^{-1/2}$, and the first passage satisfies $\log(1 + \hat{u}) \geq \frac{1}{2}\log(1/\epsilon) - O(1)$ while the comparison horizon bounds it from above, so the clock integral is $\frac{a_q}{2}\log(1/\epsilon)$ up to $o(\log(1/\epsilon))$ (lemma SI.10). This gives eq. (SI.37). At $\varsigma = \alpha + 1$, SIGNSVD is saturated ($r_{\text{s-SVD}} = 2$) and SIGNSGD is critical, and

$$\frac{\mathcal{S}_{\text{s-SGD}} T_0^{\text{s-SGD}}\, \epsilon^{-1/2}\log(1/\epsilon)/(4\sqrt{8\alpha})}{\mathcal{S}_{\text{s-SVD}} T_0^{\text{s-SVD}}\, J(p, \alpha/2)/(2\sqrt{\epsilon})} = \frac{D_N \log(1/\epsilon)}{2\sqrt{8\alpha}\, J(p, \alpha/2)}.$$

For $\varsigma > \alpha + 1$ both are saturated, and the quotient is $D_N J(p, \alpha)/J(p, \alpha/2)$. ◇

Proposition M.1 (Limits of the ratio)**.** *Write* $\mathsf{R} \overset{\text{def}}{=} t^{\text{s-SGD}}_{T\epsilon}/t^{\text{s-SVD}}_{T\epsilon}$*, and consider sequences with* $N \to \infty$*,* $\epsilon \downarrow 0$ *and* $B_N \geq N$ *in the indicated windows. The constants in* $\asymp$ *and* O *depend only on* α*,* ς *and* T*.*

(i) The range $\varsigma < 1$ *(window of theorem SI.8). With* $\Xi_N \overset{\text{def}}{=} \epsilon N^{p-1}$ *and* $r_{\text{s-SVD}} = (p-1)/\alpha$*,*

$$\mathsf{R} \asymp \begin{cases} \Xi_N^{-1/(2r_{\text{s-SVD}})}, & 0 < \alpha < 1, \\ (\log N)\, \Xi_N^{-1/(2r_{\text{s-SVD}})}, & \alpha = 1, \\ N^{\alpha-1}\, \Xi_N^{-1/(2r_{\text{s-SVD}})}, & 1 < \alpha < 2, \\ (N/\log N)\, \Xi_N^{-1/(2r_{\text{s-SVD}})}, & \alpha = 2, \\ \epsilon^{-1/(2r_{\text{s-SVD}})}, & \alpha > 2. \end{cases}$$

Hence $\mathsf{R} \to 0$ *for* $0 < \alpha < 1$ *along sequences with* $\Xi_N \to \infty$*, and* $\mathsf{R} \to \infty$ *for* $\alpha > 2$*. For* $\alpha = 1$ *and* $\epsilon_N = N^{-r_{\text{s-SVD}}}(\log N)^b$ *with* $b > 0$*,* $\mathsf{R} \to \infty$ *if* $b < 2r_{\text{s-SVD}}$ *and* $\mathsf{R} \to 0$ *if* $b > 2r_{\text{s-SVD}}$*. For* $1 < \alpha < 2$ *and* $\epsilon_N = N^{-\omega}$ *with* $0 < \omega < p-1$*, put* $\omega_c \overset{\text{def}}{=} r_{\text{s-SVD}}(2-\alpha) \in (0, p-1)$*; then* $\mathsf{R} \to 0$ *if* $\omega < \omega_c$ *and* $\mathsf{R} \to \infty$ *if* $\omega > \omega_c$*. For* $\alpha = 2$ *and* $\epsilon_N = (\log N)^{-\omega}$ *with* $\omega > 0$*,* $\mathsf{R} \to 0$ *if* $\omega < 2r_{\text{s-SVD}}$ *and* $\mathsf{R} \to \infty$ *if* $\omega > 2r_{\text{s-SVD}}$*. All these sequences lie in the window.*

(ii) $\alpha > 2$*,* $\varsigma \geq 1$ *(window of theorem SI.9).* $\mathsf{R} \to \infty$ *if* $\varsigma \leq \alpha + 1$*, and* $\mathsf{R} \to \pi\zeta(\alpha)J(p,\alpha)/(\zeta(\alpha/2)J(p,\alpha/2)) \in (0,\infty)$ *if* $\varsigma > \alpha + 1$*; this limit exceeds* $\pi/2$ *if* $\alpha \geq 4$*.*

(iii) $1 < \alpha \leq 2$*,* $\varsigma \geq 1$ *(window of theorem SI.9).* $\mathsf{R} \to 0$ *if* $\varsigma > \alpha + 1$*. If* $\varsigma = \alpha + 1$ *and* $1 < \alpha < 2$*, then* $\mathsf{R} \to 0$*. If* $\alpha = 2$*,* $\varsigma = 3$ *and* $\log(1/\epsilon)/\log N \to c_\infty$*, then* $0 \leq c_\infty \leq 1$ *and* $\mathsf{R} \to \pi\zeta(2)c_\infty/(8J(5,1))$*.*

(iv) $0 < \alpha \leq 1$*,* $\varsigma \geq 1$ *(low-exponent window).* $\mathsf{R} \to 0$*; more precisely* R *is* $O(\log(1/\epsilon)^{-2})$*,* $O(\epsilon^{(\varsigma-1)/(p-1)})$*,* $O(\sqrt{\epsilon}\log(1/\epsilon))$ *and* $O(\sqrt{\epsilon})$ *for* $\varsigma = 1$*,* $1 < \varsigma < \alpha + 1$*,* $\varsigma = \alpha + 1$ *and* $\varsigma > \alpha + 1$ *respectively.*

(v) $1 < \alpha \leq 2$*,* $1 \leq \varsigma < \alpha + 1$ *(window of theorem SI.9). Put*

$$e_\varsigma \overset{\text{def}}{=} \frac{\alpha + 1 - \varsigma}{2(p-1)}, \qquad \varsigma_c(\alpha) \overset{\text{def}}{=} \frac{\alpha^2 - 2\alpha + 3}{3 - \alpha},$$

so that $1 < \varsigma_c(\alpha) < \alpha + 1$ *for* $1 < \alpha < 2$*, and* $\varsigma_c(2) = 3$*.*

(a) If $1 < \alpha < 2$ *and* $\varsigma_c(\alpha) \leq \varsigma < \alpha + 1$*, then* $\mathsf{R} \to 0$ *along every sequence with* $N \to \infty$*,* $\epsilon \downarrow 0$ *and* $\epsilon N \to \infty$*.*

(b) If $1 < \alpha < 2$ *and* $1 \leq \varsigma < \varsigma_c(\alpha)$*, put* $\omega_c \overset{\text{def}}{=} (2-\alpha)(p-1)/(\alpha + 1 - \varsigma) \in (0,1)$*. For* $\epsilon_N = N^{-\omega}$ *with* $0 < \omega < 1$*,* $\mathsf{R} \to 0$ *if* $\omega < \omega_c$ *and* $\mathsf{R} \to \infty$ *if* $\omega > \omega_c$*. For* $\epsilon_N = 1/\log N$*,* $\mathsf{R} \to 0$*.*

(c) *If* $\alpha = 2$ *and* $1 \le \varsigma < 3$, *then* $\mathsf{R} \to \infty$ *for* $\epsilon_N = N^{-\omega}$ *with* $0 < \omega < 1$, *and* $\mathsf{R} \to 0$ *for* $\epsilon_N = 1/\log N$.

All the sequences named in (b) and (c) lie in the window for large N.

Full statement and proof: proposition SI.6 of the supplement. Lean: proved: `Muon.Phase.prop_app_phase_winners_i`, `Muon.Phase.prop_app_phase_winners_i_zero`, *+13 more.*

The ratio law of corollary M.1 is a competition between the dimension factor D_N, which tends to zero for $\alpha \le 2$, and the accuracy factor, which grows as $\epsilon \downarrow 0$. This proposition decides it inside the windows: SIGNSVD wins for $\alpha > 2$ and $\varsigma \le \alpha + 1$, the row of Phase A in fig. 7, and SIGNSGD wins for $\alpha \le 1$ and $\varsigma \ge 1$. Elsewhere the winner depends on how N and ϵ move together, through a critical exponent ω_c and the curve $\varsigma_c(\alpha)$; these are windowed forms of the accuracy-exponent map of corollary M.3.

Proof sketch. The order of the dimension factor decides each case. Comparing the harmonic sums with integrals (lemma SI.8), D_N is of order $N^{-\alpha/2}$, $N^{-1/2}\log N$, $N^{\alpha/2-1}$, $1/\log N$ for $\alpha < 1$, $\alpha = 1$, $1 < \alpha < 2$, $\alpha = 2$, and tends to $\pi\zeta(\alpha)/\zeta(\alpha/2)$ for $\alpha > 2$ (eq. (SI.31)). For $\alpha > 2$ the accuracy factor tends to infinity unless $\varsigma > \alpha + 1$, which gives (ii); for $\alpha \ge 4$ the limit exceeds $\pi/2$ because $\zeta(\alpha) > 1$, $\zeta(\alpha/2) < 2$, and $J(p,\alpha) \ge J(p,\alpha/2)$ term by term. In (iii), $D_N \to 0$ and $\log(1/\epsilon) \le \log N$ in the window. For $\varsigma < 1$, write $\epsilon = \Xi_N N^{-(p-1)}$; since $(p-1)/(2r_{\text{s-SVD}}) = \alpha/2$,

$$\mathsf{R} \asymp D_N\, \epsilon^{-1/(2r_{\text{s-SVD}})} = D_N\, N^{\alpha/2}\, \Xi_N^{-1/(2r_{\text{s-SVD}})},$$

and the five rows of (i) are the five orders of $D_N N^{\alpha/2}$; for $1 < \alpha < 2$ and $\epsilon_N = N^{-\omega}$ this is $N^{(\omega-\omega_c)/(2r_{\text{s-SVD}})}$.

For $1 < \alpha \le 2$ and $1 \le \varsigma < \alpha + 1$ the sign of one exponent locates the curve ς_c. These rows of the ratio law read $\mathsf{R} \asymp D_N \epsilon^{-e_\varsigma} \ell_\epsilon$, with $\ell_\epsilon = 1/\log(1/\epsilon)$ at $\varsigma = 1$ and $\ell_\epsilon = 1$ otherwise, and (eq. (SI.58))

$$\frac{\alpha}{2} - 1 + e_\varsigma = \frac{(3-\alpha)\big(\varsigma_c(\alpha) - \varsigma\big)}{2(p-1)}.$$

For $1 < \alpha < 2$, $\mathsf{R} \asymp (\epsilon N)^{-e_\varsigma} N^{\alpha/2-1+e_\varsigma}$, which tends to zero when $\varsigma \ge \varsigma_c(\alpha)$ because $\epsilon N \to \infty$; this is (v)(a). Below the curve, at $\epsilon_N = N^{-\omega}$, the exponent $\frac{\alpha}{2} - 1 + \omega e_\varsigma$ changes sign at $\omega_c = (1 - \frac{\alpha}{2})/e_\varsigma < 1$, which is (v)(b); (v)(c) is the same with $D_N \asymp 1/\log N$. For $\alpha \le 1$ and $\varsigma \ge 1$ the window itself decides: $D_N = \frac{\pi}{2} k_{N,\text{s-SVD}}$, and the window condition $k_{N,\text{s-SVD}} C_{\text{s-SVD}} I_r(H) \le \sqrt{\epsilon}$ makes this $O(\sqrt{\epsilon})$, or $O(\sqrt{\epsilon}/\log(1/\epsilon))$ at $\varsigma = 1$, which outweighs the accuracy factor and gives (iv). ◇

Remark M.1 (The adopted SIGNSGD array and the corrected modes)**.** *The* SIGNSGD *section computes the raw annealed row volatility of the actual sign update,* $\mathfrak{v}_i^{\text{corr}}$, *for a Haar eigenbasis independent of the batch, comparable dimensions and a covariance bounded below (theorem N.1). At* $N_{\text{out}} = N_{\text{in}} = N$, *with* $\Sigma_{\text{out}} = U\Lambda U^\top$, *the state* $\Delta = UD$ *and* $\mathscr{R} = \frac{1}{2}\operatorname{Tr}(\Lambda D D^\top)$, *its difference from the adopted array is*

$$\mathfrak{v}_i^{\text{corr}} - \mathfrak{v}_i^{\textit{s-SGD}} = \frac{2}{\pi} N \Big[\frac{2\mu_i}{\operatorname{Tr}\Sigma_{\text{out}}} + \frac{(B + 2N + 3)\,(\Lambda D D^\top \Lambda)_{ii}}{2\mathscr{R}\,\operatorname{Tr}\Sigma_{\text{out}}} \Big] \ge 0,$$

where $\mathfrak{v}_i^{\textit{s-SGD}} = N(1 - \frac{2}{\pi}) + \frac{2}{\pi} N^2 \mu_i / \operatorname{Tr}\Sigma_{\text{out}}$. *The corrected array depends on the state, its total exceeds* N^2, *and the power-law family has no covariance lower bound fixed in* N. *The phase theorems of this subsection concern the adopted array eq.* (SI.25). *Replacing it by another nonnegative array with total* N^2 *and the same drift leaves* $\mathcal{S}_{\textit{s-SGD}}$ *unchanged, but not the kernel* $g_{\textit{s-SGD}}$ *on which the joint estimates rest.*

Full statement and proof: remark SI.4 of the supplement. Lean: not formalized (remark).

For power-law spectra the main text uses SIGNSGD kernels without the nonnegative, state-dependent term that the SIGNSGD analysis finds in the per-row volatility (the footnote to eq. (14)); this remark gives that term and the reason it is dropped. The corrected array is proved only for a covariance bounded below, which the power-law family is not, and it depends on the state, so it does not fit a model with fixed coefficients. It is at least the adopted array, and its total exceeds N^2. Any array with

total N^2 and the same drift has the same noise constant $\mathcal{S}_{\text{s-SGD}}$, hence the same matched step, but the joint estimates use the shape of the rescaled kernel $g_{\text{s-SGD}}$ (appendix G.6), so the phase theorems are statements about the adopted array only.

Remark M.2 (Momentum arrays)**.** *For $\beta \in [0,1)$ put $B_{\text{eff}} \overset{\text{def}}{=} B(1+\beta)/(1-\beta) = B/s_2$. The adopted* MUON *array with momentum β has the μ_i and $\Omega_i(0)$ of definition SI.3, the batch B_{eff} in place of B in the drift, and the volatility multiplied by B_{eff}/B:*

$$\mathfrak{d}_i^{\beta} \overset{\text{def}}{=} \sqrt{\mu_i}\sqrt{\frac{B_{\text{eff}}}{N}} = \sqrt{\frac{B_{\text{eff}}}{B}}\,\mathfrak{d}_i^{\text{s-SVD}}, \qquad \mathfrak{v}_i^{\beta} \overset{\text{def}}{=} \frac{B_{\text{eff}}}{B} = \frac{B_{\text{eff}}}{B}\,\mathfrak{v}_i^{\text{s-SVD}}.$$

At $\beta = 0$ it is the SIGNSVD *array. The evidence for it is the resolved-window endpoint of corollary L.1(ii), quoted in remark SI.1, which reads $\sqrt{2}\,\mathfrak{d}_i(0) \sim \sqrt{\mu_i B_{\text{eff}}/N}$ and $\mathfrak{v}_i(0) \to B_{\text{eff}}/B$ in this notation.*

The momentum array at a step η is the SIGNSVD *array at the step $\eta\sqrt{B_{\text{eff}}/B}$. For every $\eta > 0$, the right side of eq.* (SI.2) *satisfies*

$$-\frac{2\eta\,\mathfrak{d}_i^{\beta}}{\sqrt{\mathcal{R}}}\,\Omega_i + \eta^2\mathfrak{v}_i^{\beta} = -\frac{2\eta'\,\mathfrak{d}_i^{\text{s-SVD}}}{\sqrt{\mathcal{R}}}\,\Omega_i + \eta'^2\mathfrak{v}_i^{\text{s-SVD}}, \qquad \eta' \overset{\text{def}}{=} \eta\sqrt{\frac{B_{\text{eff}}}{B}}.$$

By eq. (SI.1) *the noise constant of the momentum array is $\mathcal{S}^{\beta} = \sqrt{B_{\text{eff}}/B}\,\mathcal{S}_{\text{s-SVD}}$, so its matched step $2\sqrt{\epsilon}/\mathcal{S}^{\beta}$ is mapped to the matched step $\eta^{\star}_{\text{s-SVD}}$. At matched steps the two models therefore solve the same row equations from the same initial values, and they have the same solution by the uniqueness in theorem SI.2. In particular $t^{\beta}_{T_\epsilon} = t^{\text{s-SVD}}_{T_\epsilon}$, and theorem M.3, corollary M.1 and proposition M.1 hold for the momentum array without change. Written with the noise constant of the momentum array, the saturated prefactor $\sqrt{\gamma}\,\mathcal{S}_{\text{s-SVD}}$ of eq.* (46) *is $\sqrt{N/B_{\text{eff}}}\,\mathcal{S}^{\beta}$.*

The oversampled isotropic case refines the volatility factor. There the volatility with momentum is $c_\beta B_{\text{eff}}/B$ times the volatility without momentum, with $c_\beta \overset{\text{def}}{=} s_2 V_\beta \in (s_2, 1)$, up to a factor $1 + O(\gamma)$ (remark K.1 and lemma K.1). A factor $c > 0$ on every volatility of an array multiplies every matched-step hitting time by c. Indeed, the noise constant becomes $c\,\mathcal{S}$ and the matched step becomes $\eta^{\star}/c$, so the right side of eq. (SI.2) *becomes $1/c$ times its value for the original array at its matched step. The risk of the modified model at time t is therefore the original risk at time t/c.*

Full statement and proof: remark SI.5 of the supplement. Lean: proved: `Muon.Phase.rem_app_phase_momentum_i`, `Muon.Phase.rem_app_phase_momentum_ii`, *+1 more.*

This remark is behind the main-text sentence that a constant factor in the volatility only rescales the time to reach a target, and behind the fact that the power-law hitting times of MUON in fig. 8 do not depend on β. The adopted array with momentum multiplies the drift by $\sqrt{B_{\text{eff}}/B}$ and the volatility by B_{eff}/B, which is the SIGNSVD array at the step $\eta\sqrt{B_{\text{eff}}/B}$; this maps the matched step to the matched step, and uniqueness of the row equations makes the two trajectories, and hence the hitting times, identical. More generally, a factor $c > 0$ on every volatility multiplies the noise constant by c and divides the matched step by c, so the right side of the row equations at the matched step is divided by c, and every hitting time is multiplied by c. In the oversampled isotropic case the volatility with momentum is $c_\beta B_{\text{eff}}/B$ times the one without, up to a factor $1 + O(\gamma)$, with $c_\beta = s_2 V_\beta \in (s_2, 1)$ of lemma K.1, the c_β of the main text.

Let $\epsilon_N = N^{-(p-1)}$ be the accuracy scale of eq. (17) and $x = \epsilon/\epsilon_N$. For an integer N let $J_N(p,q) = \int_0^{\infty} \sqrt{f_{N,q}(u)}\,\mathrm{d}u$ with $f_{N,q}(u) = \frac{1}{2}\sum_{i \le N} i^{-p} e^{-2ui^{-q}}$, and let I_j be the clock integral of appendix G.6, so that $t^j_{T_\epsilon} = \mathcal{S}_j T_0^j\,\mathsf{I}_j/(2\sqrt{\epsilon})$ exactly. Put $\rho_N = D_N\,J_N(p,\alpha)/J_N(p,\alpha/2)$, the limit of $t^{\text{s-SGD}}_{T_\epsilon}/t^{\text{s-SVD}}_{T_\epsilon}$ as $\epsilon \downarrow 0$ at fixed N (corollary SI.2).

Theorem M.4 (Frozen regime)**.** *Fix $\alpha > 0$, ς with $p = \alpha + \varsigma > 1$, and $T \ge 4$. Consider a sequence of (N, B, ϵ) with $N \to \infty$, integers $B \ge N$ and $x = \epsilon/\epsilon_N \to 0$. For all large indices both first hits are attained, and for both j*

$$\frac{\mathsf{I}_j}{J_N(p,q_j)} \longrightarrow 1, \qquad t^j_{T_\epsilon} \sim \frac{\mathcal{S}_j T_0^j\,J_N(p,q_j)}{2\sqrt{\epsilon}}, \qquad \frac{t^{\text{s-SGD}}_{T_\epsilon}}{t^{\text{s-SVD}}_{T_\epsilon}} \sim \rho_N.$$

Moreover, if $r_j < 1$, there are N_2 and c such that

$$\Big|\frac{\mathsf{I}_j}{J_N(p,q_j)} - 1\Big| \le c\left(x^{1/4} + x^{1/2}\log\left(e + x^{-1/2}\right)\right)^{1/2}$$

for all $N \ge N_2$, $B \ge N$, $\epsilon < 1/(2T)$ and $x \le 1$.

Full statement and proof: theorem SI.13 of the supplement. Lean: proved: `Muon.Phase.thm_app_phase_frozen_i`, `Muon.Phase.thm_app_phase_frozen_ii`, +4 more.

Below the accuracy scale ϵ_N, which bounds the risk carried by the modes beyond N, more accuracy no longer changes the comparison. The clock integral of each model freezes at $J_N(p,q_j)$, the integral of the square root of its noiseless risk, so each time is $\epsilon^{-1/2}$ times a constant that depends on N, and the ratio is asymptotic to its accuracy-first value ρ_N. This is the levelling off of eq. (17) below ϵ_N, and the tie of Phase D in fig. 7 once $\epsilon \lesssim \epsilon_N$. The theorem needs no window, and for $r_j < 1$ it comes with a rate in $x = \epsilon/\epsilon_N$.

Proof sketch. At finite N every rate is bounded below, which pins the clock integral to J_N at small accuracy. The forcing f_{N,q_j} and the kernel $g_j = \sum_i w_i c_i e^{-c_i u}$, whose weights sum to one, are finite sums of exponentials with rates at least $b_{\min} = 2N^{-q_j}$. The envelope $h = \sqrt{F_0}\, e^{-b_{\min}u/4}$ satisfies $g_j * h \le \frac{4}{3}h$, so $(h + \sqrt{\epsilon})^2$ dominates the equation and the first passage occurs before $(2/b_{\min})\log(F_0/\epsilon)$. After it, the forcing decays at rate at least $b_{\min}$ from a value at most $T\epsilon$, which bounds the part of J_N that is missed; and the Cauchy–Schwarz inequality with $\int g_j = 1$ gives $\mathsf{I}_j \le J_N + \sqrt{\mathsf{d}\,\mathsf{I}_j}$. With $\mathsf{y}_j = \epsilon N^{2q_j}$ this gives (lemma SI.11 and eq. (SI.59))

$$J_N - \sqrt{T\mathsf{y}_j} \le \mathsf{I}_j \le \left(\sqrt{J_N} + \sqrt{\mathsf{d}}\right)^2, \qquad \mathsf{d} = \sqrt{\mathsf{y}_j}\,\log\frac{F_0}{\epsilon}.$$

The errors are small relative to J_N when $r_j \ge 1$, and a restart removes a logarithm when $r_j < 1$. Since $2q_jr_j = p - 1$, $\sqrt{\mathsf{y}_j} = \sqrt{x}\,N^{q_j(1-r_j)}$. For $r_j > 1$, $J_N \to J(p,q_j) > 0$ and $\mathsf{y}_j \to 0$, so both errors vanish. For $r_j = 1$, $\mathsf{y}_j = x$, $J_N \sim \sqrt{q_j/8}\,\log N$ and $\mathsf{d} = o(\log N)$. For $r_j < 1$, $J_N \asymp N^{q_j(1-r_j)}$ (proposition SI.8), so the lower error has relative order $\sqrt{x}$, but d/J_N is of order $\sqrt{x}\log N$, because $\log(F_0/\epsilon)$ contains $(p-1)\log N$. The proof therefore first runs the power-law comparison of appendix G.6 up to a fixed multiple u_0 of N^{q_j}, where the kernel condition $g_j * h \le 2h$ still holds, at a cost $O(x^{1/4}+x^{1/2})N^{q_j(1-r_j)}$. At u_0 the memory term $g_j * \sqrt{Y_j}$ is itself a finite sum of exponentials with the same rates, so the equation starts afresh from $Y_j(u_0) \le \epsilon(1 + cx^{-1/2})^2$, and the bound above now carries $\log(Y_j(u_0)/\epsilon) = O(\log(e + x^{-1/2}))$ in place of $\log(F_0/\epsilon)$ (lemmas SI.21 and SI.22). This gives the stated rate. The times follow from $t^j_{T\epsilon} = \mathcal{S}_j T_0^j \mathsf{I}_j/(2\sqrt{\epsilon})$, and dividing $D_N \mathsf{I}_{\text{s-SGD}}/\mathsf{I}_{\text{s-SVD}}$ by ρ_N gives the ratio. ◇

For model j, with $q = q_j$ and $r = r_j$, put

$$\Upsilon_j(N,\epsilon) = \begin{cases} \min\left\{\epsilon^{-(1-r)/(2r)},\ N^{q(1-r)}\right\}, & r < 1, \\ \min\left\{\log(1/\epsilon),\ \log N\right\}, & r = 1, \\ 1, & r > 1. \end{cases}$$

Let $\mathcal{U}$ be the set of triples (α,ς,j) with $p > 1$ such that $r_j < 1$, or $j =$ s-SVD and $\alpha \ne 1$, or $j =$ s-SGD and $\alpha \notin \{\frac{1}{2}, 1\}$. Let $e_j = \max\{\frac{1}{2}, \frac{1}{2r_j}\}$, which are e_{Muon} and $e_{\text{s-SGD}}$ of eq. (18), and $\chi = e_{\text{s-SGD}} - e_{\text{s-SVD}}$, the κ of theorem 5.1. Finally $e_D = -\frac{\alpha}{2}, -\frac{1}{2}, \frac{\alpha}{2} - 1, 0$ for $0 < \alpha < 1$, $\alpha = 1$, $1 < \alpha < 2$, $\alpha \ge 2$, and $L = [\alpha = 1] - [\alpha = 2] + [\varsigma = \alpha + 1] - [\varsigma = 1]$, where $[\cdot]$ is 1 if the condition holds and 0 otherwise.

Corollary M.2 (Ratio at every accuracy)**.** *Fix $\alpha > 0$, ς with $\alpha + \varsigma > 1$, and $T \ge 4$, and let N_0 and ϵ_0 be thresholds of theorem SI.14 valid for both models. If $(\alpha,\varsigma,$ s-SVD$)$ and $(\alpha,\varsigma,$ s-SGD$)$ both lie in $\mathcal{U}$, then for $N \ge N_0$, $B \ge N$ and $0 < \epsilon < \epsilon_0$,*

$$\frac{t^{\text{s-SGD}}_{T\epsilon}}{t^{\text{s-SVD}}_{T\epsilon}} \asymp D_N\,\frac{\Upsilon_{\text{s-SGD}}(N,\epsilon)}{\Upsilon_{\text{s-SVD}}(N,\epsilon)}, \tag{49}$$

and if moreover $\varsigma \notin \{1, \alpha+1\}$, the right side equals $D_N \max\{\epsilon, \epsilon_N\}^{-\chi}$. If only $(\alpha, \varsigma,$ s-SGD$) \in \mathcal{U}$, the upper half of eq. (49) holds; if only $(\alpha, \varsigma,$ s-SVD$) \in \mathcal{U}$, the lower half holds. For $\epsilon = N^{-\omega}$ with fixed $\omega > 0$,

$$D_N \frac{\Upsilon_{\text{s-SGD}}(N, \epsilon)}{\Upsilon_{\text{s-SVD}}(N, \epsilon)} \asymp N^{e_D + \chi \min\{\omega, p-1\}} (\log N)^L,$$

with constants depending also on ω.

Full statement and proof: corollary SI.6 of the supplement. Lean: proved: `Muon.Phase.cor_app_phase_unified_ratio_i`, `Muon.Phase.cor_app_phase_unified_ratio_ii`, *+3 more.*

This is the precise form of theorem 5.1. For all large N, all $B \geq N$ and all small ϵ, with no window, the ratio is, when both models lie in $\mathcal{U}$, D_N times the quotient of the clock orders Υ_j, and off the lines $\varsigma = 1$ and $\varsigma = \alpha + 1$ this is $D_N \max\{\epsilon, \epsilon_N\}^{-\kappa}$. Above the accuracy scale the ratio grows like $\epsilon^{-\kappa}$ as ϵ decreases, the main text's reading of Phase C, and below ϵ_N it levels off. The set $\mathcal{U}$ excludes SIGNSVD at $\alpha = 1, \varsigma \geq 1$ and SIGNSGD at $\alpha \in \{\frac{1}{2}, 1\}, \varsigma \geq \alpha + 1$, hence the main text's exclusion of $\alpha \in \{\frac{1}{2}, 1\}$; with one model excluded, one half of the order survives. At $\epsilon = N^{-\omega}$ the ratio is a power of N times a power of $\log N$, the input of theorem M.5.

Proof sketch. The clock integral of each model is of order Υ_j at every accuracy: it grows as in appendix G.6 until ϵ reaches the accuracy scale, then freezes at the order $N^{q_j(1-r_j)}$ of J_N (theorem SI.14). The lower bound needs no kernel estimate. Since $Y_j \geq f_{N,q_j}$, the first passage is no earlier than that of the forcing, and the lower envelope of the forcing, valid up to $(N/4)^{q_j}$, keeps it above $T\epsilon$ until $\min\{\sigma_*, (N/4)^{q_j}\}$, where $1 + \sigma_* = (A_p/T\epsilon)^{1/(2r_j)}$. Integrating the envelope up to that time gives the two branches of Υ_j, which for $r_j < 1$ cross at ϵ_N because $q_j(1-r_j) = (p-1)(1-r_j)/(2r_j)$:

$$\Upsilon_j(N, \epsilon) = \max\{\epsilon, \epsilon_N\}^{-(1-r_j)/(2r_j)} \qquad (r_j < 1).$$

The upper bound needs the kernel controlled up to times of order N^{q_j}, and this is where $\mathcal{U}$ comes from. For $r_j < 1$ the comparison of appendix G.6 covers $\epsilon \geq E_* \epsilon_N$, and the comparison followed by a restart, as in theorem M.4, covers the band below. For $r_j \geq 1$ the forcing envelope decays faster than a slow part of the kernel: each kernel in $\mathcal{U}$ is bounded by an integrable fast part plus a slow part of height of order N^{-q} (lemma SI.25),

$$g_j(v) \leq K_f (1+v)^{-\upsilon} + K_s N^{-q} \Big(\frac{1+v}{N^q}\Big)^{-\mathsf{b}}, \qquad \upsilon > 1,\ 0 < \mathsf{b} < 1,\ 0 \leq v \leq N^q - 1.$$

The slow part acting on the envelope h leaves a remainder $w^2/\sqrt{\epsilon}$, where w is a power-law envelope of exponent $\mathsf{b}/2 < 1$ that the kernel does not amplify; then $(h + w + \sqrt{\epsilon})^2$ dominates the equation (lemma SI.20), and a restart as in theorem M.4 closes the bound. This bound is not available for the SIGNSVD kernel at $\alpha = 1$ and the SIGNSGD kernel at $\alpha \in \{\frac{1}{2}, 1\}$, the triples excluded from $\mathcal{U}$; SIGNSVD at $\alpha \geq 2$ is treated directly.

The ratio law is then a division. Exactly, $t^{\text{s-SGD}}_{T\epsilon}/t^{\text{s-SVD}}_{T\epsilon} = D_N \mathsf{l}_{\text{s-SGD}}/\mathsf{l}_{\text{s-SVD}}$ (eq. (SI.41)), and the lower bound holds for every model, which gives the one-sided statements. For $r_j \neq 1$, $e_j - \frac{1}{2}$ is $(1-r_j)/(2r_j)$ or 0 as $r_j < 1$ or $r_j > 1$, so $\Upsilon_j = \max\{\epsilon, \epsilon_N\}^{-(e_j - 1/2)}$ and the quotient is $\max\{\epsilon, \epsilon_N\}^{-\chi}$. At $\epsilon = N^{-\omega}$, $\Upsilon_j \asymp N^{\min\{\omega, p-1\}(e_j - 1/2)} (\log N)^{[r_j = 1]}$ and $D_N \asymp N^{e_D} (\log N)^{[\alpha=1]-[\alpha=2]}$, which gives the last display. ◇

At the accuracy $\epsilon = N^{-\omega}$, with arbitrary integers $B_N \geq N$, write $\mathsf{R} = t^{\text{s-SGD}}_{T\epsilon}/t^{\text{s-SVD}}_{T\epsilon}$. With e_D, χ and L as before corollary M.2, put $S(\omega) = e_D + \chi \min\{\omega, p-1\}$ and $\mathsf{E} = \{\alpha = 1,\ \varsigma \geq 1\} \cup \{\alpha = \frac{1}{2},\ \varsigma \geq \frac{3}{2}\}$.

Theorem M.5 (Accuracy-exponent map)**.** *Fix $\alpha > 0$, ς with $p = \alpha + \varsigma > 1$, $\omega > 0$ and $T \geq 4$. As $N \to \infty$:*

(i) if $\omega > p - 1$, then $\mathsf{R} \sim \rho_N$, so every conclusion of theorem SI.12 and corollary SI.4 holds for R;

(ii) if $(\alpha, \varsigma) \notin \mathsf{E}$, then $\mathsf{R} \asymp N^{S(\omega)} (\log N)^L$;

(iii) if $\alpha = 1$ and $1 \leq \varsigma < 2$, then $\mathsf{R} \lesssim N^{S(\omega)} (\log N)^L$.

Full statement and proof: theorem SI.15 of the supplement. Lean: proved: `Muon.Phase.thm_app_phase_omega_i`, `Muon.Phase.thm_app_phase_omega_i_transfer`, *+2 more.*

Setting $\epsilon = N^{-\omega}$ turns the ratio law into one power of N, and its exponent $S(\omega)$ is the competition of the main text's paragraph on which algorithm wins: the dimension exponent e_D, negative for $\alpha < 2$, against the accuracy exponent $\chi \min\{\omega, p-1\}$, which is nondecreasing in ω until the accuracy reaches ϵ_N at $\omega = p-1$ and is constant beyond. Part (ii) holds whenever both models lie in $\mathcal{U}$, and (iii) is the one-sided bound for $\alpha = 1$, $1 \le \varsigma < 2$, where only SIGNSGD does. Part (i) is sharper in the frozen range $\omega > p-1$: the ratio is asymptotic to the accuracy-first value ρ_N, whose limits and orders theorem SI.12 computes; where the order only says that the two are balanced, this gives the limit, the tie constants of Phases D and E and of the chord. It is the theorem the main text cites for Phase B in fig. 7.

Proof sketch. Parts (ii) and (iii) are corollary M.2 at $\epsilon = N^{-\omega}$, which lies below its threshold ϵ_0 for large N. Both models lie in $\mathcal{U}$ exactly when $(\alpha, \varsigma) \notin \mathsf{E}$, and for $\alpha = 1$, $1 \le \varsigma < 2$ only SIGNSGD does, which leaves the upper half. The exponent is explicit (eq. (SI.48)):

$$\begin{aligned}\chi(p-1) &= \max\left\{\frac{p-1}{2}, \alpha\right\} - \max\left\{\frac{p-1}{2}, \frac{\alpha}{2}\right\} \\ &= \frac{\alpha}{2}, \quad \frac{\alpha+1-\varsigma}{2}, \quad 0 \quad (\varsigma < 1,\ 1 \le \varsigma \le \alpha+1,\ \varsigma \ge \alpha+1),\end{aligned}$$

so $S(\omega) = e_D + \min\{\omega/(p-1), 1\}\,\chi(p-1)$.

Part (i) is theorem M.4, since $\omega > p-1$ means $x = N^{p-1-\omega} \to 0$; what it needs in addition is the behaviour of $\rho_N = D_N J_N(p,\alpha)/J_N(p,\alpha/2)$ as $N \to \infty$. For $r_q < 1$ the finite forcing integral grows like a power of N with an explicit constant. The substitution $u = N^q z$ turns the rescaled forcing into a Riemann sum,

$$2N^{p-1} f_{N,q}(N^q z) = \frac{1}{N}\sum_{i \le N}\left(\frac{i}{N}\right)^{-p} e^{-2z(i/N)^{-q}} \longrightarrow \int_0^1 t^{-p} e^{-2zt^{-q}}\,\mathrm{d}t = \frac{(2z)^{-2r_q}}{q}\,\Gamma(2r_q, 2z),$$

with $\Gamma(a,y) = \int_y^\infty t^{a-1}e^{-t}\mathrm{d}t$. Halving every rate gives a majorant $4^r C_{p,q}^2 z^{-2r} e^{-z}$ that does not depend on N (lemma SI.19), so dominated convergence gives (proposition SI.8)

$$\frac{J_N(p,q)}{N^{q(1-r_q)}} \longrightarrow (8q)^{-1/2}\mathcal{K}(r_q), \qquad \mathcal{K}(r) = \int_0^\infty y^{-r}\,\Gamma(2r,y)^{1/2}\,\mathrm{d}y, \qquad \mathcal{K}(\tfrac{1}{2}) = \sqrt{2\pi}.$$

For $r_q = 1$, $J_N(p,q) \sim \sqrt{q/8}\,\log N$, and for $r_q > 1$, $J_N(p,q) \to J(p,q)$. Multiplying the quotient of these by the order of D_N gives each limit and order of theorem SI.12. For example, at $\alpha = \varsigma = 1$ the factors $D_N \sim \frac{\pi}{2} N^{-1/2}\log N$ and $J_N(p,1)/J_N(p,\frac{1}{2}) \sim 2\sqrt{\pi}\,N^{1/2}/\log N$ multiply to $\pi^{3/2}$. $\diamond$

SIGNSVD wins, SIGNSGD wins, or the two are balanced according as $\mathsf{R} \to \infty$, $\mathsf{R} \to 0$ or $\mathsf{R} \asymp 1$. For $1 < \alpha < 2$ and $\varsigma < 2\alpha - 1$, ω_c is the critical exponent of the main text: $\omega_c = (2-\alpha)(p-1)/\alpha$ if $\varsigma < 1$ and $\omega_c = (2-\alpha)(p-1)/(\alpha+1-\varsigma)$ if $1 \le \varsigma < 2\alpha - 1$, so that $0 < \omega_c < p-1$. The constants K_A and K_c are the explicit positive limits of ρ_N given by theorem SI.12 for $0 < \alpha < 1$, $\varsigma < 1$ and on the chord $\varsigma = 2\alpha - 1$, $1 < \alpha < 2$.

Corollary M.3 (Phases of the accuracy-exponent map)**.** *Under the hypotheses of theorem M.5:*

- *(a)* Phase A, $\alpha > 2$, $\varsigma \le \alpha + 1$*:* SIGNSVD *wins for every* ω.
- *(b)* Phase B, $0 < \alpha < 1$ *and* $\varsigma \ge 1$ *with* $(\alpha, \varsigma) \notin \mathsf{E}$*, or* $1 < \alpha < 2$ *and* $\varsigma > 2\alpha - 1$*:* SIGNSGD *wins for every* ω.
- *(c)* Phase C, $1 < \alpha < 2$, $\varsigma < 2\alpha - 1$*:* SIGNSGD *wins if* $\omega < \omega_c$ *and* SIGNSVD *wins if* $\omega > \omega_c$*; at* $\omega = \omega_c$ *the two are balanced if* $\varsigma \ne 1$*, and* SIGNSGD *wins, with order* $1/\log N$*, if* $\varsigma = 1$.
- *(d)* Phase D, $0 < \alpha < 1$, $\varsigma < 1$*:* SIGNSGD *wins if* $\omega < p-1$*, and the two are balanced if* $\omega \ge p-1$*, with limit* K_A *if* $\omega > p-1$.
- *(e)* Tie, $\alpha > 2$, $\varsigma > \alpha + 1$*: balanced for every* ω.
- *(f)* Chord, $1 < \alpha < 2$, $\varsigma = 2\alpha - 1$*:* SIGNSGD *wins if* $\omega < p-1$*, and the two are balanced if* $\omega \ge p-1$*, with limit* K_c *if* $\omega > p-1$.

(g) $\alpha = 1$*: for* $\varsigma < 1$, SIGNSGD *wins if* $\omega < p - 1$ *and* SIGNSVD *wins, with order* $\log N$*, if* $\omega \geq p - 1$*; for* $\varsigma = 1$, SIGNSGD *wins if* $\omega < 1$*, and* $\mathsf{R} \to \pi^{3/2}$ *if* $\omega > 1$*; for* $1 < \varsigma < 2$, SIGNSGD *wins for every* ω.

(h) $\alpha = 2$*:* SIGNSVD *wins for* $\varsigma < 3$*, the two are balanced for* $\varsigma = 3$*, and* SIGNSGD *wins for* $\varsigma > 3$*, for every* ω.

Full statement and proof: corollary SI.7 of the supplement. Lean: proved: `Muon.Phase.cor_app_phase_omega_winners_a`*,* `Muon.Phase.cor_app_phase_omega_winners_b`*, +6 more.*

This is the phase diagram of fig. 7 in exact form, with its boundary lines; the Tie is the main text's Phase E. In Phase A MUON wins at every accuracy, in Phase B SIGNSGD does, in Phase C SIGNSGD wins at low accuracy and MUON once $\epsilon \ll N^{-\omega_c}$, and in Phase D SIGNSGD wins above the accuracy scale and the two tie below it. The chord $\varsigma = 2\alpha - 1$ between B and C is where ω_c reaches $p - 1$, that is, where the critical accuracy meets the accuracy scale, so as $\omega \to \infty$ the boundary between B and C becomes the chord. The level $\omega_c = 1$ is the curve $\varsigma_c(\alpha)$ of proposition M.1, the cut of the map at $\epsilon = N^{-1}$, where that proposition's window ends. Not covered are the winner for $\omega \leq p - 1$ on $\{\alpha = 1, \varsigma \geq 2\} \cup \{\alpha = \frac{1}{2}, \varsigma \geq \frac{3}{2}\}$, and the limit constants for $\omega \leq p - 1$ in the balanced cases; at $\alpha = \varsigma = \omega = 1$ only $\mathsf{R} \lesssim 1$ is proved.

Proof sketch. Off E the winner is the sign of $S(\omega)$, and of L where $S(\omega) = 0$ (theorem M.5(ii)). Write $\omega' = \min\{\omega, p - 1\}$. For $0 < \alpha < 1$, $e_D = -\frac{\alpha}{2}$. If $\varsigma < 1$, then $S(\omega) = \frac{\alpha}{2}(\omega'/(p-1) - 1)$ and $L = 0$; this is negative below the accuracy scale and zero at and beyond it, which is Phase D, with the limit K_A of theorem M.5(i). If $\varsigma \geq 1$, then $\chi(p-1) = \max\{\frac{\alpha+1-\varsigma}{2}, 0\} \leq \frac{\alpha}{2}$, so $S(\omega) \leq 0$, with equality only at $\varsigma = 1$, $\omega \geq p - 1$, where $L = -1$; this is the first part of Phase B. For $1 < \alpha < 2$, $e_D = \frac{\alpha}{2} - 1 < 0$, and for $1 \leq \varsigma \leq \alpha + 1$

$$S(\omega) = \frac{\alpha}{2} - 1 + \frac{\omega'}{p-1} \cdot \frac{\alpha + 1 - \varsigma}{2},$$

which vanishes at $\omega' = \omega_c$. At $\omega' = p - 1$ it equals $(2\alpha - 1 - \varsigma)/2$, so $\omega_c < p - 1$ exactly when $\varsigma < 2\alpha - 1$. Below the chord SIGNSGD wins for $\omega < \omega_c$ and SIGNSVD for $\omega > \omega_c$, which is Phase C; at $\omega = \omega_c$, $L = -[\varsigma = 1]$ gives the balance or the win of order $1/\log N$. On the chord $S(\omega) < 0$ for $\omega < p - 1$ and $S = L = 0$ beyond, with the limit K_c; above it, and for $\varsigma > \alpha + 1$, $S(\omega) < 0$, which is the second part of Phase B. The case $\varsigma < 1$ is the same with α in place of $\alpha + 1 - \varsigma$.

For $\alpha \geq 2$ the dimension factor does not decay, $e_D = 0$, and only the accuracy factor and the logarithms count. For $\alpha > 2$, $S(\omega) > 0$ when $\varsigma < \alpha + 1$, which is Phase A; for $\varsigma \geq \alpha + 1$, $S(\omega) = 0$ and $L = [\varsigma = \alpha + 1]$, which gives SIGNSVD on the line and the Tie beyond it. At $\alpha = 2$, $S(\omega) > 0$ for $\varsigma < 3$, and $S(\omega) = 0$ with $L = -1 + [\varsigma = 3]$ for $\varsigma \geq 3$. At $\alpha = 1$ and $1 \leq \varsigma < 2$ only the upper bound of theorem M.5(iii) is available. There $S(\omega) = -\frac{1}{2} + \frac{\omega'}{p-1} \cdot \frac{2-\varsigma}{2} < 0$ unless $\varsigma = 1$ and $\omega \geq p - 1$, which forces $\mathsf{R} \to 0$; at $\varsigma = 1$, $\omega > 1$ the frozen limit $\pi^{3/2}$ of ρ_N applies. ◇

N SignSGD

SIGNSGD keeps no momentum buffer, so this chapter has no momentum version, and every statement concerns one step from a given state Δ with one fresh batch; none is a statement about a trajectory. The step is $\Delta^+ = \Delta - \eta S$ with $S = \operatorname{sign}(G^\circ)$, the entrywise sign of the minibatch gradient $G^\circ = \frac{1}{B}\sum_a (x_{\text{out}}^{(a)})^\top \Delta x_{\text{in}}^{(a)}\, x_{\text{out}}^{(a)} \otimes x_{\text{in}}^{(a)}$, which is, when the risk is positive, also the sign of the normalized gradient G_{t+1} of eq. (5). Since $\|S\|_{\text{F}}^2 \leq N_{\text{out}} N_{\text{in}}$, the exact one-step identities of the risk and of the row energy $\mathscr{Q}_i(\Delta) = \|\Delta^\top u_i\|^2$ apply to it (eqs. (SJ.2) and (SJ.3)), with $\langle V, W\rangle_\Sigma = \operatorname{Tr}(\Sigma_{\text{out}} V \Sigma_{\text{in}} W^\top)$ and $\|V\|_\Sigma^2 = \langle V, V\rangle_\Sigma$:

$$\mathbb{E}\,\mathscr{R}(\Delta^+) = \mathscr{R}(\Delta) - \eta\,\langle \Delta, \mathbb{E}\,S\rangle_\Sigma + \frac{\eta^2}{2}\,\mathbb{E}\,\|S\|_\Sigma^2,$$
$$\mathbb{E}\,\mathscr{Q}_i(\Delta^+) = \mathscr{Q}_i(\Delta) - 2\eta\,\mathbb{E}\langle \Delta^\top u_i, S^\top u_i\rangle + \eta^2\,\mathbb{E}\,\|S^\top u_i\|^2.$$

The drift constant of the main text is $\mathcal{N}_B = \mathbb{E}\left[(\sum_{b=1}^B x_b^2 y_b^2)^{1/2}\right]$ for independent standard Gaussians x_b, y_b (eq. (SJ.4)). The chapter bounds $\mathcal{N}_B$, evaluates the two expectations in the risk identity for

isotropic data, and then, for a Haar eigenbasis, evaluates the row volatility and assembles the row recursion.

Proposition N.1 (The batch constant)**.** *For every integer $B \geq 1$, the constant eq.* (SJ.4) *satisfies $\mathcal{N}_1 = 2/\pi$ and*

$$\max\Big\{\frac{2}{\pi}\sqrt{B},\ \sqrt{B} - \frac{4}{\sqrt{B}}\Big\} \leq \mathcal{N}_B \leq \sqrt{B}, \qquad 0 \leq 1 - \frac{\mathcal{N}_B}{\sqrt{B}} \leq \frac{4}{B}. \tag{50}$$

In particular $\mathcal{N}_B/\sqrt{B} \to 1$ as $B \to \infty$.

Full statement and proof: proposition SJ.1 of the supplement. Lean: proved: `Muon.SignSGD.prop_app_sgd_batch_coefficient_i`, `Muon.SignSGD.prop_app_sgd_batch_coefficient_ii`, *+2 more.*

The batch constant is the drift coefficient of SIGNSGD: the main text writes the isotropic drift as $\mathfrak{d} = \mathcal{N}_B/\sqrt{\pi}$ and then uses $\mathcal{N}_B \approx \sqrt{B}$ throughout, which gives the tabulated drift $\mu_i\sqrt{B/(\pi\bar{\mu})}$ of table 1. The proposition justifies this replacement with an explicit error: $\mathcal{N}_B$ differs from $\sqrt{B}$ by a factor in $[1 - 4/B, 1]$, and it is never smaller than $\frac{2}{\pi}\sqrt{B}$, a bound attained at $B = 1$.

Proof sketch. The constant is the mean of the square root of a sum $T_B = \sum_b x_b^2 y_b^2$ of B independent terms with mean 1 and second moment $(\mathbb{E}\,x^4)(\mathbb{E}\,y^4) = 9$, so T_B/B has mean 1 and variance $8/B$. At $B = 1$ independence gives $\mathcal{N}_1 = (\mathbb{E}\,|g|)^2 = 2/\pi$. Jensen's inequality gives $\mathcal{N}_B \leq (\mathbb{E}\,T_B)^{1/2} = \sqrt{B}$, and the Cauchy–Schwarz comparison $(\sum_b z_b^2)^{1/2} \geq B^{-1/2}\sum_b |z_b|$ with $z_b = x_b y_b$ gives $\mathcal{N}_B \geq \frac{2}{\pi}\sqrt{B}$.

The Jensen deficit is at most half the variance. Rationalizing $\sqrt{s} - 1$ gives the exact identity, for $s \geq 0$,

$$\sqrt{s} = 1 + \frac{s-1}{2} - \frac{(s-1)^2}{2(\sqrt{s}+1)^2},$$

whose last term lies between 0 and $(s-1)^2/2$. At $s = T_B/B$ the linear term has mean zero, so $0 \leq 1 - \mathbb{E}\,\sqrt{T_B/B} \leq \frac{1}{2}\cdot\frac{8}{B} = \frac{4}{B}$, which is the second inequality of eq. (50) and the bound $\mathcal{N}_B \geq \sqrt{B} - 4/\sqrt{B}$. ◇

In the next proposition constants $K \geq 1$, $0 < f_- \leq f_+$ and $C_\Delta > 0$ are fixed, and a state Δ with batch size B is in the isotropic regime of definition SJ.2 if, with $k \stackrel{\text{def}}{=} \min\{N_{\text{out}}, N_{\text{in}}, B\}$,

$$\max\{N_{\text{out}}, N_{\text{in}}, B\} \leq Kk, \qquad f_- \leq \|\Delta\|_{\text{F}} \leq f_+, \qquad \max_{i,j}|\Delta_{ij}| \leq \frac{C_\Delta}{k}.$$

The threshold is $k_0 = \max\{4, 8KC_\Delta^2/f_-^2\}$, C_2 is the universal constant of the sign comparison lemma SJ.5, and C_r depends only on K, f_-, C_Δ.

Proposition N.2 (Isotropic SIGNSGD recursion)**.** *Let $\Sigma_{\text{in}} = I_{\text{in}}$ and $\Sigma_{\text{out}} = I_{\text{out}}$, and let Δ_t be the state at step t. Suppose that, almost surely, Δ_t with the batch size B is in the isotropic regime of definition SJ.2 and $k \geq k_0$. Then the* SIGNSGD *step satisfies*

$$\mathbb{E}_t\left[\mathscr{R}(t+1)\right] = \mathscr{R}(t) - \eta\,\mathcal{N}_B\sqrt{\frac{4\mathscr{R}(t)}{\pi}} + \frac{\eta^2}{2}\,N_{\text{in}}N_{\text{out}} + E_t, \qquad |E_t| \leq \eta\,K\sqrt{2\mathscr{R}(t)}\,\Big(C_2 + \frac{C_r}{\sqrt{k}}\Big), \tag{51}$$

with C_2 universal and C_r from theorem SJ.1. The error is at most $\pi K(C_2 + C_r)/(2\sqrt{2/\pi}\sqrt{k})$ times the drift term.

Full statement and proof: proposition SJ.4 of the supplement. Lean: proved: `Muon.SignSGD.prop_app_sgd_iso_filtered`, `Muon.SignSGD.prop_app_sgd_iso_frozen`.

This is the SIGNSGD case of the scalar recursion eq. (8), with drift $\mathfrak{d} = \mathcal{N}_B/\sqrt{\pi}$ and volatility $\mathfrak{v} = N_{\text{in}}N_{\text{out}}$ (N^2 in the square case). It holds one step at a time, from any state in the isotropic regime, whose entries are all of order $1/k$, with no hypothesis on its rank or operator norm. The volatility is exact, and the drift carries a relative error of order $k^{-1/2}$.

Proof sketch. The volatility is exact. With identity covariances $\|S\|_\Sigma = \|S\|_{\text{F}}$, and every entry of G° is a polynomial in the Gaussian batch that is not identically zero, hence nonzero almost surely.

So $\|S\|_{\mathrm{F}}^2 = N_{\mathrm{in}}N_{\mathrm{out}}$ (proposition SJ.2), and the risk identity reads $\mathbb{E}_t[\mathscr{R}(t+1)] = \mathscr{R}(t) - \eta\,\mathsf{D} + \frac{\eta^2}{2}N_{\mathrm{in}}N_{\mathrm{out}}$ with $\mathsf{D} = \langle \Delta_t, \mathbb{E}_t\, S\rangle_{\mathrm{F}}$.

Each expected sign is compared with that of a shifted Gaussian. With $F = \|\Delta\|_{\mathrm{F}}$, the entry BG_{ij}° is a sum of B independent copies of $X = (z^\top \Delta w) z_i w_j$, of mean Δ_{ij} and variance $s^2 = F^2 + O(1/k)$ (lemma SJ.3). Writing the sign as a limit of $\frac{2}{\pi}\int_0^\infty e^{-\varepsilon u}\sin(uv)\,\mathrm{d}u/u$ compares expected signs through the characteristic function ϕ_Y of the standardized coordinate Y, uniformly in the shift $x = \sqrt{B}\Delta_{ij}/s$ and with universal C_1, C_2 (lemma SJ.5):

$$\Big|\,\mathbb{E}\,\mathrm{sign}\Big(\frac{1}{\sqrt{B}}\sum_{a=1}^{B} Y_a + x\Big) - \mathbb{E}\,\mathrm{sign}(g+x)\Big| \le \frac{C_1|\,\mathbb{E}\,Y^3|}{\sqrt{B}} + \frac{C_2}{B}.$$

This needs $|\phi_Y(t)|$ to decay like $|t|^{-1/2}$, and that is where small entries enter. Given the other coordinates, X is Gaussian in the coordinates of z other than z_i, with a variance controlled by the block $\bar{\Delta}$ of Δ without row i and column j; every row and column of Δ has squared norm at most KC_Δ^2/k, so $\|\bar{\Delta}\|_{\mathrm{F}}^2 \ge \frac{3}{4}F^2$ for $k \ge k_0$ (theorem SJ.1), more than the $\|\bar{\Delta}\|_{\mathrm{F}} \ge F/2$ that lemma SJ.4 needs. The third cumulant $\mathbb{E}\,Y^3$ is $O(1/k)$ as well. With $\mathbb{E}\,\mathrm{sign}(g+x) = \sqrt{2/\pi}\,x + O(|x|^3)$ and $|\sqrt{B} - \mathcal{N}_B| \le 4/\sqrt{B}$ (proposition N.1), every entry satisfies (theorem SJ.1)

$$\Big|\,\mathbb{E}\,\mathrm{sign}(G_{ij}^{\circ}) - \sqrt{\frac{2}{\pi}}\,\mathcal{N}_B\,\frac{\Delta_{ij}}{F}\Big| \le \frac{C_2}{B} + \frac{C_r}{k^{3/2}}.$$

The sum over entries costs the ℓ^1 norm of the state, which is why the error is relative of order $k^{-1/2}$. Since $\sum_{ij}\Delta_{ij}^2/F = F$, the leading terms add up to $\mathsf{D}_0 = \sqrt{2/\pi}\,\mathcal{N}_B F = \mathcal{N}_B\sqrt{4\mathscr{R}/\pi} \ge (2/\pi)^{3/2}\sqrt{k}F$, while $\sum_{ij}|\Delta_{ij}| \le \sqrt{N_{\mathrm{out}}N_{\mathrm{in}}}\,F \le KkF$ bounds the summed errors by $KF(C_2 + C_r/\sqrt{k})$, which is of order one. ⋄

The Haar model of definition SJ.3 fixes $K \ge 1$, $0 < c \le L$, $0 < f_- \le f_+$ and $C_{\mathrm{col}} > 0$, dimensions with $k \overset{\mathrm{def}}{=} \min\{N_{\mathrm{out}}, N_{\mathrm{in}}, B\} \ge 2$ and $\max\{N_{\mathrm{out}}, N_{\mathrm{in}}, B\} \le Kk$, and deterministic $\Lambda = \mathrm{diag}(\mu_1, \dots, \mu_{N_{\mathrm{out}}})$ and $D \in \mathbb{R}^{N_{\mathrm{out}}\times N_{\mathrm{in}}}$ with

$$c\,I_{\mathrm{out}} \preceq \Lambda \preceq L\,I_{\mathrm{out}}, \qquad f_- \le \|D\|_{\mathrm{F}} \le f_+, \qquad \max_j \|D_{\cdot j}\|^2 \le \frac{C_{\mathrm{col}}}{k}.$$

With U Haar distributed on $O(N_{\mathrm{out}})$ and independent of the whitened batch $(z_b, w_b)_{b\le B}$, it sets $\Sigma_{\mathrm{out}} = U\Lambda U^\top$, $\Sigma_{\mathrm{in}} = I_{\mathrm{in}}$, $\Delta = UD$, $x_{\mathrm{out}}^{(b)} = U\Lambda^{1/2}z_b$ and $x_{\mathrm{in}}^{(b)} = w_b$. Then $u_i = Ue_i$, $\mathscr{Q}_i(\Delta) = \|D_{i\cdot}\|^2$, the risk $2\mathscr{R} = \mathrm{Tr}(\Lambda D D^\top)$ is deterministic, $\bar{\mu} = \mathrm{Tr}\,\Lambda/N_{\mathrm{out}}$ and $T_D = \Lambda D D^\top \Lambda$. Expectations integrate the batch and U, a constant is admissible if it depends only on $K, c, L, f_\pm, C_{\mathrm{col}}$, and the main text's row volatility is $\mathfrak{v}_i^{\text{s-SGD}} = N_{\mathrm{in}}\big[1 + \frac{2}{\pi}\big(\frac{\mu_i}{\bar{\mu}} - 1\big)\big]$.

Theorem N.1 (Corrected row volatility)**.** *In the Haar model of definition SJ.3, define the raw annealed row volatility of the update and the corrected array*

$$\mathfrak{v}_i^{(0)} \overset{\mathrm{def}}{=} \mathbb{E}\,\|S^\top u_i\|^2, \tag{52}$$

$$\mathfrak{v}_i^{\mathrm{corr}} \overset{\mathrm{def}}{=} N_{\mathrm{in}}\Big(1 - \frac{2}{\pi}\Big) + \frac{2}{\pi}\,N_{\mathrm{out}}\Big[\frac{(N_{\mathrm{in}}+2)\,\mu_i}{\mathrm{Tr}\,\Lambda} + \frac{(B + 2N_{\mathrm{in}} + 3)\,(T_D)_{ii}}{2\mathscr{R}\ \mathrm{Tr}\,\Lambda}\Big]. \tag{53}$$

There is an admissible constant C_{row} such that, for every i,

$$|\mathfrak{v}_i^{(0)} - \mathfrak{v}_i^{\mathrm{corr}}| \le C_{\mathrm{row}}\,\frac{N_{\mathrm{in}}}{\sqrt{k}}, \qquad \Big|\frac{\mathfrak{v}_i^{(0)}}{\mathfrak{v}_i^{\mathrm{corr}}} - 1\Big| \le \frac{C_{\mathrm{row}}}{(1-2/\pi)\sqrt{k}}. \tag{54}$$

The actual total is exact, $\sum_i \mathfrak{v}_i^{(0)} = N_{\mathrm{out}}N_{\mathrm{in}}$, while

$$\sum_i \mathfrak{v}_i^{\mathrm{corr}} - N_{\mathrm{out}}N_{\mathrm{in}} = \frac{2}{\pi}\,N_{\mathrm{out}}\Big[2 + \frac{(B + 2N_{\mathrm{in}} + 3)\,\mathrm{Tr}\,T_D}{2\mathscr{R}\ \mathrm{Tr}\,\Lambda}\Big] \in \Big(0,\ \frac{2}{\pi}N_{\mathrm{out}}\big[2 + (3K+3)L/c\big]\Big). \tag{55}$$

The corrected array exceeds the main text's array row by row:

$$\mathfrak{v}_i^{\mathrm{corr}} - \mathfrak{v}_i^{\textit{s-SGD}} = \frac{2}{\pi}\,N_{\mathrm{out}}\Big[\frac{2\mu_i}{\mathrm{Tr}\,\Sigma_{\mathrm{out}}} + \frac{(B + 2N_{\mathrm{in}} + 3)\,(\Lambda D D^\top \Lambda)_{ii}}{2\mathscr{R}\ \mathrm{Tr}\,\Sigma_{\mathrm{out}}}\Big] \ge 0. \tag{56}$$

Full statement and proof: theorem SJ.4 of the supplement. Lean: proved: `Muon.SignSGD.thm_app_sgd_modes_i`, `Muon.SignSGD.thm_app_sgd_modes_ii`, *+2 more.*

The theorem computes the SIGNSGD row volatility $\mathfrak{v}_i$ of eq. (14) on a Haar eigenbasis and is the source of the footnote of section 4: the actual row volatility is the corrected array $\mathfrak{v}_i^{\mathrm{corr}}$ up to a relative error of order $k^{-1/2}$, and this array exceeds the main text's $\mathfrak{v}_i^{\text{s-SGD}}$ by a nonnegative term that depends on the state through $(T_D)_{ii}$. In total the excess is of order N_{out} against $N_{\mathrm{out}}N_{\mathrm{in}}$, but for a single row it can be of the same order as $\mathfrak{v}_i^{\text{s-SGD}}$: for $\Lambda = \mathrm{Id}$, $B = N$ and $D = e_1 \otimes \mathbf{1}/\sqrt{N}$ the first row has $\mathfrak{v}_1^{\mathrm{corr}} = (1 + \frac{6}{\pi})N + \frac{10}{\pi}$, whereas $\mathfrak{v}_1^{\text{s-SGD}} = N$.

Proof sketch. Averaging over the eigenbasis reduces the row volatility exactly to the column directions of a matrix that does not involve U. The residuals do not involve U, so $G^\circ = UH$ with $H = \frac{1}{B}\sum_b (z_b^\top \Lambda^{1/2} D w_b)\, \Lambda^{1/2} z_b \otimes w_b$ independent of U, and the jth column of $U^\top S$ is $U^\top \operatorname{sign}(Uh_j)$. For a fixed nonzero $a \in \mathbb{R}^n$ and $Y = U^\top \operatorname{sign}(Ua)$, invariance under the rotations that fix a leaves two unknown numbers in $\mathbb{E}_U[Y \otimes Y]$; the trace is $\|Y\|^2 = n$, and $a^\top Y/\|a\|$ is the ℓ^1 norm of a uniform unit vector, whose second moment is $1 + 2(n-1)/\pi$. This gives (lemmas SJ.7 and SJ.11)

$$\mathbb{E}_U\left[Y \otimes Y\right] = \Big(1 - \frac{2}{\pi}\Big)\,\mathrm{Id} + \frac{2n}{\pi}\,\frac{a \otimes a}{\|a\|^2}, \qquad \mathfrak{v}_i^{(0)} = N_{\mathrm{in}}\Big(1 - \frac{2}{\pi}\Big) + \frac{2}{\pi}\,N_{\mathrm{out}} \sum_{j=1}^{N_{\mathrm{in}}} \mathbb{E}\,\frac{(h_j)_i^2}{\|h_j\|^2}.$$

Every entry of UH is nonzero almost surely, so $\|S\|_{\mathrm{F}}^2 = N_{\mathrm{out}}N_{\mathrm{in}}$ and the total $\sum_i \mathfrak{v}_i^{(0)} = \|S\|_{\mathrm{F}}^2$ is exact.

The random column norm can be replaced by $x_0 = 2\mathscr{R}\,\mathrm{Tr}\,\Lambda/B$. The column h_j is the mean of B independent vectors of mean $m_j = \Lambda D_{\cdot j}$, so $\|h_j\|^2$ differs from x_0 by a bias of order $1/k$ and a fluctuation whose fourth moment is of order k^{-2} (a fourth-moment count over the samples), so $\mathbb{E}\,|(h_j)_i^2/\|h_j\|^2 - (h_j)_i^2/x_0| = O(k^{-3/2})$ (lemma SJ.13), and summed over the N_{in} columns with the factor $\frac{2}{\pi}N_{\mathrm{out}} \le \frac{2}{\pi}Kk$ this is the error $C_{\mathrm{row}}N_{\mathrm{in}}/\sqrt{k}$.

The replaced sum is an exact Gaussian moment, and the mean of the columns produces the state term. Gaussian pairings give (eq. (SJ.34))

$$\sum_j \mathbb{E}\left[h_j \otimes h_j\right] = \mathbb{E}\left[HH^\top\right] = \frac{2\mathscr{R}\,(N_{\mathrm{in}} + 2)}{B}\,\Lambda + \Big(1 + \frac{2N_{\mathrm{in}} + 3}{B}\Big)T_D,$$

where $T_D = \sum_j m_j \otimes m_j$ collects the squared column means. The (i,i) entry of $\mathbb{E}[HH^\top]$ divided by x_0 is the bracket of eq. (53). Summing over i and subtracting the main text's array give eqs. (55) and (56), with $\mathrm{Tr}\,T_D \le 2\mathscr{R}L$ and $\mathrm{Tr}\,\Lambda \ge cN_{\mathrm{out}}$ for the upper bound. ◇

In the next proposition the Haar model and its notation are those set before theorem N.1. The threshold is $k_* = \max\{4, \lceil 4LC_{\mathrm{col}}/(cf_-^2)\rceil\}$, C_* is the admissible constant of the half-anisotropic drift theorem SJ.2, and C_{row} is the constant of theorem N.1.

Proposition N.3 (Half-anisotropic SIGNSGD recursion)**.** *In the Haar model of definition SJ.3 with $k \ge k_*$, let $\eta \ge 0$ be deterministic and $\Delta^+ = \Delta - \eta S$. Then, with $\mathfrak{d}_i \overset{\mathrm{def}}{=} \mu_i\mathcal{N}_B/\sqrt{\pi\bar{\mu}}$, there are real numbers $E_1, \dots, E_{N_{\mathrm{out}}}$ such that*

$$\mathbb{E}\;\mathscr{Q}_i(\Delta^+) = \mathscr{Q}_i(\Delta) - \frac{2\eta\,\mathfrak{d}_i}{\sqrt{\mathscr{R}}}\,\mathscr{Q}_i(\Delta) + \eta^2\,\mathfrak{v}_i^{\mathrm{corr}} + E_i, \qquad \sum_i |E_i| \le 2\eta f_+\sqrt{C_*} + \eta^2 C_{\mathrm{row}}\,\frac{N_{\mathrm{out}}N_{\mathrm{in}}}{\sqrt{k}}. \tag{57}$$

The risk satisfies

$$\mathbb{E}\;\mathscr{R}(\Delta^+) = \mathscr{R} - \eta\sum_i \frac{\mu_i\,\mathfrak{d}_i}{\sqrt{\mathscr{R}}}\,\mathscr{Q}_i(\Delta) + \frac{\eta^2}{2}\sum_i \mu_i\,\mathfrak{v}_i^{\text{s-SGD}} + \mathcal{E}_\eta, \tag{58}$$

with $\mathcal{E}_\eta$ bounded as in corollary SJ.2.

Full statement and proof: proposition SJ.8 of the supplement. Lean: proved: `Muon.SignSGD.prop_signsgd_recursion_anisotropic`.

This is the one-step row recursion eq. (14) for SIGNSGD on a Haar eigenbasis, which the main text cites for the claim that SIGNSGD admits the same row recursion as MUON. The drift $\mathfrak{d}_i = \mu_i \mathcal{N}_B/\sqrt{\pi\bar{\mu}}$ is the tabulated $\mu_i\sqrt{B/(\pi\bar{\mu})}$ of table 1 with $\sqrt{B}$ replaced by $\mathcal{N}_B$. The per-row volatility is the state-dependent $\mathfrak{v}_i^{\mathrm{corr}}$ of theorem N.1, and the main text's array $\mathfrak{v}_i^{\text{s-SGD}}$ is correct in the covariance-weighted total, the risk identity eq. (58). The hypotheses of the Haar model, among them a covariance bounded below and a state independent of the eigenbasis, are not shown to persist after the step.

Proof sketch. The whitened batch is independent of U, so the row identity of the chapter introduction holds given U; averaged over U, its quadratic term is $\eta^2\mathfrak{v}_i^{(0)}$, which theorem N.1 replaces by $\eta^2\mathfrak{v}_i^{\mathrm{corr}}$. It remains to compute the mean update $M(U) = \mathbb{E}[S \mid U]$.

Each entry of $M(U)$ is an isotropic coordinate drift for a whitened state with small entries. Rotating the output feature so that its pth coordinate becomes a multiple of the first shows that $\operatorname{sign}(G^{\circ}_{pj})$ has the law of an entry of the isotropic gradient of a state of Frobenius norm $\sqrt{2\mathscr{R}}$ whose entry is $\mu_{pj} = (\Sigma_{\mathrm{out}}\Delta)_{pj}/\sqrt{v_p}$, with $v_p = (\Sigma_{\mathrm{out}})_{pp}$ (lemma SJ.8). The coordinate estimate of the isotropic case (proposition SJ.3) then gives $\sqrt{2/\pi}\,\mathcal{N}_B\mu_{pj}/\sqrt{2\mathscr{R}}$. Its hypothesis, a row and a column that carry a small part of the risk, is supplied by the rotation: with ω_p the pth row of U, a uniform unit vector, $\mu_{pj} = \omega_p^{\top}\Lambda D_{\cdot j}/\sqrt{v_p}$ has $\mathbb{E}\,|\mu_{pj}|^{2q} = O((N_{\mathrm{out}}k)^{-q})$, and the row is heavy only with probability $O({N_{\mathrm{out}}}^{-4})$ (lemma SJ.9). The coordinate errors are squared and summed before averaging over U, and their total stays bounded (proposition SJ.6).

The diagonal of the covariance concentrates, $\mathbb{E}(v_p - \bar{\mu})^4 \le 60L^4/{N_{\mathrm{out}}}^2$ (lemma SJ.6), and replacing v_p by $\bar{\mu}$ gives (theorem SJ.2)

$$\mathbb{E}_U \left\| U^{\top}M(U) - L_B\,\Lambda D \right\|_{\mathrm{F}}^2 \le C_*, \qquad L_B = \frac{\mathcal{N}_B}{\sqrt{\pi\mathscr{R}\,\bar{\mu}}}, \qquad \|L_B\Lambda D\|_{\mathrm{F}}^2 \asymp B.$$

The row pairing given U is $\langle D_{i\cdot}, (U^{\top}M)_{i\cdot}\rangle$, so the Cauchy–Schwarz inequality bounds the summed row errors by $\|D\|_{\mathrm{F}}\sqrt{C_*} \le f_+\sqrt{C_*}$, against a total of order $\sqrt{B}$ (corollary SJ.1), and $L_B\mu_i = \mathfrak{d}_i/\sqrt{\mathscr{R}}$. For the risk, the Haar sign matrix (lemma SJ.11) reduces the weighted volatility to $N_{\mathrm{in}}(1 - \frac{2}{\pi})\operatorname{Tr}\Lambda + \frac{2}{\pi}N_{\mathrm{out}}\sum_j \mathbb{E}[h_j^{\top}\Lambda h_j/\|h_j\|^2]$, with $h_j = U^{\top}G^{\circ}_{\cdot j}$, and the mean of each ratio is $\operatorname{Tr}\Lambda^2/\operatorname{Tr}\Lambda$ up to $O(k^{-1/2})$ by concentration of quadratic forms of a column, which gives $\sum_i \mu_i\mathfrak{v}_i^{\text{s-SGD}}$ (theorem SJ.3 and corollary SJ.2). ◇

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: We are explicit in the abstract and introduction that we are studying a simplified model, matrix-valued least squares problem, with both MUON and SIGNSVD.

2. **Limitations**

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

   Answer: [Yes]

   Justification: In the conclusion paragraph, we discuss the limitations of the work including the simplified setting as well as only analyzing the dynamics of SIGNSVD over MUON and SIGNSGD over ADAM. We also only use the terms theorem, lemma, proposition, etc when indeed these statements are proven. For instance, we do not claim a theorem for the deterministic equivalent for the risk as we do not **prove** the concentration result.

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: We have explicitly stated assumptions labeled within the text itself (e.g., the assumption on the Gaussianity of the data). Statements for results are referenced to the appropriate section in the appendix.

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: We provide a description of the least-squares problem in Section 1.1, the algorithms implemented in Section 1.2, and all hyperparameter choices in figure captions, sufficient to reproduce the synthetic experiments.

   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: Code is not yet publicly released, but we intend to release the code for numerically solving the self-consistent equations for the risk. The synthetic least-squares experiments can be reproduced directly from the descriptions in the paper and figure captions without additional code.

   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: All hyperparameters for the synthetic experiments—including the power law exponents, problem dimensions N_{in}, N_{out}, B, and optimizer settings—are specified in figure captions. Details for the numerical experiments with MUON and momentum on the least squares problem are given in Section D.

   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 report error bars in the synthetic experiments (see e.g., Fig. 4), computed over multiple random seeds; figure captions specify the number of trials and what variability is shown.

   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: [Yes]

   Justification: The synthetic least-squares experiments are inexpensive: each run completes in seconds to a few minutes on a single CPU core; total compute for all figures is negligible. A small number of preliminary runs ($\sim$10) were used to verify the numerical solver before generating the final figures. Res-Net with CIFAR-5M [49] data were also used to generate plots. These are generally inexpensive and standard tests for algorithms.

   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: We have reviewed the code of ethics. We have made our utmost attempt to adhere to the guidelines provided by NeurIPS. We do not use any human subjects nor any datasets. We did our best to cite all the relevent related work. Given that our work is in the foundational research, it is difficult to mitigate all the risks as the downstream effects of theory are long, but we have done our best. The model is completely synthetic using the standard stochastic spectral algorithms; thus we don't, to the best of our knowledge, anticipate any risks. We have included a "Broader Impact" statement at the beginning of the "Supplemental Materials."

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: [Yes]

Justification: The work presented is purely foundational research and is not directly tied to any particular application. We study a simple matrix-valued least squares problem on synthetic dataset and we solve the model using common algorithms such as MUON and SIGNSGD. Given the theoretical nature of this work, we do not anticipate any direct ethical and societal issues. We do state some potential negative and positive societal effects because the implications of our results are on scaling behavior of these commonly used algorithms. See our Broader Impact Statement in the appendix.

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. The majority of the work is theoretical and focuses on the matrix-valued least squares problem using only synthetic data generated from a normal distribution. We do not release any pretrained models.

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: We cite the original papers for MUON [34] and ADAM, and include a detailed comparison with prior work in Section A. We additionally cite datasets such as CIFAR-5M [49] and software [11] when we use them.

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: We do not release new datasets or models. We will release code for solving the deterministic equivalent for the risk; this will be documented in the repository.

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 work does not involve crowdsourcing nor research with human subjects. The data used in this work is generated synthetically.

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 work does not involve crowdsourcing nor research with human subjects. The data used in this work is generated synthetically.

    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 include a declaration of LLM usage at the beginning of the appendix. LLMs were used to assist with code implementation for validating theoretical results (e.g., generating code to solve the deterministic equivalent), as well as to support parts of the proof development and simplify complex expressions. All LLM-generated outputs, particularly proofs, were carefully verified and corrected by the authors to ensure accuracy.

    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.