FREEZE OR DITHER: ROUNDING-AWARE PRECISION FLOORS FOR LOW-PRECISION TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Convergence theory for low-precision training has produced only *upper* bounds, in a worst-case relative-error model blind to the rounding rule. We give the first *lower* bounds for training with floating-point weight storage, organized by one principle: *a coordinate survives round-to-nearest (RN) when its update magnitude either exceeds the local grid spacing u or carries noise at that scale, and freezes or floors when it does neither — freeze or dither.* Sign-type updates can never dither. Below an explicit threshold they freeze bitwise, weight growth stops at a hard "binade wall," and their *movement-weighted* noisy floor is exactly $\sqrt{\pi/2}/2 \cdot u\sigma/L$ under both RN and stochastic rounding (SR), for arbitrary predictable step-size schedules, so at the stationary floor SR provably cannot help sign descent. SGD *self-dithers*: RN's rounding bias collapses as $e^{-2\pi^2(\eta\sigma/u)^2}$, which puts homogeneous-noise SGD–RN on the universal scale. But un-ditherable low-noise coordinates escape the rescue, and under heterogeneous noise with a shared constant step SGD–RN floors at $\Theta\big((u\sigma^2/L^2)^{2/3}\big)$, strictly worse than SGD–SR, so no rounding-blind analysis can be tight for both rules. A representation floor separately forces $M = \Omega(\log T)$ mantissa bits; with the $O(\log T)$ sufficiency of Tang et al. (2026) the requirement is $\Theta(\log T)$. The floors are measured to their constants, and the wall is visible in real training: in a 124M-parameter GPT-2 storing weights in bf16 with no master copy, RN pins *every* LayerNorm gain at exactly its initialization of 1.0 (a binade boundary) in 8 of 9 (optimizer, seed) cells, SR pins none; at the tuned budget the held-out RN/SR ratios order, in every seed, as ditherability predicts, and SR recovers the fp32 master's loss at a third of the weight memory. All code, logs and figures behind every number reported here are in the supplementary material and will be released publicly upon acceptance.

1 INTRODUCTION

Modern large-scale training increasingly stores weights, gradients, and optimizer states in low-precision floating-point formats. When a high-precision master copy of the weights is kept (Micikevicius et al., 2018), quantization is essentially a forward-pass detail; but that copy is the largest item low-precision training has not given up, and practice keeps pushing on it: FP8 for gradients, activations and optimizer states (Peng et al., 2023), then for the parameters and Adam moments at trillion-token scale (Fishman et al., 2024). Storing weights directly in bf16 or FP8 is reported to need compensation, Kahan summation (Zamirai et al., 2020) or stochastic rounding (Liu et al., 2025). What does optimization theory say about this regime?

The state of the art is a collection of *upper* bounds. Tang et al. (2026) analyze Adam and Muon when gradients, weights, and optimizer states are all quantized, under the standard relative-error model $|Q(x) - x| \le q|x|$ with $q = \Theta(2^{-M})$ for M mantissa bits, and show that full-precision rates are preserved if M grows like $\log T$. Ozkara et al. (2025) analyze stochastic rounding (SR) of the weight update. Two questions are left open, the first explicitly so in Tang et al. (2026): (i) in practice the bit-width is *fixed*, so is the resulting "convergence to a neighborhood" an artifact of the analysis, or a real, quantifiable floor? (ii) sign-based optimizers such as signSGD (Bernstein et al., 2018) and Lion (Chen et al., 2023b) are not covered. A third question is implicit in the model itself: the relative-error bound is satisfied by round-to-nearest (RN) *and* by SR with the same q (Lemma 3.1),

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

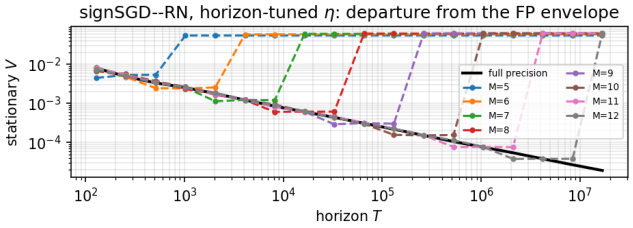

Figure 1: **Fixed mantissa length has a hard horizon.** Each point is an *independent* run at horizon T with the step size tuned to T: final error against T, not a trajectory. It follows the full-precision envelope until a mantissa-dependent $T^\star(M)$ and *equals its initial error* beyond, where the tuned step has fallen under Theorem 4.2's threshold and every update is absorbed bitwise. $\log_2 T^\star$ is linear in M, slope 2.000 over $M \in \{5, \dots, 12\}$: the tuned step falls as $T^{-1/2}$ while the threshold is $\Theta(u)$, so $T^\star \propto u^{-2}$ (Section 8).

so no analysis in that model can distinguish the two rounding rules, yet practitioners consistently report that the choice matters (Gupta et al., 2015; Zamirai et al., 2020; El Arar et al., 2026).

This paper. We answer all three questions with the first lower-bound theory for floating-point weight *storage*. (Memory lower bounds (Marsden et al., 2022) bound what an algorithm can remember, not what a rounding rule does to the iterate it stores.) The organizing principle is *freeze or dither*: a coordinate survives RN storage when its update magnitude either exceeds the local grid spacing or carries noise at that scale, and is frozen or floor-limited when neither holds; no rounding-blind analysis can see this distinction (Corollary 6.7), and once the analysis is made rounding-aware, the floors become exactly characterizable. We prove the two endpoints (ulp-scale noise, and an exactly noiseless coordinate) and *measure* the interpolation: a second coordinate with noise σ_2 moves the exponent from 0.662 (σ_2=0) through 0.671, 0.703, 0.786 to 0.998 at 3×10^{-2}, so $2/3$ is left gradually over a $30\times$ window, not at a knife edge; the sweep resolves the departure from $2/3$, not the approach to 1 (Appendix G.1). Concretely:

(1) $\Theta(\log T)$ mantissa bits, tight (Section 5). A representation floor (Theorem 5.1) forces $M \geq \log_2(LW/4\varepsilon)$ for any algorithm whose iterates live on the grid, where W is the weight scale of the optimum; matching any polynomial full-precision rate therefore requires $M = \Omega(\log T)$, and Tang et al. (2026) prove $O(\log T)$ suffices. The proof is deliberately elementary; its value is the pairing, which closes the fixed-precision question.

(2) Sign descent freezes, exactly (Sections 4, 6 and 7). Sign-type updates have magnitude exactly η, so they can never dither: below the threshold $\eta < |w|2^{-M-2}$ they are absorbed *deterministically, for every noise realization* (Theorem 4.2); no coordinate can grow past a hard *binade wall* at $\Theta(\eta 2^M)$ (Corollary 4.3), verified bitwise on a transformer at three step sizes. Above the threshold, RN sign descent inside a binade *is* an integer lattice chain (Lemma 6.1) and SR sign descent is the same chain on a slowed clock (Lemma 6.2), giving a floor of exactly $C^\star u\sigma/L$, $C^\star = \sqrt{\pi/2}/2 \approx 0.6267$, for *arbitrary predictable step-size schedules* under both rounding rules (Theorem 6.3): *SR provably does not help sign descent*. Sign methods are also exactly invariant to gradient and preconditioner quantization at any $q < 1$ (Proposition 4.4), while momentum accumulation loses a factor $1-\beta$ and cannot decay at all once $\beta > 1 - 2^{-M_m-2}$ (Proposition 7.1).

(3) SGD dithers, conditionally (Section 5). SGD's update ηg_t carries the gradient noise, and noise at ulp scale is a free dither: we prove RN's conditional rounding bias collapses as $\frac{u}{\pi} e^{-2\pi^2(\eta\sigma/u)^2}$ (a Sripad–Snyder-type argument; Theorem 5.4), so homogeneous-noise SGD–RN *reaches* the universal $O(u\sigma/L)$, an upper bound; the matching lower bound at that scale is open (Appendix G.3). The rescue is conditional in two ways. The admissible step-size window collapses to $[\Theta(u/\sigma), \Theta(1/L)]$ and horizon-tuned schedules exit it at $T^\star = \Theta(4^M)$ (Corollary 5.5). And coordinates whose updates are deterministically sub-ulp escape the rescue entirely, proved at the endpoint $\sigma_2 = 0$ with the interpolation in σ_2 open: on a two-coordinate instance with noise $(\sigma, 0)$, SGD–RN floors at $\Theta\big((u\sigma^2/L^2)^{2/3}\big)$, matched, while SGD–SR attains $\Theta(u\sigma/L)$ on the same instance (Theorem 5.6),

Table 1: **Four zero-cost diagnostics, read on the two no-master storage formats.** η_{eff} is the realized update magnitude (Adam: $\eta\hat{m}/(\sqrt{\hat{v}}+\epsilon)$), $u_i = |w_i|2^{-M}$, M_m the momentum mantissa, $\beta_{\text{crit}} = 1 - 2^{-M_m-2}$ the decay-phase (g=0) case. Rows are Thms. 4.2, 5.4, Cor. 4.3, Prop. 7.1; ditherability lists the proved threshold $\rho \geq 1$, while measured floors are reached at $\rho \approx 0.2$.

Diagnostic	Read it as	bf16 (M=7)	FP8-E4M3 (M=3)
freeze margin $\frac{\eta_{\text{eff}}}{\lvert w\rvert 2^{-M-2}}$	< 1: the RN update is dead	dies below $\frac{\lvert w\rvert}{512}$	dies below $\frac{\lvert w\rvert}{32}$
binade wall $\Theta(\eta 2^M)$	RN caps $\lvert w\rvert$ in $(2,4]\cdot 2^M\eta$	256–$512\,\eta$	16–$32\,\eta$
ditherability $\frac{\eta\sigma_i}{u_i}$	$\gtrsim 1$ proved, $\gtrsim 0.2$ measured	$\eta\sigma_i \gtrsim \frac{\lvert w\rvert}{128}$	$\eta\sigma_i \gtrsim \frac{\lvert w\rvert}{8}$
momentum absorption	$\frac{\lvert g-m\rvert}{\lvert m\rvert} < \frac{2^{-M_m-2}}{1-\beta}$ freezes m	$\beta_{\text{crit}} = 0.998$	$\beta_{\text{crit}} = 0.969$

a divergent gap $(\sigma/(Lu))^{1/3}$ that proves the incompleteness of the rounding-blind model (Corollary 6.7).

(4) Experiments (Section 8). Every predicted exponent is confirmed (Figure 2: homogeneous slopes 0.99–1.02 for all four pairs; heterogeneous SGD–RN 0.667 against $2/3$ and SGD–SR 0.998 against 1, on one mantissa window, constants stable to 0.6%); the sign constant $C^\star$ to ± 0.007; the fragility ordering at 124M in every one of three seeds; the $\Theta(\log T)$ horizon with slope 2 for sign descent *and* for SGD via window collapse; the dither-bias collapse over five orders of magnitude; and, on real data, the fragility ordering *adaptive methods most, sign descent least* (Appendix G, Table 2), with an fp32-master control matching full precision throughout.

2 Related work

Convergence under quantization. Early work analyzed SGD with unbiasedly quantized gradients (Alistarh et al., 2017) or repaired biased compression with error feedback (Karimireddy et al., 2019). For quantized *weights*, Li et al. (2017) is the classical reference: on *fixed-point* grids, convex problems, they show SR converges to an $O(\Delta)$ neighborhood while deterministic rounding can stall. Our Lemma 4.1 is the classical *stagnation* (swamping) effect of floating-point accumulation, whose RN-versus-SR contrast is analyzed quantitatively by Higham & Pranesh (2019); Connolly et al. (2021); Croci & Giles (2023); in deep learning Wang et al. (2018) identified it as the binding failure of 8-bit floating-point weight updates and fixed it with chunked accumulation and SR, the empirical counterpart of our freeze branch, in the format Appendix G's Table 2 targets. New here is what the effect implies for *optimization*. Our setting differs from Li et al. (2017) in three ways that matter for modern training: the grid is floating point, so the absorption threshold scales with the weight magnitude, which is the large-weight/small-update regime of LLMs and the source of the binade wall (Corollary 4.3) and the ℓ_1 law (Corollary 6.6), neither of which has a fixed-point analogue; the objects are sign-type methods and SGD side by side; and we prove matching *lower* bounds, a tight $\Theta(\log T)$ mantissa law, and an exact constant. Tang et al. (2026) give the strongest upper bounds for fully quantized Adam/Muon under the relative-error model; our results are complementary: their sufficiency plus our necessity yields tightness, and our mode-separation shows their model cannot be sharpened to distinguish RN from SR. Ozkara et al. (2025) analyze SR upper bounds (Adam with $\beta_1 = 0$) and observe RN stalling for gradient-proportional updates; the lower-bound theory, the exact constants, and the sign-method side, which behaves the opposite way, are new here. Kim (2026) prove lower bounds for *communication*-constrained oracles, explicitly excluding storage precision; our floors arise from the dynamics of storage rounding itself.

Sign methods and low-precision practice. signSGD and its analyses (Bernstein et al., 2018; 2019; Safaryan & Richtárik, 2021) form the template for our upper bounds; Lion (Chen et al., 2023b;a) is the practical face of momentum-sign updates. On the systems side, SR hardware and practice are surveyed by El Arar et al. (2026); Croci et al. (2022); bf16-only training is known to need SR or compensated summation (Zamirai et al., 2020), master weights are standard (Micikevicius et al., 2018), and the FP8 formats we quote are those of Micikevicius et al. (2022), and quantized optimizer states show staleness effects (Dettmers et al., 2022; Li et al., 2023; Topollai & Choromanska, 2026). Error feedback (Richtárik et al., 2021) has recently made master-weight-free training practical at scale (Nikdan et al., 2026), microscaling formats (Rouhani et al., 2023) narrow the element type under a power-of-two block scale (which leaves $\mathbb{F}_M$ invariant, so our criteria apply within a block,

up to the underflow of its small elements), and gradient low-rank projection (Zhao et al., 2024) shrinks optimizer state instead; our floors characterize the *uncompensated* limit, and read as a lower-bound explanation of why *some* compensation (a master copy, SR, bit-centering or averaging (De Sa et al., 2018; Yang et al., 2019), or error feedback) is necessary rather than optional, with closed-form criteria attached.

3 SETUP

Floating-point storage. For mantissa length $M \in \mathbb{N}$, the (unbounded-exponent) grid is $\mathbb{F}_M = \{0\} \cup \{\pm 2^e(1 + k2^{-M}) : e \in \mathbb{Z}, k \in \{0, \dots, 2^M - 1\}\}$. For $x \neq 0$ with $2^e \leq |x| < 2^{e+1}$, the local spacing is $\mathrm{ulp}_M(x) = 2^{e-M}$, so

$$|x|\, 2^{-M-1} < \mathrm{ulp}_M(x) \leq |x|\, 2^{-M}. \tag{1}$$

$\mathrm{RN}_M(y)$ rounds to the nearest grid point (ties to even; our lower bounds use strict inequalities, so tie-breaking is immaterial). $\mathrm{SR}_M(y)$ returns the upper neighbor with probability proportional to the lower offset, so that $\mathbb{E}[\mathrm{SR}_M(y)] = y$. Following Tang et al. (2026) we assume no overflow/underflow (per-tensor scaling in practice); this isolates the mantissa. A per-tensor scale s (storing w/s) cancels from every criterion below, each a ratio of update to weight or to η: absorption reads $|\eta U| < |w|2^{-M-2}$ in either representation and the wall of Corollary 4.3 stays in $(2, 4] \cdot 2^M \eta$ in weight units, so rescaling cannot lift a frozen coordinate.

Lemma 3.1 (basic facts)**.** *(a)* $|\mathrm{RN}_M(y) - y| \leq \mathrm{ulp}_M(y)/2$. *(b)* $\mathbb{E}[\mathrm{SR}_M(y)] = y$, $|\mathrm{SR}_M(y) - y| < \mathrm{ulp}_M(y)$, *and* $\mathrm{Var}[\mathrm{SR}_M(y)] \leq \min\{\mathrm{ulp}_M(y)^2/4,\ \mathrm{ulp}_M(y)\,\mathrm{dist}(y, \mathbb{F}_M)\}$. *(c) Both operators satisfy the relative-error bound* $|Q(y) - y| \leq q|y|$ *with* $q = 2^{-M}$*, i.e., both are admissible instantiations of Assumption 3.1 of Tang et al. (2026) at the same* q*.*

Part (c) is the pivot of the paper: everything we prove separates two instantiations of the *same* assumption.

Quantized-storage training. Weights are stored in $\mathbb{F}_M$; the update is computed in high precision and rounded once on storage (the "no master weights" regime, identical to the formalization of Ozkara et al., 2025):

$$W_{t+1} = Q_M\big(W_t - \eta_t U_t\big) \quad \text{coordinatewise}, \qquad Q_M \in \{\mathrm{RN}_M, \mathrm{SR}_M\}. \tag{2}$$

Optimizers differ by the update direction: signSGD takes $U_t = \mathrm{sign}(g_t)$; SGD takes $U_t = g_t$; momentum-sign (Lion-style) takes $U_t = \mathrm{sign}(m_t)$ with $m_t = Q_{M_m}(\beta m_{t-1} + (1 - \beta)g_t)$, the momentum state stored with M_m mantissa bits; Adam takes $U_t = \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$ with moments in $\mathbb{F}_{M_m}$ and enters through Proposition 7.1 and the experiments, not a floor theorem. Gradients are $g_t = \nabla f(W_t) + \sigma\xi_t$ with $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$; lower bounds use Gaussian noise $\xi_t \sim \mathcal{N}(0, 1)$ or, where boundedness is needed, a truncated Gaussian ($|\xi_t| \leq 3$, density bounded below on $[-1, 1]$). A quantized (FP8) gradient makes ξ_t discrete: sign updates are unaffected (Proposition 4.4); Theorem 5.4 is Gaussian, its discrete analogue open (Appendix G.3).

Hard instances. Lower bounds are proved on one-dimensional quadratics $f_\mu(x) = \frac{L}{2}(x - \mu)^2$ with the optimum at a *mid-cell* point of a binade of scale W (so $\mathrm{dist}(\mu, \mathbb{F}_M) = \mathrm{ulp}/2$), embedded in $\mathbb{R}^d$ when needed, where x_t is the one-dimensional instance of the iterate W_t of equation 2; upper bounds are stated for general smooth f (Appendix E). Throughout, $e_t := x_t - \mu$, $u := 2^{e_W - M} \in [W2^{-M}, 2W2^{-M})$ denotes the relevant spacing, and $V := \liminf_T \frac{1}{T}\sum_{t<T} \mathbb{E}\, e_t^2$ the stationary error. One-dimensional hard instances are standard in lower-bound arguments (Arjevani et al., 2023); Corollary 6.6 gives a genuinely d-dimensional statement.

4 ABSORPTION, FREEZING, AND THE BINADE WALL

Lemma 4.1 (absorption)**.** *Let* $w \in \mathbb{F}_M \setminus \{0\}$ *and* $|\delta| < |w|\, 2^{-M-2}$*. Then* $\mathrm{RN}_M(w + \delta) = w$.

Theorem 4.2 (freeze–or–move dichotomy for sign descent under RN)**.** *Run equation 2 with* $Q_M = \mathrm{RN}_M$*,* $U_t = \mathrm{sign}(g_t)$*, constant* η*, on* f_μ *with* $|x_t| \in [W, 2W)$ *throughout,* W *a power of two (so the spacing is the single value* $u = W2^{-M}$*). If* $\eta < W2^{-M-2}$ *then* $x_t = x_0$ *for all* t*,*

deterministically, for every noise realization and noise law, *so the run is pinned at its* initial *error* e_0^2, *which for a generic start exceeds the representation floor by* $\Theta((e_0/u)^2)$. *If instead* $\eta \geq W2^{-M-2}$ *the iterates move but remain on the grid, so Theorem 5.1 still applies. Either way, for every* η *and* T, $\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 \geq \frac{L^2W^2}{16}4^{-M}$, *a bound that holds for any grid-valued sequence; the content of the dichotomy is the freeze branch, which is far worse.*

The noise-independence is specific to sign-type updates, whose magnitude is exactly η, whereas SGD's update ηg_t occasionally rides a large noise sample across the rounding boundary (Section 5: ulp-scale noise is a *dither* that rescues SGD but never sign updates). Decaying schedules inherit the failure: once η_t anneals below the threshold, training is over (Section 8).

Corollary 4.3 (binade wall)**.** *Under RN, a coordinate updated with magnitudes at most* η *can never grow past the first power of two* 2^E *such that* $\eta < 2^{E-1-M}$*: every attempted crossing rounds back, and from* $w = 2^E$ *the upward gap doubles. Tasks whose loss decrease requires growing weight norm (e.g. cross-entropy logit scaling) stall at* $W_{\text{wall}} = \Theta(\eta\, 2^M)$ *under RN sign descent, while SR crosses in expectation.*

On a transformer trained by sign descent with $M{=}4$ storage, the measured growth ceiling equals the predicted wall *exactly* at all three step sizes tested ($0.125/0.5/1.0$ for $\eta \in \{0.003, 0.01, 0.03\}$), while SR reaches $2.75/3.63/9.5$ on the same runs, 7–22$\times$ past the wall (Appendix G.4).

Proposition 4.4 (free robustness of the sign)**.** *Let* Q *have relative error* $q < 1$. *Then* $\operatorname{sign}(Q(x)) = \operatorname{sign}(x)$ *for all* x*, and sign-descent trajectories are* exactly invariant *to gradient quantization; likewise to any positive per-coordinate rescaling (preconditioner) quantized at* $q < 1$. *By contrast, Tang et al. (2026) require* $q_G = O(1/T)$ *for Adam.*

All proofs for this section are short and given in Appendix A.

5 Floors for SGD: representation, self-dither, and heterogeneity

Theorem 5.1 (representation floor)**.** *For every* $M \geq 1$ *and weight scale* $W > 0$ *there exists* μ *with* $|\mu| \in [W, 2W)$ *such that any sequence* $x_0, x_1, \cdots \in \mathbb{F}_M$ *(any algorithm, any rounding rule) satisfies* $\|\nabla f_\mu(x_t)\| \geq \frac{L}{4}W2^{-M}$ *for all* t*; when* W *is a power of two the constant is* $\frac{1}{2}$.

Corollary 5.2 (necessity of $\Theta(\log T)$ mantissa bits)**.** *Guaranteeing* $\min_t \|\nabla f(x_t)\| \leq \varepsilon$ *on the class* $\{f_\mu : |\mu| \in [W, 2W)\}$ *requires* $M \geq \log_2 \frac{LW}{4\varepsilon}$. *Hence matching a full-precision rate* $\varepsilon(T) = \Theta(T^{-\alpha})$ *requires* $M \geq \alpha \log_2 T - O(1)$. *With the* $M = O(\log T)$ *sufficiency of Tang et al. (2026), the mantissa requirement of quantized training is* $\Theta(\log T)$.

Figure 1 shows the constructive version: with horizon-tuned steps the failure is not a graceful degradation but a jump to the initial error at $T^\star \propto 4^M$, exactly as Theorem 4.2 predicts ($\eta_T \propto T^{-1/2}$ crosses the freeze threshold at $T \propto 4^M$).

Theorem 5.3 (SGD–SR reaches the universal scale; upper bound)**.** *On the instance of Theorem 5.1 with truncated-Gaussian noise and any* $0 < \eta \leq 1/(2L)$*, SGD with SR storage satisfies* $V \leq C\big(\eta\sigma^2/L + u\sigma/L + u^2\big)$*; tuning* η *gives* $V \leq C'(u\sigma/L + u^2)$. *The same holds coordinatewise on separable instances, including noiseless coordinates (which contract to an* $O(u)$*-neighborhood).*

SGD under RN is not a single rate but a dichotomy, governed by whether the update can *dither*.

Theorem 5.4 (self-dithering: homogeneous SGD–RN reaches the universal scale; upper bound)**.** *Let the gradient noise be exactly Gaussian. Inside a binade with spacing* u*, the conditional rounding bias of an SGD–RN step obeys*

$$\big|\mathbb{E}[r_t \mid \mathcal{F}_t]\big| \;\leq\; \frac{u}{\pi}\,\frac{e^{-2\pi^2\rho^2}}{1-e^{-6\pi^2\rho^2}}, \qquad \rho = \frac{\eta\sigma}{u},$$

a Sripad–Snyder-type collapse (Sripad & Snyder, 1977; Gray & Neuhoff, 1998): ulp-scale noise makes RN behave like an unbiased quantizer. Consequently, for $\eta \in [u/\sigma,\, 1/(2L)]$*, homogeneous SGD–RN satisfies* $V \leq C(\eta\sigma^2/L + u\sigma/L)$ *up to an additive bias term* $(b/(\eta L))^2 < 2(\sigma/(\pi L))^2 e^{-4\pi^2}$*, so its tuned error is* $O(u\sigma/L)$*, the universal scale. (We prove the upper bound*

only.) (Proof in Appendix B.5; measured, the bias tracks the bound over five orders of magnitude, and at ρ=0.7 the RN residual variance is $0.0831u^2$ against the uniform $u^2/12$, SR's $2.002\times$ that.)

Corollary 5.5 (window collapse and the SGD horizon)**.** *The rescue is conditional: with bounded noise, every update is absorbed once $\eta(Le_0 + 3\sigma) < u/2$ (within-binade lattice), so SGD–RN's admissible step-size window is $[\Theta(u/\sigma), \Theta(1/L)]$ and shrinks with precision. A horizon-tuned schedule $\eta_T \propto T^{-1/2}$ exits the window at $T^\star = \Theta(4^M)$: the $\Theta(\log T)$ horizon binds SGD–RN through its schedule exactly as it binds sign descent through freezing (measured in Section 8).*

Theorem 5.6 (heterogeneous noise: the $2/3$ floor, and the separation)**.** *Consider $f(x) = \frac{L}{2}\sum_{i=1,2}(x_i - \mu_i)^2$ with both μ_i mid-cell at scale W (spacing u), noise $(\sigma_1, \sigma_2) = (\sigma, 0)$, initial errors $e_{i,0} = \Theta(W)$, a shared constant step $0 < \eta \le 1/(16L)$, and the noise-dominated regime $\sigma \ge cLu$ (used in Leg 2 of the proof; for $\sigma \lesssim Lu$ the conclusion survives via Leg 1 alone, with a worse absolute constant). Under RN storage,*

$$V \;\ge\; c\,\min\Big\{e_{2,0}^2,\; \big(u\sigma^2/L^2\big)^{2/3},\; \sigma^2/L^2\Big\},$$

matched at $\eta = \Theta\big((u^2/(L\sigma^2))^{1/3}\big)$, while SGD–SR attains $C(u\sigma/L + u^2)$ on the same instance (Theorem 5.3): a divergent gap $(\sigma/(Lu))^{1/3}$.

Proof idea (Appendix B.2; the two legs overlap, so the proof has *no* step-size gaps): the noiseless coordinate cannot dither. Its update $\eta L|e_2|$ is deterministic, so it stalls at $e_2^\star = \Theta(u/(\eta L))$, already above the $2/3$ floor for all $\eta \lesssim \eta^\star = (u^2/(L\sigma^2))^{1/3}$; for $\eta \gtrsim \eta^\star$ the exact linear recursion of the noisy coordinate forces $V_1 \gtrsim \eta\sigma^2/L$. The legs meet at $\eta^\star$, and balancing yields the $2/3$ exponent. Experimentally the tuned floor is $0.747\,(u\sigma^2/L^2)^{2/3}$, constant to 0.63% over five mantissa settings, with fitted exponent 0.667 (0.662–0.671 across M-windows; seed-bootstrap CI $[0.664, 0.669]$) (Figure 2, right).

Together, Theorems 5.4 and 5.6 replace the tempting blanket claim "RN hurts SGD" with a sharper one: *RN hurts the coordinates whose update magnitude is deterministically below the grid spacing, and only those.* The theorems establish the endpoints; the σ_2 sweep of Appendix G.1 measures the interpolation (exponent $2/3$ to near 1 over a $30\times$ window, the tuned step migrating from the balance point to the u scale), and a theorem covering that window is open. Where ulp-scale noise is present it is a free dither; where it is absent (noiseless or low-noise coordinates, late-phase adaptive methods whose normalized updates become deterministic, and all sign-type updates) the freezes and amplified floors apply in full. Real gradient noise is strongly heterogeneous across coordinates, which is why the separation, not the rescue, is what real training exhibits (Appendix G, Table 2).

6 Exact characterization for sign descent, and what it implies for SGD

For sign descent the floor is not merely order-tight but *exactly* characterized, uniformly over rounding rules and schedules; two structural lemmas do the work.

Lemma 6.1 (lattice reduction for RN)**.** *Inside a binade with spacing u (away from its endpoints), for any $x \in \mathbb{F}_M$, $s \in \{\pm 1\}$, $\eta \ge 0$ with $\eta/u \notin \mathbb{Z}+\frac{1}{2}$ (exact ties round parity-dependently), and $m := \mathrm{round}(\eta/u)$: $\mathrm{RN}_M(x - \eta s) = x - mus$. RN sign descent inside a binade* is *the integer lattice chain with step mu; every $\eta \in (u/2, 3u/2)$ realizes the* same *Markov chain, and $m = 0$ is the freeze regime of Theorem 4.2.*

Lemma 6.2 (clock invariance for SR)**.** *For $0 < \eta < u$, one SR step moves to the adjacent lattice point (in the direction of the update) with probability η/u and stays put otherwise, independently of the gradient draw. The embedded jump chain is exactly the m=1 lattice chain of Lemma 6.1, and since the holding probability is state-independent, the time-stationary law is* invariant to $\eta \in (0, u)$*: shrinking the step only slows the clock.*

Together: the tuned dynamics of RN and SR sign descent are *the same Markov chain*, the mechanism behind the empirical mode-equivalence of Section 8 (ratios in $[1.001, 1.027]$); step sizes below u buy nothing.

Theorem 6.3 (exact noisy floor; arbitrary predictable schedules; both modes)**.** *On the mid-cell instance with Gaussian noise, run sign descent equation 2 with* any *predictable step-size sequence*

$\{\eta_t\}$ (η_t chosen before observing ξ_t; adaptive schedules allowed) and $Q \in \{\mathrm{RN}_M, \mathrm{SR}_M\}$, iterates in the binade. With movement weights $w_t := \mathrm{round}(\eta_t/u)\,u$ (RN) or $w_t := \eta_t$ (SR), for every T:

$$\frac{\sum_{t<T}\mathbb{E}[w_t e_t^2]}{\sum_{t<T}\mathbb{E}[w_t]} \;\geq\; C^\star\,\frac{u\sigma}{L}\left(1 - \frac{\left(\mathbb{E}e_T^2 - \mathbb{E}e_0^2\right)_+}{u\sum_{t<T}\mathbb{E}[w_t]}\right), \qquad C^\star = \frac{\sqrt{\pi/2}}{2}.$$

The RN branch inherits Lemma 6.1's hypotheses ($\eta_t/u \notin \mathbb{Z}+\frac{1}{2}$, and iterates at distance $> (m_t{+}1)u$ from the binade endpoints); the ratio is read as $0/0$ on a fully frozen run. The correction vanishes whenever the run ends no farther from the optimum than it started, and is $O\big(R_{\mathrm{conf}}^2/(u\sum_t \mathbb{E}[w_t])\big)$ in general for the chain's confinement radius R_{conf}. Moreover the bound is attained: for the $m{=}1$ lattice chain, $V = C^\star\,\frac{u\sigma}{L}\big(1 + O(uL/\sigma)\big)$ (Appendix C.4, via the chain's explicit birth–death stationary law).

Proof sketch. Telescoping $e_{t+1}^2 - e_t^2$ and taking expectations gives $-2\sum_t \mathbb{E}[e_t d_t] = \sum_t \mathbb{E}[d_t^2] + \mathbb{E}e_0^2 - \mathbb{E}e_T^2$ for the displacement d_t. Under RN, $d_t = -m_t u s_t$ (Lemma 6.1), so $-\mathbb{E}[e_t d_t] = u\,\mathbb{E}[m_t e_t(2\Phi(Le_t/\sigma) - 1)] \leq u\sqrt{2/\pi}\,(L/\sigma)\mathbb{E}[m_t e_t^2]$ by the Gaussian sign-bias inequality $0 \leq z(2\Phi(z) - 1) \leq \sqrt{2/\pi}z^2$, while $\mathbb{E}[d_t^2] = u^2\mathbb{E}[m_t^2] \geq u^2\mathbb{E}[m_t]$. Rearranging yields the claim; the SR case replaces exactness by conditional unbiasedness and $\mathbb{E}[d_t^2 \mid \cdot] \geq u\eta_t$. The matching upper bound uses the exact stationary identity $\mathbb{E}[es] = u/2$ for the lattice chain together with the third-order expansion of $2\Phi(z) - 1$ and fourth-moment control. Full proofs in Appendix C. □

Remark 6.4 (unweighted averages)**.** *For constant step sizes the weights are constant, so Theorem 6.3 is an ordinary average-iterate bound; if every step moves and $\max_t w_t/\min_t w_t \leq \rho$, the plain average obeys the floor up to ρ (Appendix C). Dropping the weights is not free: solving the chain exactly, a schedule* that sees *e_t drives the plain average to the representation floor $u^2/4$ by moving to the nearest grid point and freezing, smaller than the weighted floor by $uL/(4C^\star\sigma) \ll 1$. This measures the value of the weighting, not a gap any implementable method realizes: locating that grid point needs μ to within $O(u)$, which a stochastic oracle at noise σ cannot do reliably (Appendix C). A tight final-iterate characterization is open.*

Corollary 6.5 (mode equivalence, and its limits)**.** *For sign descent, RN at any η in the invariance window of Lemma 6.1 ($m = \mathrm{round}(\eta/u) = 1$) and SR at any $\eta < u$ floor at the same $C^\star u\sigma/L$ with the same constant:* SR provably does not improve sign descent*, because gradient noise is already a dither. The constant is the $m{=}1$ one: at $m \geq 2$ the RN chain steps by mu, the stationary identity reads $\mathbb{E}_\pi[e\,(2\Phi(Le/\sigma) - 1)] = mu/2$, and the floor is $C^\star\,m\,u\sigma/L$, so a larger step multiplies the floor rather than lowering it. The reduction needs the sign to see* fresh *noise, so this is a statement about memoryless sign updates: Lion, whose sign reads an EMA, floors at the same scale with RN's constant $3.0\times$ SR's (Appendix G.5). The statement concerns the stationary floor near an interior optimum in the noise-dominated regime $\sigma \gtrsim Lu$; outside it (very coarse precision, or weight-growth-dominated phases) the two modes can deviate in either direction, with RN protected by the wall of Corollary 4.3 (Section 8).*

Corollary 6.6 (d-dimensional ℓ_1 law)**.** *For separable instances with per-coordinate scales W_i (spacings $u_i = \Theta(W_i 2^{-M})$) the identities decouple and sum. With* per-coordinate *steps $\eta_i \in (u_i/2, 3u_i/2)$ the floor is $\gtrsim C^\star\,\frac{\sigma}{L}\,2^{-M}\sum_i W_i$: sign descent's noisy floor scales with the ℓ_1* norm of the weight scales*. A* common *η does not reach it: coordinate i runs at $m_i = \mathrm{round}(\eta/u_i)$ and floors at $\approx C^\star\eta\sigma/L$ whenever $u_i \ll \eta$, independent of W_i, while coordinates with $u_i > 2\eta$ freeze ($m_i = 0$). That gap is the quantitative account of why per-channel scaling helps: one step size cannot sit in the invariance window of a spread-out set of scales.*

Corollary 6.7 (incompleteness of the relative-error model)**.** *There is no function $B(q, L, \sigma, \Delta, T)$ that (i) upper-bounds the tuned stationary error of quantized SGD, on a class containing the heterogeneous-noise instances of Theorem 5.6, for every quantizer with relative error q, and (ii) is tight up to constants for the SR quantizer: by Theorem 5.6, (i) forces $B \geq c(qW\sigma^2/L^2)^{2/3}$ (RN is admissible at $q = 2^{-M}$), which exceeds the SGD–SR rate $C\,qW\sigma/L$ by the divergent factor $(\sigma/(LqW))^{1/3}$. Rounding-aware analysis is necessary.*

7 Momentum: the $1/(1-\beta)$ amplification

Gradient quantization is free for sign methods (Proposition 4.4); momentum *accumulation* is not, because the increment $(1-\beta)(g-m)$ competes with the ulp of the *state*.

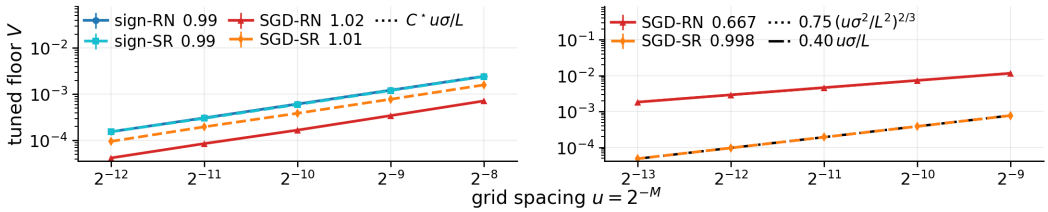

Figure 2: **Freeze or dither, measured** (Appendix G's protocol, 5 seeds, seed-level bootstrap CIs). Left, homogeneous noise: *all four* pairs floor at $\Theta(u\sigma/L)$, every CI containing 1. Right, one noiseless coordinate added, both arms on the *same* window: SGD–RN's 0.667 (0.662–0.671 across M-windows) contains $2/3$; SGD–SR keeps its *homogeneous* constant, so it costs SR nothing.

Proposition 7.1 (momentum absorption; β_{crit})**.** *One RN accumulation step* $m^+ = \text{RN}_{M_m}(\beta m + (1-\beta)g)$ *is absorbed whenever* $(1-\beta)|g-m| < |m|2^{-M_m-2}$*, i.e. the momentum ignores relative gradient changes up to* $2^{-M_m-2}/(1-\beta)$*. In particular, with* $g = 0$ *the decay step* $\text{RN}(\beta m)$ *is absorbed for all* m *once* $\beta > \beta_{\text{crit}} := 1 - 2^{-M_m-2}$*:* the momentum can never decay. *For bf16 states (*$M_m = 7$*),* $\beta_{\text{crit}} \approx 0.998$*:* $\beta = 0.999$ *freezes,* $\beta = 0.99$ *does not.*

A companion bound (Appendix D) shows the accumulated error of quantized momentum is $O\big(q/(1-\beta)\big)$ times an exponentially weighted peak gradient $\tilde{M}_t$, so sign-of-momentum updates are safe iff $q \ll (1-\beta)\,|m_t|/\tilde{M}_t$. This gives closed-form criteria for the staleness reported for 8-bit/4-bit optimizer states (Dettmers et al., 2022; Li et al., 2023; Topollai & Choromanska, 2026). Experiments confirm the absorption band, the $1/(1-\beta)$ scaling (log-log slope 0.99), the decay freeze at β_{crit}, and a Lion-on-MLP degradation specific to RN momentum above β_{crit} (Section 8).

8 Experiments

The synthetic suite runs on CPU with an exact simulator of $\mathbb{F}_M$ (validated bit-for-bit against `ml_dtypes` bfloat16); the transformers use one consumer GPU. Code, seeds, logs and details: supplement and Appendix F.

Floors: universality, the dither valley, and the restored separation (Fig. 2). With homogeneous noise all four pairs floor at $u\sigma/L$ (fitted exponents in Fig. 2, constants in Appendix G). Getting there needs a horizon scaling with $1/(\eta L)$ and an initialization well above the grid, else frozen coarse-M and unconverged fine-M cells *both* bias the fitted exponent below 1 (Appendix G). The mechanism is visible in the SGD–RN step-size landscape (Appendix G, Figure 4): a freeze cliff at the predicted $\eta = u/(2(Le_0 + 3\sigma)) = 0.166u$ (frozen at $0.16u$, moving at $0.22u$), a dither valley on $\eta \in [0.2, 0.7]u$, and the variance ramp $\eta\sigma^2/L$. With 3σ-truncated noise the dither bias falls from $9.8 \times 10^{-2}u$ at $\rho = 0.2$ to $5.2 \times 10^{-4}u$ at $\rho = 0.8$; untruncated it spans five orders (Appendix B.5). On the heterogeneous instance the separation is exact: SGD–RN sits on $0.747\,(u\sigma^2/L^2)^{2/3}$, the constant stable to 0.63%, with $\eta^\star$ on the predicted $(u^2/(L\sigma^2))^{1/3}$ balance point in all five cells; the exponent is 0.6669 on the full window, 0.662–0.671 across three-point M-windows, a drift the seed CI does not cover, reproduced on a non-anchored grid (Appendix G). SGD–SR stays on $0.396\,u\sigma/L$ (exponent 0.998, CI $[0.991, 1.003]$), its *homogeneous* constant 0.39–0.41, so the unditherable coordinate costs SR nothing and RN a divergent factor. Corollary 6.6's ℓ_1 law is exact to 0.5% on a separable 8-coordinate instance, and the price of a *shared* step grows from $0.99\times$ at equal scales to $1.71\times$ over four binades: RN's excess is a cost of step-size sharing, which per-channel scaling buys off (Appendix G.5). Sign-RN and sign-SR straddle the truncation-corrected 0.6250 (0.624–0.634 and 0.613–0.633, per-cell ratios 1.001–1.027), the SR side low by up to 2%, a grid-minimum estimator's downward bias.

A real transformer: bf16 storage, no master weights (Appendix G.6). We test the diagnostics at two scales: a 10.8M GPT on character-level text and a 124M GPT-2-small on BPE-tokenized TinyStories, both storing weights in bf16 with no master copy (the step is literally $w \leftarrow Q(w + \Delta)$, $Q \in \{\text{RN}_7, \text{SR}_7\}$), against an fp32-master control, three seeds each. *The binade wall lands on a*

tensor every transformer has. LayerNorm and RMSNorm gains initialize at exactly 1.0, a binade boundary with bf16 ulp 2^{-7} above and 2^{-8} below, where the tuned update is sub-ulp on *both* sides, so Theorem 4.2 pins them and Corollary 4.3 makes the upward side the harder to escape. It does: at 124M, at the tuned step size over three seeds, RN leaves *all* 25 gain tensors at $\max|w| = 1.0000$ in 8 of the 9 (optimizer, seed) cells and $24/25$ in the ninth, while SR pins *not one gain in any cell* and the master control pins all 25 in none.

The freeze margin measures exposure, not damage. The $\phi < 1$ fraction predicts the fraction whose *stored bits do not change* ($r = 0.991$ and 0.987 at the two scales), but its mean is the same under RN, SR and the master. The damage is realized freezing at matched exposure: RN stops 0.065 more coordinates than $\phi < 1$ accounts for, against 0.021 (SR) and 0.024 (master). *Momentum absorption* isolates cleanly: with a bf16 buffer the frozen fraction of m rises with β_1 ($0.017 \to 0.089$); the fp32-state control stays below 10^{-5}. Proposition 7.1's switch at β_{crit} is a $g{=}0$ statement two β_1 values cannot test (Appendix G.6).

Loss moves much less, but three seeds resolve it. At 124M (25M tokens, under half the split) the held-out RN/SR ratios are 1.0245 ± 0.0011 (Adam), 1.011 ± 0.003 (SGD) and 0.993 ± 0.004 (sign): the first two exceed 1 by 22.7 and 3.4 seed standard deviations, sign's does not clear 1%, and the **ordering holds in every seed**, at the tuned budget that scopes it (Appendix G.6). **SR also recovers the fp32 master** (0.999, 0.995, 1.001 held-out, each within 1.5 seed sd of 1). Repeating the grid on WikiText-103, which memorizes far less, reproduces all of it (Appendix G.6). At 10.8M, also three seeds, the held-out ranking splits on the memorization gap: Adam and sign end $9.7\times$ and $4.4\times$ above train and RN *wins* held-out in $3/3$ seeds, beating the master; SGD ends $1.09\times$ and RN loses in $3/3$. On *training* loss, which is what the floors are about, RN is worse in all nine cells at 10.8M and the six Adam and SGD cells at 124M; sign at 124M sits just below 1, as mode indifference allows. Freezing is a cost the generalization gap may refund.

Exact constant and schedule robustness (Appendix G, Fig. 3). $V/(u\sigma/L)$ matches $C^\star = 0.6267$ to within ± 0.007 in all 36 cells ($M \in [9, 14]$, $\sigma \in \{0.25, 0.5, 1.0\}$, both modes, untruncated noise). The RN window is bitwise: five step sizes spanning $[0.6u, 1.4u]$ give identical trajectories (Lemma 6.1); SR is flat on $\eta \in [u/8, 3u/4]$ (Lemma 6.2). Annealed and *oracle* schedules (seeing e_t) do not beat the floor: movement-weighted error is 1.7–$5.3\times$ Theorem 6.3's bound over four schedules, and a fully informed scheme buys only 0.52–$0.81\times$ in plain average (Remark 6.4).

Necessity (Fig. 1). $T^\star$ quadruples per mantissa bit over $M \in \{5, \dots, 12\}$: slope of $\log_2 T^\star$ vs M is 2.000 on a dyadic horizon grid (the exponent, not its decimals; Appendix G). Corollary 5.5 binds SGD–RN at the same horizon ($T^\star = 4096/16384/65536$ for $M{=}6/7/8$, slope 2.00).

Real data (Appendix G, Table 2). Across a 440-run grid ($M \in \{3..7\} \times 11$ cells $\times$ eight configurations), the *same* ordering appears with far larger separations, as these cells reach stationarity: RN/SR training-loss floor ratios of 20–$322\times$ (Adam), 2.2–$21.9\times$ (SGD) and 0.84–$3.8\times$ (sign), $M{=}3$ being the FP8-E4M3 *mantissa* budget — rerun on the real finite-range codes the SGD gap is $5\times$ smaller, because E4M3's subnormals stop SR's floor improving (Appendix G.5). The largest Adam ratios divide an RN floor by a *task* floor, so they bound the gap rather than measure it.

Beyond the idealizations (Appendix G.5). Real `float8` formats, truncated noise, a clipped exponent and a longer budget are all checked and none overturns a claim; a uniformly placed μ only *worsens* the SGD–RN floor.

9 Discussion

Table 1 collects the four run-time criteria. Gradients and preconditioners are free for sign methods ($q < 1$), weight storage binds, and momentum tolerance shrinks by $1 - \beta$. All are per-tensor, one reduction per step, and acting on them pays. At 124M the diagnostic's top 25 tensors are *exactly* the 25 LayerNorm gains, and compensating those (0.0154% of parameters) recovers much of the RN–SR gap, beating 17–18 of 20 random draws carrying 300–$3300\times$ more parameters (Appendix G.2). We recommend SR, which matches an fp32 master at a third of the memory and error feedback at half (Appendix G.6). **Limitations.** Appendix G.3 details three. The $2/3$ floor needs constant steps, homogeneous SGD–RN has no lower bound, and Adam and Lion have no floor theorem.

REPRODUCIBILITY STATEMENT

Every figure and table is generated from committed JSON logs with fixed seeds, and the simulator is validated bit-for-bit against `ml_dtypes`. The core synthetic suite (`run_all.sh`) is CPU-only and takes 8–14 CPU-hours, dominated by the MLP grid; the horizon-scaled floors of Figure 2 take a few CPU-hours, and the two bf16 transformer grids of Appendix G.6 take about four GPU-hours each on one consumer card, and the seed, horizon, selective-compensation and random-control follow-ups bring the total to about sixteen. Appendix F lists every hyperparameter.

AI USAGE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023. arXiv:1912.02365.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations (ICLR)*, 2019.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As Lyapunov predicts. *arXiv preprint arXiv:2310.05898*, 2023a.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, et al. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv:2302.06675.

Michael P. Connolly, Nicholas J. Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021.

Matteo Croci and Michael B. Giles. Effects of round-to-nearest and stochastic rounding in the numerical solution of the heat equation in low precision. *IMA Journal of Numerical Analysis*, 43 (3):1358–1390, 2023.

Matteo Croci, Massimiliano Fasi, Nicholas J. Higham, Theo Mary, and Mantas Mikaitis. Stochastic rounding: Implementation, error analysis and applications. *Royal Society Open Science*, 9(3): 211631, 2022.

Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R. Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations (ICLR)*, 2022.

El-Mehdi El Arar, Massimiliano Fasi, Silviu-Ioan Filip, and Mantas Mikaitis. What is new in stochastic rounding: a survey on theory, hardware, and applications. *arXiv preprint arXiv:2603.06060*, 2026.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 training to trillion-token LLMs. *arXiv preprint arXiv:2409.12517*, 2024.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning (ICML)*, 2015.

Nicholas J. Higham and Srikara Pranesh. Simulating low precision floating-point arithmetic. *SIAM Journal on Scientific Computing*, 41(5):C585–C602, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signSGD and other gradient compression schemes. In *International Conference on Machine Learning (ICML)*, 2019.

Munsik Kim. Information-theoretic lower bounds for bit-constrained stochastic optimization via a reduction to compressed Gaussian mean estimation. *arXiv preprint arXiv:2606.00703*, 2026.

Bingrui Li, Jianfei Chen, and Jun Zhu. Memory efficient optimizers with 4-bit states. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.02379.

Taowen Liu, Marta Andronic, Deniz Gündüz, and George A. Constantinides. Training with fewer bits: Unlocking edge LLMs training with stochastic rounding. *arXiv preprint arXiv:2511.00874*, 2025.

Annie Marsden, Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Efficient convex optimization requires superlinear memory. In *Conference on Learning Theory (COLT)*, 2022.

Paulius Micikevicius, Sharan Narang, Jonah Alben, et al. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al. FP8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

Mahdi Nikdan, Amir Zandieh, Dan Alistarh, and Vahab Mirrokni. ECO: Quantized training without full-precision master weights. *arXiv preprint arXiv:2601.22101*, 2026.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. arXiv:2502.20566.

Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. FP8-LM: Training FP8 large language models. *arXiv preprint arXiv:2310.18313*, 2023.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.05203.

Bita Darvish Rouhani, Ritchie Zhao, Ankit More, Mathew Hall, Alireza Khodamoradi, Summer Deng, Dhruv Choudhary, Marius Cornea, Eric Dellinger, Kristof Denolf, et al. Microscaling data formats for deep learning. *arXiv preprint arXiv:2310.10537*, 2023.

Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *International Conference on Machine Learning (ICML)*, 2021.

Anekal B. Sripad and Donald L. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5): 442–448, 1977.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.21314.

Kristi Topollai and Anna Choromanska. Understanding quantization of optimizer states in LLM pre-training: Dynamics of state staleness and effectiveness of state resets. *arXiv preprint arXiv:2603.16731*, 2026.

Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, and Kailash Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Christopher De Sa. SWALP: Stochastic weight averaging in low-precision training. In *International Conference on Machine Learning (ICML)*, 2019.

Pedram Zamirai, Jian Zhang, Christopher R. Aberger, and Christopher De Sa. Revisiting BFloat16 training. *arXiv preprint arXiv:2010.06192*, 2020.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03507.

A Proofs for Sections 3 and 4

A.1 Proof of Lemma 3.1

(a) is the definition of rounding to the nearest point of a grid with local spacing $\mathrm{ulp}_M(y)$. (b) With neighbors $\ell = \lfloor y \rfloor_M$, $r = \lceil y \rceil_M$ and $\theta = (y-\ell)/(r-\ell)$, $\mathrm{SR}_M(y) = r$ w.p. θ, so $\mathbb{E}[\mathrm{SR}_M(y)] = \theta r + (1-\theta)\ell = y$ and $\mathrm{Var} = (y-\ell)(r-y)$, which is at most $(r-\ell)^2/4$ and at most $(r-\ell)\min(y-\ell, r-y) = (r-\ell)\,\mathrm{dist}(y, \mathbb{F}_M)$. (c) The error is at most $\mathrm{ulp}_M(y) \le |y| 2^{-M}$ for SR and half that for RN, by equation 1. □

A.2 Proof of Lemma 4.1

The distance from w to its nearest distinct grid point is at least the smaller of the two adjacent spacings. If $|w|$ is not a power of two, both spacings equal $\mathrm{ulp}_M(w) > |w| 2^{-M-1}$ by equation 1. If $|w| = 2^e$, the spacing on the smaller-magnitude side is $2^{e-1-M} = |w| 2^{-M-1}$. In all cases the nearest distinct grid point is at distance $\ge |w| 2^{-M-1} > 2|\delta|$, so $w + \delta$ lies strictly inside the half-gap around w and $\mathrm{RN}_M(w+\delta) = w$. □

A.3 Proof of Theorem 4.2

If $\eta < W2^{-M-2}$ and x_t lies in the binade (so $|x_t| \geq W$), then $|\eta \operatorname{sign}(g_t)| = \eta < |x_t|2^{-M-2}$, and Lemma 4.1 gives $x_{t+1} = x_t$; induction from x_0 freezes the run. The magnitude of a sign update is η for every realization of g_t, so no noise law changes this. If instead $\eta \geq W2^{-M-2}$, Theorem 5.1 applies to all iterates. The displayed bound is the minimum of $L^2 e_0^2$ (freeze case, using $e_0 \geq \frac{1}{2}W2^{-M}$), and the representation floor: W is a power of two here, so $\operatorname{dist}(\mu, \mathbb{F}_M) = \frac{1}{2}W2^{-M}$ and the floor is $\frac{L^2W^2}{4}4^{-M}$, above the displayed $\frac{L^2W^2}{16}4^{-M}$, which is the square of Theorem 5.1's general constant. □

A.4 Proof of Corollary 4.3

Consider the binade $[2^E, 2^{E+1})$ with $\eta < 2^{E-1-M}$, i.e. η below half of that binade's spacing 2^{E-M}. (i) To *enter* the binade from below, an iterate $w < 2^E$ must round to a point $\geq 2^E$; since the spacing below 2^E is 2^{E-1-M} and $\eta < 2^{E-1-M}$, any $w + \eta < 2^E + \eta$ is within half a gap of a point $\leq 2^E$, so the entering step can at most land exactly on 2^E (from the grid point immediately below, when η exceeds half that gap). (ii) From $w = 2^E$, the upward gap is 2^{E-M}, and $\eta < 2^{E-1-M} = \frac{1}{2}\cdot 2^{E-M}$ is below half of it, so $\mathrm{RN}(2^E + \eta) = 2^E$: the coordinate is pinned. Hence no coordinate with bounded update magnitude η ever exceeds 2^E, where 2^E is the smallest power of two with $\eta < 2^{E-1-M}$; this is $\Theta(\eta 2^M)$. Under SR, $\mathbb{E}[\mathrm{SR}(w + \eta)] = w + \eta$: crossing occurs with positive probability at every step and the expected drift is unimpeded. □

A.5 Proof of Proposition 4.4

$|Q(x) - x| \leq q|x| < |x|$ implies $Q(x)$ lies strictly on the same side of zero as x, so $\operatorname{sign}(Q(x)) = \operatorname{sign}(x)$. A sign-descent step depends on the gradient only through $\operatorname{sign}(g)$, hence trajectories coincide for any noise realization. For (positive) preconditioners, $\operatorname{sign}(Q(g)/\sqrt{Q(v)}) = \operatorname{sign}(Q(g)) = \operatorname{sign}(g)$ for $v > 0$, since $Q(v) > 0$ when $q < 1$. □

B Proofs for Section 5

B.1 Proof of Theorem 5.1 and Corollary 5.2

Let $e = \lfloor \log_2 W \rfloor$, so $2^e \leq W < 2^{e+1}$, and let μ be the smallest cell midpoint of $\mathbb{F}_M$ that is $\geq W$. The midpoints of the binade $[2^e, 2^{e+1})$ are $2^e(1 + (k + \frac{1}{2})2^{-M})$, $k < 2^M$, spaced 2^{e-M} apart, and the first midpoint of the next binade, $2^{e+1}(1 + 2^{-M-1})$, lies $\frac{3}{2} \cdot 2^{e-M}$ above the last of them, so $\mu < W + \frac{3}{2} \cdot 2^{e-M} \leq 2W$ (using $2^e \leq W$ and $M \geq 1$): $\mu \in [W, 2W)$. Since μ lies in binade e or $e + 1$, $\operatorname{dist}(\mu, \mathbb{F}_M) \geq 2^{e-M-1} > \frac{1}{4}W2^{-M}$, because $2^e > W/2$; when $W = 2^e$ the first midpoint of binade e itself qualifies and $\operatorname{dist}(\mu, \mathbb{F}_M) = \frac{1}{2}W2^{-M}$. Every grid point is at least this far from μ, and $\|\nabla f_\mu(x)\| = L|x - \mu|$. (The constant $\frac{1}{4}$ is the one squared in Theorem 4.2's bound $\frac{L^2W^2}{16}4^{-M}$.) The corollary follows by solving $\frac{L}{4}W2^{-M} \leq \varepsilon$ for M and inserting $\varepsilon(T) = \Theta(T^{-\alpha})$. □

B.2 Proof of Theorem 5.6 (heterogeneous lower bound)

Coordinates decouple. Write e_1 for the noisy coordinate, $e_{1,t+1} = (1 - \eta L)e_{1,t} - \eta\sigma\xi_t + r_{1,t}$ with $|r_{1,t}| \leq u/2$ and truncated-Gaussian ξ_t ($|\xi_t| \leq 3$, $v_\xi := \mathbb{E}\xi_t^2 \geq c_0$), and e_2 for the noiseless coordinate, $x_{2,t+1} = \mathrm{RN}_M(x_{2,t} - \eta L\, e_{2,t})$.

Leg 1 (the noiseless coordinate; every η). The update magnitude $\eta L|e_{2,t}|$ is deterministic — there is no noise to dither it. By Lemma 4.1 (with $|x_2| \geq W$ and $u \leq 2W2^{-M}$, so $|x_2|2^{-M-2} \geq u/8$), the step is absorbed whenever $\eta L|e_{2,t}| < u/8$. While moving, the pre-rounding target lies within the lattice cell below $x_{2,t}$ (for $\eta L e_2 < u$), so the iterate moves down by 0 or one lattice step: e_2 is non-increasing and decreases by at most u per step. Hence: if $e_{2,0} \leq u/(8\eta L)$ the coordinate is frozen at its initial error; otherwise it descends until the first time $e_{2,t} < u/(8\eta L)$ and stalls there

forever, at a value at least $u/(8\eta L) - u \geq u/(16\eta L)$ (using $\eta \leq 1/(16L)$). In all cases

$$V \;\geq\; \min\Big\{e_{2,0}^2,\; \Big(\frac{u}{16\eta L}\Big)^2\Big\} \qquad \text{for every } \eta.$$

In particular, for $\eta \leq 6u/(v_\xi\sigma)$ this gives $V \geq \min\{e_{2,0}^2,\, (v_\xi\sigma/(96L))^2\} = \Omega(\min\{e_{2,0}^2, \sigma^2/L^2\})$: *the entire dither window is covered by the coordinate that cannot dither.*

Leg 2 (the noisy coordinate; $\eta \geq K\eta^\star$, where $\eta^\star := (u^2/(L\sigma^2))^{1/3}$, $K := (2/v_\xi)^{1/3}$). Since $\sigma \geq cLu$, every $\eta \geq K\eta^\star$ satisfies $\eta\sigma/u \geq K(\sigma/(Lu))^{1/3} \geq 6/v_\xi$, i.e. $u \leq \frac{v_\xi}{6}\eta\sigma$. Cesàro-averaging the exact linear recursion for e_1 exactly as in Theorem 5.3's proof but as a lower bound,

$$2\eta L V_1 + u\sqrt{V_1} \;\geq\; \eta^2\sigma^2 v_\xi - 3u\eta\sigma \;\geq\; \tfrac{v_\xi}{2}\eta^2\sigma^2 \;=: R.$$

At least one term on the left is $\geq R/2$. If $2\eta L V_1 \geq R/2$, then $V_1 \geq R/(4\eta L) = \frac{v_\xi}{8}\,\eta\sigma^2/L$. If $u\sqrt{V_1} \geq R/2$, then $V_1 \geq R^2/(4u^2) = \frac{R}{4\eta L}\cdot\frac{R\eta L}{u^2}$, and $R\eta L/u^2 = \frac{v_\xi}{2}\,\eta^3\sigma^2 L/u^2 \geq 1$ precisely because $\eta \geq K\eta^\star$. Either way $V_1 \geq \frac{v_\xi}{8}\,\eta\sigma^2/L$ on this range. (For $\eta \in [6u/(v_\xi\sigma), K\eta^\star)$ the same split would only give $V_1 \gtrsim \eta^4\sigma^4/u^2$, which can dip below $\eta\sigma^2/L$; the combination never uses Leg 2 there — Leg 1 already exceeds the $2/3$ floor on all of $\eta \leq K\eta^\star$.)

Combination. For every $\eta \leq 1/(16L)$,

$$V \;\geq\; c\,\min\Big\{e_{2,0}^2,\; \max\Big\{\Big(\frac{u}{16\eta L}\Big)^2,\; \frac{\eta\sigma^2}{L}\,\mathbf{1}\big[\eta \geq K\eta^\star\big]\Big\},\; \frac{\sigma^2}{L^2}\Big\}.$$

For $\eta \leq K\eta^\star$ the first branch gives $(u/(16\eta L))^2 \geq (u/(16K\eta^\star L))^2 = c'\,(u\sigma^2/L^2)^{2/3}$; for $\eta \geq K\eta^\star$ the second gives $\eta\sigma^2/L \geq K\eta^\star\sigma^2/L = c''\,(u\sigma^2/L^2)^{2/3}$. The two legs meet at $\eta^\star = \Theta\big((u^2/(L\sigma^2))^{1/3}\big)$ (admissible whenever $\sigma \geq cLu$), and every step size is covered by one of them: no intermediate window escapes. □

B.3 Matching upper bounds on the heterogeneous instance

At $\eta = \Theta((u^2/(L\sigma^2))^{1/3})$, coordinate 2 stalls at $O(u/(\eta L)) = O((u\sigma^2/L^2)^{1/3})$, contributing $O((u\sigma^2/L^2)^{2/3})$, and coordinate 1 contributes $O(\eta\sigma^2/L) = O((u\sigma^2/L^2)^{2/3})$ by the adversarial-rounding upper recursion (bound $|r| \leq u/2$ one-sidedly and solve the fixed point as in the SR proof below). For SGD–SR: coordinate 2 contracts in expectation with $\mathbb{E}[\zeta^2 \mid \cdot] \leq u\eta L|e_2|$, giving $V_2 \leq Cu^2$; coordinate 1 obeys Theorem 5.3; tuned, $V \leq C(u\sigma/L + u^2)$. □

B.4 Proof of Theorem 5.3

With SR, $e_{t+1} = (1-\eta L)e_t - \eta\sigma\xi_t + \zeta_t$, $\mathbb{E}[\zeta_t \mid \mathcal{F}_t, \xi_t] = 0$ and, by Lemma 3.1(b), $\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_t, \xi_t] \leq u\,|\eta g_t| \wedge u^2/4 \leq u\eta(L|e_t| + 3\sigma)$. The cross term vanishes exactly, so

$$\mathbb{E}e_{t+1}^2 \leq (1-\eta L)^2\mathbb{E}e_t^2 + \eta^2\sigma^2 v_\xi + u\eta L\,\mathbb{E}|e_t| + 3u\eta\sigma,$$

and the fixed-point inequality $2\eta L V \leq \eta^2\sigma^2 + u\eta L\sqrt{V} + 3u\eta\sigma$ yields $V \leq C(\eta\sigma^2/L + u\sigma/L + u^2)$. The noiseless-coordinate variant drops the σ terms and gives $V \leq Cu^2$. □

B.5 Proof of Theorem 5.4 and Corollary 5.5

Write the pre-rounding point as $y = c - \eta\sigma\xi$ with $\xi \sim \mathcal{N}(0,1)$ and c deterministic given $\mathcal{F}_t$. Inside a binade the grid is locally the lattice $u\mathbb{Z}$ (shifted), and the rounding residual $s(y) = \mathrm{RN}_M(y) - y$ is the u-periodic sawtooth equal to $-y$ on $(-u/2, u/2)$, with Fourier series $s(y) = \sum_{k\geq 1}\frac{u(-1)^k}{\pi k}\sin(2\pi k y/u)$. Using $\mathbb{E}[\sin(a+bZ)] = \sin(a)\,e^{-b^2/2}$ for $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[s(y) \mid \mathcal{F}_t] = \sum_{k\geq 1}\frac{u(-1)^k}{\pi k}\,\sin\Big(\frac{2\pi k c}{u}\Big)\,e^{-2\pi^2k^2\rho^2}, \qquad \rho = \frac{\eta\sigma}{u},$$

so $|\mathbb{E}[s]| \leq \frac{u}{\pi}\sum_{k\geq 1}e^{-2\pi^2k^2\rho^2} \leq \frac{u}{\pi}\,\frac{e^{-2\pi^2\rho^2}}{1-e^{-6\pi^2\rho^2}}$ (consecutive terms decay by at least $e^{-6\pi^2\rho^2}$). This is the bias bound. For the variance, $\mathrm{Var}[s_t \mid \cdot] \leq u^2/4$ always.

Now the recursion. With $e_t = x_t - \mu$ and $g_t = Le_t + \sigma\xi_t$, $e_{t+1} = (1-\eta L)e_t - \eta\sigma\xi_t + s_t$, so

$$\mathbb{E}[e_{t+1}^2 \mid \mathcal{F}_t] = (1-\eta L)^2 e_t^2 + \eta^2\sigma^2 + \mathbb{E}[s_t^2] + 2(1-\eta L)e_t\mathbb{E}[s_t] \underbrace{-\,2\eta\sigma\,\mathbb{E}[\xi_t s_t]}_{\text{cross term}}.$$

The cross term does *not* vanish (s_t is a deterministic function of $\eta\sigma\xi_t$), and it is exactly the noise-residual correlation that separates RN from SR, so we bound it rather than drop it. Stein's identity $\mathbb{E}[\xi f(\xi)] = \mathbb{E}[f'(\xi)]$ applied termwise to the sawtooth series (legitimate because the partial sums of $\sum_k \sin(kx)/k$ are uniformly bounded and converge to the sawtooth almost everywhere, so dominated convergence, with dominating function $C|\xi|$, passes the expectation through the sum; the resulting series converges absolutely by the Gaussian factor), with $\mathbb{E}[\cos(a+bZ)] = \cos(a)e^{-b^2/2}$, gives

$$\mathbb{E}[\xi_t s_t] = -2u\rho\sum_{k\geq 1}(-1)^k \cos\Big(\frac{2\pi kc}{u}\Big)e^{-2\pi^2k^2\rho^2}, \qquad |2\eta\sigma\,\mathbb{E}[\xi_t s_t]| \;\leq\; 4\eta^2\sigma^2\,\frac{e^{-2\pi^2\rho^2}}{1-e^{-6\pi^2\rho^2}},$$

using $u\rho = \eta\sigma$. In the dither regime the prefactor is tiny: at $\rho = 1/2$ it is below 0.03 and at $\rho \geq 1$ below 10^{-8}, so the cross term is absorbed into the $\eta^2\sigma^2$ variance term at the cost of a factor $(1+o(1))$. (Under SR the same term vanishes identically, since $\mathbb{E}[s_t \mid \mathcal{F}_t, \xi_t] = 0$; this is the one place the two rounding rules enter the recursion differently.) Write b for the bias bound, so $|\mathbb{E}[s_t \mid \mathcal{F}_t]| \leq b$. The remaining cross term is controlled by AM–GM, $2(1-\eta L)|e_t|\,b \leq \eta L\, e_t^2 + b^2/(\eta L)$, which folds it into the contraction: taking expectations and Cesàro-averaging,

$$V \leq \big[(1-\eta L)^2 + \eta L(1-\eta L)\big]V + C\big(\eta^2\sigma^2 + u^2/4\big) + \frac{b^2}{\eta L} = (1-\eta L)\,V + C\big(\eta^2\sigma^2 + u^2/4\big) + \frac{b^2}{\eta L},$$

and dividing by ηL gives $V \leq C\big(\eta\sigma^2/L + u^2/(\eta L) + (b/(\eta L))^2\big)$. At $\eta \geq u/\sigma$ the middle term is $\leq u\sigma/L$ and the bias term is $\leq (\sigma/(\pi L))^2 e^{-4\pi^2}(1-e^{-6\pi^2})^{-2} < 2(\sigma/(\pi L))^2 e^{-4\pi^2}$ (using $\rho \geq 1$ and $u/\eta \leq \sigma$), which is below $u\sigma/L$ whenever $\sigma/(Lu) \leq \frac{1}{2}\pi^2 e^{4\pi^2} \approx 7\times 10^{17}$, i.e. in every regime we consider. Measured collapse (M=12, pure Gaussian, 4×10^6 samples per point): $|b|/u = 1.1\times10^{-1}$, 2.1×10^{-2}, 1.8×10^{-3}, 2.3×10^{-5}, 6.8×10^{-7} at $\rho = 0.2, 0.35, 0.5, 0.65, 0.8$, tracking the bound; beyond $\rho \approx 0.9$ the measurement hits the Monte Carlo noise floor. *The noise model matters, and Corollary 5.5 uses a different one.* Theorem 5.4 is a Gaussian statement: the harmonic collapse $e^{-2\pi^2k^2\rho^2}$ is the Gaussian characteristic function, and no compactly supported noise reproduces it. Truncating at $c\sigma$ perturbs each harmonic by an additive $O(e^{-c^2/2})$ term, which at $c = 3$ is $\approx 10^{-2}$ — four orders of magnitude above the $\sim 10^{-6}u$ the Gaussian bound gives at $\rho = 0.8$, and consistent with the $5.2\times10^{-4}u$ we measure there (Appendix G.5). So under bounded noise the collapse is a floor, not an exponential: the self-dither rescue is *not* uniform in $M \to \infty$ at fixed truncation, and the crossover is at $\sigma/(Lu) \approx (\eta L/b)^2$, beyond our grids but not beyond the theorem's reach. Corollary 5.5 below is stated for bounded noise ($|\xi| \leq 3$), and is therefore not a corollary of Theorem 5.4 in the same noise model; it is a separate absorption argument, and its conclusion runs the other way (bounded noise makes freezing *easier* to certify). Our chain experiments use truncated noise throughout, for consistency with the lower-bound noise class.

Corollary 5.5: with $|\xi| \leq 3$, every update satisfies $|\eta g| \leq \eta(Le_0 + 3\sigma)$; if this is below the within-binade half-gap $u/2$, Lemma 4.1-type absorption freezes the run at x_0 (the conservative binade-edge-safe threshold is $u/8$). Solving $\eta_T(Le_0 + 3\sigma) < u/2$ with $\eta_T = e_0\sqrt{L/(\sigma T)}$ gives $T^\star = \Theta(1/u^2) = \Theta(4^M)$. □

C Proofs for Section 6

C.1 Proof of Lemma 6.1

Inside a binade the grid is the arithmetic lattice $x + u\mathbb{Z}$ (locally, at distance $> (m+1)u$ from the endpoints). The point $x - \eta s$ lies within distance $|\eta - mu| \leq u/2$ of $x - mus$ and at distance $|\eta - ku| > u/2$ from every other lattice point $x - kus$, $k \neq m$; rounding to nearest selects $x - mus$. The window claim: $\mathrm{round}(\eta/u) = 1$ for all $\eta \in (u/2, 3u/2)$, so all such η generate the same transition kernel. □

Three freeze thresholds, and why they differ. Three distinct constants appear, and they are not interchangeable. (i) The *sharp within-binade* threshold is $u/2$: with spacing u on both sides, RN absorbs an update iff $|\eta U| < u/2$, and this is what Lemma 6.2 and $m = \mathrm{round}(\eta/u) = 0$ record. (ii) The *binade-edge-safe* threshold is $|w|2^{-M-2}$, used in Theorem 4.2 and Table 1; inside a binade of spacing u it runs over $[u/4, u/2)$, equal to (i) at the top of the binade and half of it at the bottom, where the downward spacing is $u/2$, the half-gap is $u/4$, and the extra factor of two makes the statement hold uniformly over the binade without tracking position. (Leg 1 of Theorem 5.6's proof uses the cruder $u/8$ because the setup's convention allows u up to $2W2^{-M}$ while $|w| \geq W$.) (iii) The *noise-inclusive* thresholds $\eta(Le_0+3\sigma) < u/2$ (Corollary 5.5), and $\eta L|e_2| < u/8$ (Theorem 5.6, Leg 1) apply (i), and (ii) respectively to the largest update the noise class permits, not to η alone. So (ii) is 1–2× more conservative than (i) by construction (up to $4\times$ under Leg 1's convention), which is also why Table 1's *proved* ditherability criterion ($\rho \gtrsim 1$) sits above the measured onset ($\rho \approx 0.2$): the proved threshold is the uniform one.

C.2 Proof of Lemma 6.2

For $0 < \eta < u$, $y = x - \eta s$ has lattice neighbors x and $x - us$, and $\mathrm{SR}(y) = x - us$ with probability η/u, else x. The probability η/u does not depend on the realized direction s (only the offset magnitude η enters), so the move indicator is independent of the gradient draw, and conditional on moving, the transition is that of the $m{=}1$ lattice chain. Holding times are geometric with state-independent parameter $1 - \eta/u$; a Markov chain subsampled at i.i.d. state-independent holding times has the same stationary law as its jump chain, hence time averages coincide with jump-chain averages and are invariant to $\eta \in (0, u)$. □

C.3 Proof of Theorem 6.3 (lower bound)

Let $d_t := e_{t+1} - e_t$. Telescoping and taking total expectations,

$$-2\sum_{t<T}\mathbb{E}[e_t d_t] \;=\; \sum_{t<T}\mathbb{E}[d_t^2] - \big(\mathbb{E}e_T^2 - \mathbb{E}e_0^2\big) \;\geq\; \sum_{t<T}\mathbb{E}[d_t^2] - \big(\mathbb{E}e_T^2 - \mathbb{E}e_0^2\big)_+. \tag{3}$$

RN. By Lemma 6.1, $d_t = -m_t u s_t$ with $m_t = \mathrm{round}(\eta_t/u)$ ($\mathcal{F}_t$-measurable since η_t is predictable), and $s_t = \mathrm{sign}(Le_t + \sigma\xi_t)$, so $\mathbb{E}[s_t \mid \mathcal{F}_t] = 2\Phi(Le_t/\sigma) - 1$ and, by the Gaussian sign-bias inequality $0 \leq z(2\Phi(z)-1) \leq \sqrt{2/\pi}\, z^2$ (valid for all z, by concavity of Φ on $[0,\infty)$ and $\Phi'(0) = 1/\sqrt{2\pi}$),

$$-\mathbb{E}[e_t d_t] = u\,\mathbb{E}\big[m_t e_t(2\Phi(Le_t/\sigma) - 1)\big] \;\leq\; u\sqrt{2/\pi}\,\tfrac{L}{\sigma}\,\mathbb{E}[m_t e_t^2], \quad \mathbb{E}[d_t^2] = u^2\mathbb{E}[m_t^2] \geq u^2\mathbb{E}[m_t]$$

(m_t integer-valued). Substituting into equation 3 and dividing by $2\sqrt{2/\pi}\,(L/\sigma)\,u\sum_t \mathbb{E}[m_t]$ gives

$$\frac{\sum_t \mathbb{E}[m_t u\, e_t^2]}{\sum_t \mathbb{E}[m_t u]} \;\geq\; \frac{\sqrt{\pi/2}}{2}\,\frac{u\sigma}{L}\Big(1 - \frac{(\mathbb{E}e_T^2 - \mathbb{E}e_0^2)_+}{u\sum_t \mathbb{E}[m_t u]}\Big),$$

which is the claim with $w_t = m_t u$. The correction term vanishes for every run that ends no farther from the optimum than it started ($\mathbb{E}e_T^2 \leq \mathbb{E}e_0^2$), and is in general at most $R_{\mathrm{conf}}^2/(u\sum_t \mathbb{E}[w_t])$ for the confinement radius R_{conf} of the chain, hence tends to zero as the total movement budget grows. *SR.* Conditioning on $(\mathcal{F}_t, \xi_t)$, SR is unbiased: $\mathbb{E}[d_t \mid \mathcal{F}_t, \xi_t] = -\eta_t s_t$, so $-\mathbb{E}[e_t d_t] = \mathbb{E}[\eta_t e_t s_t] \leq \sqrt{2/\pi}(L/\sigma)\mathbb{E}[\eta_t e_t^2]$. Writing $\eta_t = k_t u + \delta_t$ with $k_t \in \mathbb{Z}_{\geq 0}$, $\delta_t \in [0, u)$, a direct computation over the two SR outcomes gives $\mathbb{E}[d_t^2 \mid \mathcal{F}_t, \xi_t] = u^2\big(k_t^2 + (2k_t+1)\delta_t/u\big) \geq u\,\eta_t$ (the difference is $u^2(k_t^2 - k_t + 2k_t\delta_t/u) \geq 0$). The same substitution yields the claim with $w_t = \eta_t$. □

C.4 Proof of Theorem 6.3 (matching upper bound)

For the $m{=}1$ lattice chain, equation 3 holds with equality in stationarity and $d_t = -us_t$, so the exact identity $\mathbb{E}_\pi[e\,(2\Phi(Le/\sigma) - 1)] = u/2$ holds under the stationary law π. We control the bracketing error through the chain's explicit structure.

Lemma C.1 (Gaussian odds decay)**.** *For $z \geq 0$,* $\dfrac{1-\Phi(z)}{\Phi(z)} \leq e^{-3z/2}$.

Proof. Both sides equal 1 at $z = 0$ and $\frac{d}{dz} \ln \frac{1-\Phi}{\Phi} = -\frac{\phi(z)}{\Phi(z)(1-\Phi(z))}$, so it suffices that $\psi(z) := \phi(z) - \frac{3}{2}\Phi(z)(1-\Phi(z)) \geq 0$. $\psi(0) = \phi(0) - \frac{3}{8} \approx 0.024 > 0$, and $\psi'(z) = \phi(z)\big(\frac{3}{2}(2\Phi(z) - 1) - z\big)$, which is ≥ 0 on $[0, z_1]$ and ≤ 0 on $[z_1, \infty)$ for a single crossing $z_1 \in (1, 1.2)$; hence ψ increases on $[0, z_1]$ and decreases after, so on $[0, z_1]$ it is bounded below by $\psi(0) > 0$. On the decreasing part we need the value at the right end of any interval we cover: for $z \in [z_1, 2]$, $\psi(z) \geq \psi(2) = \phi(2) - \frac{3}{2}\Phi(2)(1-\Phi(2)) \approx 0.0540 - \frac{3}{2}(0.9772)(0.0228) \approx 0.0206 > 0$. For $z > 2$, $1 - \Phi(z) \leq \phi(z)/z$ gives $\psi(z) \geq \phi(z)(1 - \frac{3}{2z}) \geq \frac{1}{4}\phi(z) > 0$. The two ranges overlap at $z = 2$, so $\psi > 0$ throughout. (Checking $\psi(z_1)$ alone would not suffice: z_1 is the *maximum* of ψ and says nothing about the decreasing branch, which is why the two ranges are treated separately.) □

Lemma C.2 (explicit stationary law; sub-Gaussian tails; fourth moment)**.** *The m=1 chain is a reversible birth–death chain on the offset lattice $e_k = (k + \frac{1}{2})u$, $k \in \mathbb{Z}$, with (for $k \geq 0$) inward probability $\Phi(z_k)$, $z_k = Le_k/\sigma$. Detailed balance gives $\pi(k{+}1)/\pi(k) = \frac{1-\Phi(z_k)}{\Phi(z_{k+1})} \leq \frac{1-\Phi(z_k)}{\Phi(z_k)} \leq e^{-3z_k/2}$ by Lemma C.1, hence*

$$\pi(k) \;\leq\; \pi(0)\,\exp\Big(-\tfrac{3}{2}\sum_{j=0}^{k-1} z_j\Big) \;=\; \pi(0)\,\exp\Big(-\tfrac{3Lu}{4\sigma}\,k^2\Big),$$

and symmetrically for $k < 0$: Gaussian tails at scale $k^\star = \Theta(\sqrt{\sigma/(Lu)})$, i.e., e-scale $\Theta(\sqrt{u\sigma/L})$.

A matching lower bound *on the same ratio is what converts this into a fourth-moment bound, and it is needed: the tail estimate alone controls only $\pi(k)/\pi(0)$, and with the trivial $\pi(0) \leq 1$ it gives $\mathbb{E}_\pi e^4 \leq Cu^4\theta^{-5/2} = C(u\sigma/L)^2\sqrt{\sigma/(Lu)}$, larger than claimed by the divergent factor $\sqrt{\sigma/(Lu)}$. Write $\theta := Lu/\sigma$, so $z_k = \theta(k + \frac{1}{2})$. Since $1 - \Phi$ is decreasing, $\pi(k{+}1)/\pi(k) \geq (1 - \Phi(z_{k+1}))/\Phi(z_{k+1})$, and on $z \in [0, 1]$ the odds ratio satisfies $(1 - \Phi(z))/\Phi(z) \geq e^{-2z}$ (both sides are 1 at $z = 0$, and $-\frac{d}{dz} \ln \frac{1-\Phi}{\Phi} = \frac{\phi}{\Phi(1-\Phi)} \leq \frac{\phi(1)}{\Phi(1)(1-\Phi(1))} \approx 1.82 < 2$ there, the ratio being increasing in z). Hence for $1 \leq k \leq K := \lfloor \theta^{-1/2} \rfloor$, a range on which $z_K \leq \theta^{1/2} + \theta/2 \leq 1$ whenever $\theta \leq \frac{1}{2}$, which we assume from here on (it is the noise-dominated regime $\sigma \gtrsim Lu$) —*

$$\frac{\pi(k)}{\pi(0)} \;\geq\; \exp\Big(-2\theta\sum_{j=1}^{k}(j + \tfrac{1}{2})\Big) \;=\; e^{-\theta k(k+2)} \;\geq\; e^{-3\theta k^2}.$$

Summing over $|k| \leq K$ and using $\sum_k \pi(k) = 1$ and $\theta K^2 \leq 1$, $1 \geq \pi(0)(2K+1)e^{-3\theta K^2} \geq e^{-3}\pi(0)\theta^{-1/2}$, since $2K + 1 > 2\theta^{-1/2} - 1 \geq \theta^{-1/2}$ for $\theta^{-1/2} \geq 1$ (the tempting $2K + 1 \geq 2\theta^{-1/2}$ fails whenever the fractional part of $\theta^{-1/2}$ exceeds $\frac{1}{2}$), so

$$\pi(0) \;\leq\; e^3\sqrt{Lu/\sigma}.$$

Feeding this into the upper tail, $\mathbb{E}_\pi e^4 \leq \pi(0)u^4\sum_k (k + \frac{1}{2})^4 e^{-\frac{3}{4}\theta k^2} \leq C\pi(0)u^4\theta^{-5/2} \leq C'u^4\theta^{-2} = C'(u\sigma/L)^2$, as claimed.

Now bracket the sign bias with $\sqrt{2/\pi}\,(z - z^3/6) \leq 2\Phi(z) - 1 \leq \sqrt{2/\pi}\,z$:

$$\sqrt{2/\pi}\,\tfrac{L}{\sigma}\Big(\mathbb{E}_\pi e^2 - \tfrac{L^2}{6\sigma^2}\mathbb{E}_\pi e^4\Big) \;\leq\; \tfrac{u}{2} \;\leq\; \sqrt{2/\pi}\,\tfrac{L}{\sigma}\,\mathbb{E}_\pi e^2.$$

By Lemma C.2 the subtracted term is at most $\frac{L^2}{6\sigma^2} \cdot C(u\sigma/L)^2 = O(u^2)$, i.e., a relative $O(uL/\sigma)$ correction to the leading term $\mathbb{E}_\pi e^2 = \Theta(u\sigma/L)$. Hence $V = C^\star\,\frac{u\sigma}{L}\,\big(1 + O(uL/\sigma)\big)$. Numerically the constant matches $C^\star$ to ± 0.007 over the whole grid and to ± 0.008 against the truncation-corrected target at $M \geq 9$ (Figure 3); the correction itself sits below seed noise at these scales, consistent with its $O(uL/\sigma)$ magnitude (Appendix G). □

C.5 Unweighted averages (Remark 6.4)

If every step moves and $\max_t w_t / \min_t w_t \leq \rho$, then

$$\frac{1}{T}\sum_{t<T}\mathbb{E}e_t^2 \;\geq\; \frac{1}{T\,w_{\max}}\sum_{t<T}\mathbb{E}[w_t e_t^2] \;\geq\; \frac{\sum_{t<T}\mathbb{E}w_t}{T\,w_{\max}} \cdot \mathrm{floor} \;\geq\; \frac{1}{\rho} \cdot \mathrm{floor},$$

where floor denotes the right-hand side of Theorem 6.3. For constant step sizes $\rho = 1$ and the two averages coincide.

C.6 Final-iterate caveat (Corollary 6.5)

The movement-weighted average is the right object for a schedule-robust statement: a scheme may move early and then freeze deliberately at a lucky grid point, making the *final* error as small as the representation floor $u/2$, but no smaller, by Theorem 5.1; and locating a lucky point requires knowing μ to within $O(u)$, which a stochastic oracle at noise σ cannot do reliably in the noise-dominated regime without paying the floor along the way. We leave a tight final-iterate statement open.

C.7 Proof of Corollary 6.6

For separable $f(x) = \sum_i \frac{L}{2}(x_i - \mu_i)^2$ the update factorizes coordinatewise, with shared η_t: each coordinate is a one-dimensional chain on its own lattice (spacing u_i), and equation 3 holds per coordinate. Summing the per-coordinate bounds with weights $w_{t,i} = \mathrm{round}(\eta_t/u_i)u_i$ (RN), and lower-bounding each by Theorem 6.3 yields the ℓ_1 law. □

C.8 Proof of Corollary 6.7

Suppose such a B exists. RN with mantissa M has relative error $q = 2^{-M}$ (Lemma 3.1(c)), so (i), and Theorem 5.6 force $B(q, \dots) \geq c\,(qW\sigma^2/L^2)^{2/3}$ (taking $u = \Theta(qW)$). But SGD–SR, also admissible at the same q, attains $C\,qW\sigma/L$ by Theorem 5.3, and $(qW\sigma^2/L^2)^{2/3}/(qW\sigma/L) = (\sigma/(LqW))^{1/3} \to \infty$ as $q \to 0$ for fixed σ/L, contradicting (ii). □

D Momentum: proofs and quantitative staleness

D.1 Proof of Proposition 7.1

$\beta m + (1-\beta)g = m + (1-\beta)(g-m)$; apply Lemma 4.1 with $\delta = (1-\beta)(g-m)$ (the lemma needs $m \neq 0$; the state $m = 0$ is trivially fixed by the decay step). With $g = 0$ the condition reads $(1-\beta)|m| < |m|2^{-M_m-2}$, i.e. $\beta > 1 - 2^{-M_m-2}$, independent of $m \neq 0$: the decay chain $m \mapsto \mathrm{RN}(\beta m)$ is frozen at every state. □

D.2 Accumulated error and sign safety

Proposition D.1. *Let $q \leq (1-\beta)/2$ and $\hat{m}_t = (1+\delta_t)(\beta\hat{m}_{t-1} + (1-\beta)g_t)$ with $|\delta_t| \leq q$. Define the half-rate weighted peak gradient $\tilde{M}_t := \sup_{0\leq j\leq t} e^{-j(1-\beta)/4}\,|g_{t-j}|$. Then*

$$|\hat{m}_t - m_t| \;\leq\; \frac{C\,q}{1-\beta}\,\tilde{M}_t \qquad \textit{for an absolute constant } C \;(C \leq 64\sqrt{e} \textit{ suffices}),$$

and consequently $\mathrm{sign}(\hat{m}_t) = \mathrm{sign}(m_t)$ *whenever $\tilde{\kappa}_t\, Cq/(1-\beta) < 1$, where $\tilde{\kappa}_t := \tilde{M}_t/|m_t|$.*

Proof. The coefficient of $\beta^j(1-\beta)g_{t-j}$ in $\hat{m}_t$ is $\prod_{i=t-j}^{t}(1+\delta_i)$, and $\left|\prod_i(1+\delta_i) - 1\right| \leq (j+1)q\,e^{(j+1)q}$. Since $q \leq (1-\beta)/2 \leq \frac{1}{2}\ln(1/\beta)$, we have $e^{jq}\beta^j \leq e^{-j(1-\beta)/2}$, hence

$$|\hat{m}_t - m_t| \;\leq\; q\,e^q\,(1-\beta)\sum_{j\geq 0}(j+1)\,e^{-j(1-\beta)/2}\,|g_{t-j}| \;\leq\; q\,e^q\,\tilde{M}_t\,(1-\beta)\sum_{j\geq 0}(j+1)e^{-j(1-\beta)/4},$$

splitting $e^{-j(1-\beta)/2} = e^{-j(1-\beta)/4}\cdot e^{-j(1-\beta)/4}$ and absorbing one factor into $\tilde{M}_t$. Finally $\sum_{j\geq 0}(j+1)x^j = (1-x)^{-2}$ with $x = e^{-(1-\beta)/4}$ and $1 - x \geq (1-\beta)/8$ give $(1-\beta)\sum_j(j+1)e^{-j(1-\beta)/4} \leq 64/(1-\beta)$, while $e^q \leq \sqrt{e}$.

Why the half-rate window is necessary. The natural-looking bound with the full-rate mass $(1-\beta)\sum_k \beta^{t-k}|g_k|$ is *not* available: per-step multiplicative errors compound, so the error contribution of the gradient at lag j is attenuated only at the half rate $e^{-j(1-\beta)/2}$ rather than β^j, an early large gradient can contribute more error than its momentum weight suggests. The half-rate peak $\tilde{M}_t$ is the correct notion; the $1/(1-\beta)$ amplification, which is the message of this section, is unchanged. □

E General d-dimensional upper bounds

Theorem E.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth with $f(x_0) - f^* \le \Delta$, per-coordinate noise with a sign-probability margin $\Pr[\mathrm{sign}(g_i) = \mathrm{sign}(\nabla_i f)] - \frac{1}{2} \ge c \min\{1, |\nabla_i f|/\sigma\}$ (Safaryan & Richtárik, 2021), and let $u_{\max} := \sup_{t,i} \mathrm{ulp}_M(x_{t,i}) \le W_\infty 2^{-M}$ under a weight bound $\|x_t\|_\infty \le W_\infty$. Then for sign descent with SR storage and any $\eta > 0$,*

$$\frac{1}{T} \sum_{t<T} \mathbb{E}\Big[\sum_i \min\Big\{ |\nabla_i f(x_t)|, \tfrac{|\nabla_i f(x_t)|^2}{\sigma} \Big\} \Big] \;\le\; C\Big[\frac{\Delta}{\eta T} + Ld\,(\eta + u_{\max}) \Big];$$

with RN storage and $\eta \ge 2C u_{\max}$ the same bound holds with an additional $C \frac{u_{\max}}{\eta} \mathbb{E}\|\nabla f\|_1$ term, which is absorbable into the left side only on coordinates with $|\nabla_i f| \gtrsim \sigma u_{\max}/\eta$; below that gradient scale it is not, and no bound of this form is available. Coordinates with $\eta < |x_{t,i}| 2^{-M-2}$ freeze outright as in Theorem 4.2.

Proof. Smoothness gives $f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2$ with $x_{t+1} - x_t = -\eta s_t + \varrho_t$. For SR, $\mathbb{E}[\varrho_t \mid x_t, s_t] = 0$, so both cross terms vanish, $\mathbb{E}\langle \nabla f, \varrho_t \rangle = 0 = \mathbb{E}\langle s_t, \varrho_t \rangle$, and $\mathbb{E}\|x_{t+1} - x_t\|^2 = d\eta^2 + \mathbb{E}\|\varrho_t\|^2$ exactly (as $\|s_t\|^2 = d$), with $\mathbb{E}\|\varrho_t\|^2 \le d\, u_{\max}\eta$; the margin condition gives $\mathbb{E}[s_{t,i}\mathrm{sign}(\nabla_i f)] = 2\Pr[\mathrm{sign}(g_i) = \mathrm{sign}(\nabla_i f)] - 1 \ge 2c \min\{1, |\nabla_i f|/\sigma\}$, hence $\mathbb{E}\langle \nabla f, -\eta s_t \rangle \le -2c\eta \sum_i \min\{|\nabla_i f|, |\nabla_i f|^2/\sigma\}$. Telescoping proves the SR bound. For RN, bound $\langle \nabla f, \varrho_t \rangle \le \|\nabla f\|_1 u_{\max}/2$ adversarially and use $\|x_{t+1} - x_t\|^2 \le 2d\eta^2 + 2\|\varrho_t\|^2$.

Why the RN branch stops at a gradient scale. On the left side the i-th term is $\min\{|\nabla_i f|, |\nabla_i f|^2/\sigma\}$. Absorbing $C\frac{u_{\max}}{\eta}|\nabla_i f|$ into a constant fraction of it needs $u_{\max}/\eta \lesssim 1$ when $|\nabla_i f| \ge \sigma$ (which is what $\eta \ge 2C u_{\max}$ delivers), but needs $|\nabla_i f| \gtrsim \sigma u_{\max}/\eta$ when $|\nabla_i f| < \sigma$, and that fails as $|\nabla_i f| \to 0$: exactly the near-stationary regime the theorem is supposed to measure. This is not a slack in the proof. It is the d-dimensional shadow of the one-dimensional floors: SR's residual is conditionally mean-zero and drops out of the descent inequality, RN's does not, and the gradient scale $\sigma u_{\max}/\eta$ at which the RN bound expires is the same $u\sigma/\eta$ scale that Theorems 5.4 and 5.6 localize. We therefore do *not* claim a clean d-dimensional upper bound for RN; the SR branch above is the clean one, and the contrast is the point.

□

F Experimental details

Simulator. $\mathrm{RN}_M/\mathrm{SR}_M$ are implemented by mantissa rounding after a `frexp` decomposition (float64 carrier; exact for $M \le 51$), vectorized in NumPy. Validation (`exp31_simulator_validation.py`): 4×10^5 random floats agree bit-for-bit with `ml_dtypes.bfloat16` at $M = 7$ on f32-representable inputs — 0 mismatches. On raw f64 inputs the rate is $4/4 \times 10^5 = 1.0 \times 10^{-5}$, and every one of the four disagreements has the signature of a double rounding: the gap is exactly one bf16 ulp, the `f64→f32` step had already moved the value in $4/4$, and ours is the strictly closer (correctly rounded) result in $4/4$. `ml_dtypes` rounds f64→f32→bf16; we round once, directly. SR unbiasedness is verified to CLT accuracy.

Synthetic floors (Figs. 2, 3). Instance $f_\mu(x) = \frac{1}{2}(x-\mu)^2$ with $\mu = 1.5 + 2^{-M-1}$ (mid-cell, binade $[1, 2)$), truncated-Gaussian noise. Figure 2 (homogeneous, published protocol): σ=1, $x_0 = \mathrm{RN}(\mu + 0.2)$, $M \in \{8, \dots, 12\}$, η-grids spanning the dither valley (SGD–RN: $\{0.2, 0.3, 0.45, 0.7, 1.2\}u$), and *per-cell* horizon $T = 60\,\tau/(1 - \text{burn}) = 120\,\tau$ with $\tau = 1/(\eta_{\min} L)$ for SGD and $\tau = e_0/\eta_{\min}$ for sign, floored at 1.2×10^5 steps, so T ranges from 1.2×10^5 at M=8 to 2.5×10^6 at M=12 and every cell gets 120–1406 contraction times, 48 replicas $\times 5$ seeds, with each seed's tuned floor stored so the slope interval can be a seed-level bootstrap. (A fixed $T = 1.5 \times 10^5$ with $x_0 = \mathrm{RN}(\mu + 0.02)$ biases the fitted exponent downward on both counts; Appendix G quantifies each, which is why the protocol scales the horizon per cell and starts well above the grid.) Heterogeneous panel: one window $M \in \{9, \dots, 13\}$ for *both* arms, second (noiseless) coordinate started at $\mu + 0.2$ so that the $e_{2,0}^2$ branch of Theorem 5.6 stays inactive, per-cell $T = 240\,\tau$ for the *noisy* coordinate (floored at 1.5×10^5), 40 replicas $\times 5$ seeds, and each arm tuned over a grid spanning its own valley, the $(u^2/(L\sigma^2))^{1/3}$ balance scale for RN, $\eta/u \in [0.12, 0.75]$ for SR. Putting both arms on the RN grid

instead measures the noiseless coordinate's stall rather than SR's floor; Appendix G.5 documents what that costs. Figure 3 (sign side): $x_0 = \mathrm{RN}(\mu + 0.1)$, $\eta = u$, T up to 4×10^5, 512 replicas, burn-in 33%, 5 seeds. In all cases the tuned floor is the grid-minimum of the time-averaged e^2, which is downward-biased by the selection (visible as sign-SR landing up to 2% below $C^\star$), so we do not read the tuned constants at more than two decimals. Every slope CI in Figures 2 and 3 is seed-level: a bootstrap over the per-seed tuned floors, resampling seeds within each mantissa cell. We report these rather than the OLS-on-means intervals because the latter treat the cell mean as the replication unit and so ignore the seed variance that actually limits the fit; on the superseded dense heterogeneous RN run the OLS interval was $2.2\times$ narrower than a seed bootstrap of the same rows and excluded $2/3$, which the bootstrap contains. The two are not uniformly ordered: on the old heterogeneous SR arm the OLS interval was $2.7\times$ *wider*, and on the heterogeneous RN arm we actually report the OLS interval is $1.9\times$ wider than the bootstrap ($[0.6623, 0.6715]$ against $[0.6644, 0.6693]$), and contains $2/3$ as well. The point is that they answer different questions, not that either is always tighter; we quote the bootstrap because the seed is the replication unit. Freeze checks run 5×10^4 steps at $0.9\times$ and $4.1\times$ the threshold (64 replicas; $0/64$ vs $64/64$ moved). Necessity (Fig. 1): $\eta_T = e_0/\sqrt{\sigma T/L}$, $M \in \{5, \ldots, 12\}$, $T \le 2^{24}$, 64 replicas; $T^\star$ is the first T with error $> 3\times$ the full-precision envelope.

Real-data grid (Table 2). MLPs on `digits`: depth $\in \{2, 3\}$ $\times$ width $\in \{32, 128\}$. The reference architecture, used for Table 2 and for the other four datasets, is depth 2, width 32 (64–32–10, ReLU); the other three `digits` architectures enter the 440-run grid but not the table, whose rows are the reference cell alone. Weight storage quantized, gradients and activations in float64 (isolating storage rounding per Proposition 4.4); optimizers sign/SGD/Adam $\times$ storage fp/master/RN/SR; learning-rate grid per cell, 5 seeds, tail-averaged training loss. Datasets: `digits` (full), MNIST (1%/10%/100%), and California housing (regression), five in all. Preregistered exclusions: cells where all modes reach the task floor (ratio of near-zeros; e.g. sign on MNIST-1%), and cells where all modes including full precision diverge (SGD on unstandardized California housing). The fp32-master control matches fp64 in every reported cell.

Transformer. Two-layer, single-head attention + ReLU MLP blocks, learned token/position embeddings, untied unembedding; manual backpropagation validated by central finite differences (worst relative error 8.5×10^{-7} at $h = 10^{-5}$; $h = 10^{-6}$ hits FD truncation noise). Task: fixed-offset induction (predict a repeated prefix), 5% label noise for the main grid (interior optimum), noiseless variant for the wall experiment. $M \in \{4, 5, 6\}$ storage, learning-rate grid, multiple seeds.

Momentum. Absorption thresholds scanned over ρ for grid-point states, $\beta \in \{0.9, 0.99, 0.999\} \times M_m \in \{5, 7, 9\}$; decay-freeze runs 5000 steps from $m = 1$; Lion-style MLP with quantized momentum states, $\beta_2 \in \{0.99, 0.999\}$.

G Additional results and protocol notes

Slopes, and how a protocol can manufacture one. On the *homogeneous* instance, a tuning grid anchored at the $2/3$-balance point produces an apparent SGD–RN slope of $0.71\,[0.69, 0.73]$ over $M \in [5, 14]$, a spurious exponent created entirely by the grid missing the dither valley; grids spanning the valley give slope $1.020\,[0.999, 1.039]$ with floors at 0.17–$0.18\,u\sigma/L$ once the horizon is scaled (below), i.e., the universal scale of Theorem 5.4. The true $2/3$ exponent lives on the *heterogeneous* instance, where valley-spanning grids on a single mantissa window yield slope $0.6669\,[0.6644, 0.6693]$ with the constant 0.747 stable to 0.63%, against SGD–SR's $0.998\,[0.991, 1.003]$ on the same window (Figure 2, right). The sign-side slopes sit at 1 within seed error ($0.994\,[0.983, 1.007]$ for RN, $0.993\,[0.979, 1.005]$ for SR); the positive $O(uL/\sigma)$ correction of Theorem 6.3 is below seed noise everywhere on this grid (Figure 3). The three intervals just quoted, and the fourth homogeneous arm (SGD–SR, $1.009\,[1.000, 1.019]$), are seed-level bootstraps over the per-seed tuned floors (`exp9e_horizon_seeds.json`), which is why they can be quoted alongside the heterogeneous ones. The 0.71 above is the superseded anchored-grid protocol and is an OLS-on-means interval; it is quoted to name the artifact, not as a measurement.

Constant-rate fit. The residuals $|V/(u\sigma/L) - C^\star|$ are at the seed-noise level ($\sim 10^{-3}$) throughout $M \in [9, 14]$: consistent with the predicted $O(uL/\sigma)$ correction magnitude ($\lesssim 10^{-3}$ over this range),

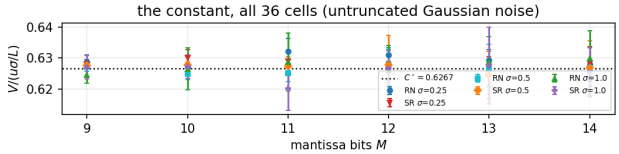

Figure 3: **The exact constant.** $V/(u\sigma/L)$ for $M \in \{9, \ldots, 14\}$, $\sigma \in \{0.25, 0.5, 1.0\}$, both rounding modes, 5 seeds each, against $C^\star = \sqrt{\pi/2}/2$ (dotted): all 36 cells within ± 0.007. These runs use untruncated Gaussian noise, the hypothesis Theorem 6.3 is stated under, so the target is $C^\star$ itself with no truncation correction. We claim the constant, not the $O(uL/\sigma)$ correction rate, which is below the seed spread in 31 of the 36 cells (see the correction-rate discussion in this appendix).

Table 2: **Who needs stochastic rounding, on real data.** Ratio of tuned *training-loss* floors, RN over SR, for a 2-layer MLP on `digits` (converged regime; best of a step-size grid, 5 seeds; an fp32 master-weight control, run for sign and SGD, matches full precision in every cell). $M{=}3$ is the FP8-E4M3 *mantissa* budget; the real finite-range codes give a $5\times$ smaller SGD ratio (Appendix G.5). Exclusions, per-cell seed spreads (within $\pm 35\%$ except †, where the Adam–SR *denominator* is pre-asymptotic (read it as "very large," not 202), and Adam–RN at $M{=}7$), and the full 440-run grid: Appendix G.

RN/SR floor ratio	$M{=}3$	$M{=}4$	$M{=}5$	$M{=}6$	$M{=}7$ (bf16)
Adam	$202\times^\dagger$	$322\times$	$107\times$	$23.0\times$	$20.1\times$
SGD	$21.9\times$	$12.3\times$	$5.8\times$	$3.5\times$	$2.2\times$
signSGD	$3.8\times$	$1.8\times$	$0.95\times$	$1.15\times$	$0.84\times$

but leaving the *rate* unresolvable at this budget (fitted slope 0.22, CI $[-0.10, 0.54]$); we claim the constant, not the rate.

Coarse-precision deviations from mode equivalence. Mode equivalence (Corollary 6.5) is a statement about the stationary floor near an interior optimum with $\sigma \gtrsim Lu$. Outside this regime we observe deviations in *both* directions, as the theory allows: on the noiseless logit-growth transformer task, sign-SR underperforms sign-RN (SR's unbiased escape drives weights into coarser binades whose larger ulp feeds back as update noise, while RN's wall caps the scale, an implicit regularizer); on the coarsest MLP cells ($M = 3$), sign-RN is worse ($3.8\times$), the freeze branch binding. Both deviations vanish by $M \geq 5$ in the converged MLP cells (ratios 0.84–1.15). Relatedly, the Adam RN/SR ratio is non-monotone in M ($202 \to 322 \to 107 \to 23$) because at $M{=}3$ the *denominator* is still elevated: Adam–SR itself suffers on the coarsest grid (1.5×10^{-3}, seed std 2.4×10^{-3}, vs. its flat $\approx 3.5 \times 10^{-4}$ floor from $M{\geq}4$ on), compressing the ratio from below; from $M{=}4$ the SR floor is constant and the ratio decays with the RN numerator alone. That constancy is worth flagging on its own: an Adam–SR floor that no longer moves with M has reached the task/optimizer floor rather than a storage-precision floor, so for $M \geq 4$ the Adam row of Table 2 is an RN floor divided by a constant rather than a ratio of two precision floors. The preregistered exclusions remove cells where *all* modes reach the task floor, not cells where only the denominator has; the unnormalized RN and SR floors are in the released `exp5_mlp.json`.

Full 440-run grid. The count: $440 = 4$ architectures $\times$ 5 M $\times$ 11 cells on `digits` plus 4 further datasets $\times$ 5 M $\times$ 11 cells at the reference architecture, where the 11 cells per (task, M) pair are {sign, SGD} $\times$ {fp, RN, SR}, two master-weight controls, and Adam $\times$ {fp, RN, SR}. Across datasets, the RN/SR fragility ordering Adam $>$ SGD $>$ sign holds in all 15 cells of the three tasks where all three optimizers are resolved (`digits`, MNIST-100%, MNIST-10%). The three tasks are not equally converged, and that shows in the *size* rather than the order of the gaps: full-MNIST at our step budget has not reached its floors, so its ratios are compressed into 1.09–1.67 against `digits`' 0.84–322, consistent with the theory that floors are stationary phenomena. The ordering survives the compression. On California housing the ordering is *not* resolved and we do not count those cells: SGD diverges at every M (excluded), and the sign–RN cells carry seed spreads of 9–97% of their means, which is wider than the Adam–sign gap they would have to establish. Raw JSON for all 440 runs is included in the supplement.

Does acting on the diagnostic change anything? Every other experiment here describes a phenomenon. This one tests the intervention the diagnostics exist for. On a 4-layer, width-64 MLP on `digits` (8 tensors, 13,130 parameters) we run plain RN storage for 400 steps, record the per-tensor fraction of coordinates with $\phi < 1$, one reduction per step, no oracle and no second training run, then compensate only the top-k tensors by that fraction and leave the rest on RN. Two compensations with different costs: SR (an RNG draw per element, no extra memory), and an fp32 master copy for the flagged tensors only (one extra fp32 word per flagged parameter). The control is the same k chosen at random.

The intervention works, and cheaply. At $M{=}3$ with Adam, RN floors at 3.7×10^{-2} against ≈ 0 for all-SR; compensating the single flagged tensor (the output matrix, 4.9% of the parameters) recovers 99.6% of that gap, and two tensors recover all of it. The same holds for the master variant, so 4.9% extra fp32 storage buys what a full master copy buys on this problem.

What the control shows is more equivocal, and we report it because it bounds the claim. A *random* single tensor also recovers essentially the whole gap here — it just costs 31% of the parameters instead of 4.9%. In the other three cells ($M{=}3$ sign, $M{=}5$ Adam and sign) the flagged and random curves recover comparable fractions at comparable cost, and the random arm is noisy enough at three seeds and one permutation (66% and 18% at $k{=}3$ and $k{=}2$ in two cells) that we would not read a difference from it. So on this model the diagnostic picks a *cheaper* sufficient subset, not a more effective one: the network is over-parameterized enough that rescuing any one weight matrix restores it. Whether ranking by ϕ beats ranking by size or at random on a model too large for one tensor to carry the rescue is settled at the 124M scale of Appendix G.6, not here; this MLP does not have the capacity to separate the two rankings.

One negative reading is worth stating separately. At $M{=}3$ under sign descent the measured $\phi < 1$ fractions are tiny (0.000–0.11), which by itself predicts almost no exposure, and yet RN floors at 1.3×10^{-1} against ≈ 0 for SR. The damage there is not freezing but the binade wall (Corollary 4.3), a different row of Table 1. It is a concrete reminder that ϕ is one of four criteria and does not subsume the others.

G.1 Leaving $\sigma_2 = 0$: the $2/3$ floor is a window, not a knife edge

Theorem 5.6 proves the $2/3$ floor with the second coordinate's noise exactly zero. That is the paper's most attackable hypothesis (real training has low-noise coordinates, not noiseless ones), and nothing above says what happens at $0 < \sigma_2 \ll \sigma_1$. We measured it (`exp26b_sigma2_interp.py`, aggregated by `exp26_sigma2_analysis.py`): same instance, same RN storage, $\sigma_1 = 1$, σ_2 swept from 0 to 1, the tuned floor's exponent in u fitted over $M \in \{9, 10, 11\}$ with a *geometric* η grid from $0.1u$ to $3\eta^\star$ so the optimum is interior at every point.

σ_2/σ_1	exponent	95% CI	$\eta^\star/u$ ($M = 9, 10, 11$)
0	0.662	[0.661, 0.662]	8.0, 9.6, 11.6
10^{-3}	0.671	[0.670, 0.673]	8.0, 9.6, 11.6
3×10^{-3}	0.703	[0.702, 0.705]	8.0, 9.6, 11.6
10^{-2}	0.786	[0.785, 0.788]	8.0, 9.6, 11.6
3×10^{-2}	0.998	[0.996, 1.000]	3.9, 4.5, 5.3
10^{-1}	0.895	[0.890, 0.899]	1.9, 2.1, 1.6
3×10^{-1}	0.927	[0.920, 0.935]	0.6, 0.7, 0.7
1	0.949	[0.922, 0.980]	0.21, 0.21, 0.22

The exponent leaves $2/3$ slowly and arrives near 1 by $\sigma_2 = 3 \times 10^{-2}\sigma_1$: a window of about $30\times$ in σ_2, not a knife edge. So the theorem's hypothesis is an idealization of a *regime*, and the regime is the one where a coordinate's noise is three orders of magnitude below the loud one, which is what "low-noise coordinate" means in practice.

Two things in that table are worth more than the exponents. First, the tuned $\eta^\star$ migrates: it sits exactly at the balance point $\eta^\star = (u^2/(L\sigma^2))^{1/3}$ while the $2/3$ floor holds, and moves to the u scale as the coordinate starts to dither, the two regimes the two theorems describe, with the tuner walking between them on its own. Second, once $\sigma_2 \gtrsim 10^{-1}$ the product $\rho_2 = \eta^\star\sigma_2/u$ stays in 0.16–0.22

across a tenfold range of σ_2 and all three mantissa settings: the tuner drives the noisy coordinate to a roughly fixed ditherability, which is the dither valley of Figure 4 found from the other side.

We do *not* claim the plateau is exactly 1. The four large-σ_2 points span 0.89–1.00 and the $\sigma_2 = 10^{-1}$ interval excludes 1. A three-point fit in M is a short lever arm, and this protocol reads 0.662 at $\sigma_2 = 0$ where EXP19's five-point, five-seed measurement reads 0.6669. That bias is ~ 0.005 and does not explain the 0.05–0.10 shortfall at the large-σ_2 end, which is a stated limit of this protocol: read those rows as the exponent having left $2/3$, not as measurements of 1. The homogeneous scale itself is measured at five mantissa points and five seeds in Figure 2, where every CI contains 1. The $2/3$ claim rests on EXP19; this sweep is evidence about the *shape* of the interpolation, and what it establishes is that there is one.

G.2 Acting on the diagnostic at 124M: which tensors, not how many

EXP22 asked whether compensating the flagged tensors beats compensating arbitrary ones, and could not answer: the MLP there is over-parameterised enough that rescuing any one weight matrix restores it. At 124M no single tensor can carry the rescue. Protocol (`exp27_selective.py`): run plain RN for 300 steps, rank the 148 tensors by the fraction of coordinates with $\phi < 1$ (one reduction per tensor per step, no oracle, no second training run) then apply SR to a chosen set and leave the rest on RN. Adam, three seeds, the same 3000-step budget and tuned step size as the main grid.

The diagnostic finds the tensors the theory names, without being told. Its top 25 by $\phi < 1$ are *exactly* the 25 LayerNorm gain tensors, $25/25$. Those are 19,200 of 124.4M parameters: 0.0154%.

Paired per-seed differences in training loss (the arms share seeds, data order and initialization, and the seed-to-seed drift here is 0.096 in loss, as large as the biggest effect below and four times the smallest, so only the paired comparison is meaningful):

comparison	mean $\pm$ sd over 3 seeds	sd from 0
flagged 25 vs. all-RN	-0.1006 ± 0.0183	5.5
flagged 25 vs. *one* random 25†	-0.0231 ± 0.0056	4.1
flagged 25 vs. all-SR	-0.0504 ± 0.0263	1.9
flagged 74 vs. all-SR	-0.0746 ± 0.0184	4.0
all-SR vs. all-RN	-0.0502 ± 0.0091	5.5

† paired across the three seeds against *one* drawn set. The seed spread it is measured against is not the spread that matters here. Across 20 draws the spread is 2.5× larger, and against that null the flagged set is 1.3 sd, not 4.1. See the random-control paragraph below; this row is kept because it is what EXP27 measured, not because it is the right test.

Three readings, and one comparison that does not resolve. Compensating 0.0154% of the parameters buys 0.101 in training loss, *twice* what compensating all of them buys. That "twice" is a ratio of two separate paired differences, each at 5.5 seed sd; the *direct* paired test of selective against blanket, flagged 25 vs. all-SR, is -0.0504 ± 0.0263, only 1.9 seed sd, and we report it rather than leave it out. At three seeds it is not resolved: what the data support is that the flagged set recovers at least as much of the gap as blanket SR while running the RNG on 0.0154% of the parameters instead of all of them (SR costs a draw per element and no extra memory, so the saving is in work, not bytes), not that it beats it. And more compensation is not better. Compensating the flagged half beats compensating everything by 0.075 ± 0.018.

Twenty random draws separate two readings that one draw cannot (`exp30_random_null.py`). Against *one* random draw the flagged 25 win by 0.023 ± 0.006, which invites the reading "which tensors matter, and not merely how many." One draw is not a null distribution, so the null is 20 draws, all at training seed 1337 so that within-condition noise is GPU nondeterminism alone. The first of the two readings below rules that interpretation out.

Among random draws, what predicts the loss is *how many parameters* you compensate, not how many LayerNorm gains you happen to catch. Regressed on $\log_{10}$ of the compensated parameter count the random draws give $r = -0.84$ (train), and -0.85 (held-out), $p < 10^{-5}$; regressed on the number of gains they give $r = +0.42$ and $+0.56$, the *wrong* sign, and in a joint regression the

gain term collapses to $t = +0.8$ and $+2.2$ while the parameter term stands at $t = -5.6$ and -6.0. The gain count and the parameter count are correlated at -0.38, because a gain is 768 parameters against tensors of up to 38.6M (the tied embedding, which four of the twenty draws contain and which anchors the high end of the fit), so the apparent "more gains is worse" is the parameter effect showing through. Within random selection, how many plainly does matter, and any reading that denies it is wrong.

What survives, and is stronger, is a statement about *efficiency*. The flagged 25 compensate 19,200 parameters (309 to 3281 times fewer than any of the 20 random draws, which span 5.9M to 63M), and still beat 18 of 20 on training loss and 17 of 20 on held-out. On the raw percentile that is $z = -1.33$ and -1.07, not significant on its own. Measured against the parameter-count trend it is not close: at 19,200 parameters the trend predicts 2.038 (train), and 2.044 (held-out), against an observed 1.896 and 1.895 — residuals of -18.3 and -18.4 residual standard deviations. That is an extrapolation 2.5 decades below the fitted range and we flag it as one; the check that the trend extrapolates at all is *top74*, which is outside the range on the high side and lands on the line (-0.3 and $+0.1$ residual sd). The claim the data support is therefore not "which, not how many" but something more useful: random selection buys loss with parameter count, and the diagnostic buys the same loss three orders of magnitude more cheaply.

We do not have a verified mechanism for that last one, and we checked two candidates that fail (all weight-scale numbers here are seed 1337; the across-seed ranges are in `exp27_selective.json`). The natural story is that SR injects rounding noise wherever it is applied, so on tensors where freezing never binds it is pure cost, but the non-gain tensors reach $\max|w| = 0.836$ under the flagged-74 arm against 0.777 under all-SR (seed 1337), i.e. the arm that wins shows *more* of the symptom, not less. The second story, that the winning arm simply freezes less, also fails: the flagged-25 arm has a *higher* frozen fraction than all-RN (0.598 against 0.584) while beating it by 0.10 in loss. What we can report is the ordering and its significance, not its cause.

What the diagnostic separates is not a matter of ranking. The 25 gain tensors have a $\phi < 1$ fraction of exactly 1.0000 (*every* coordinate's update is sub-ulp) while the largest value among the other 123 tensors is 0.508 and the median is 0.095. The cliff falls between rank 25 and rank 26; there is nothing to tune.

Implementation check, at the level of individual tensor indices: the random-25 set happens to contain 4 of the gains, and in that run those four leave the wall ($\max|w| = 0.97, 1.16, 1.30, 1.20$) while the other 21 gains stay pinned at exactly 1.0000. The per-tensor storage map does what it says.

Two limits. Held-out loss moves the same way but is not resolved at three seeds (-0.122 ± 0.045 for flagged 25 against all-RN, 2.7 sd). And "flagged 25" and "top 25" are the *same* set, run twice; the 0.006 mean difference between those two arms at fixed seed is this GPU's run-to-run nondeterminism, and it is the floor against which the numbers above should be read.

G.3 Extended limitations

Schedules. The heterogeneous $2/3$ floor (Theorem 5.6) is proved for constant step sizes. The sign-side Theorem 6.3 already covers arbitrary predictable schedules, including adaptive ones; whether the lattice-chain reduction that makes that possible has a two-coordinate analogue is the obstruction.

The open homogeneous lower bound. Theorem 5.4 shows homogeneous SGD–RN *attains* the universal floor; we have no matching lower bound at that scale. Simulations put the constant at $0.176\,u\sigma/L$, stable to 6.7% over $M \in [8, 12]$, below SGD–SR's 0.40, consistent with dithered-RN residuals carrying variance $u^2/12$ against SR's $u^2/6$, which is what a matching bound would have to reproduce.

Gaussianity. It is used twice: for the characteristic-function collapse of Theorem 5.4, where truncating at 3σ perturbs each harmonic by an additive $O(e^{-9/2})$ (Appendix B.5), and the collapse persists numerically; and for the sign-bias inequality $0 \le z(2\Phi(z) - 1) \le \sqrt{2/\pi}z^2$, where any noise with a linear signal-to-flip-probability margin changes only constants.

Averaging and the final iterate. The exact constant is proved for the movement-weighted average. A final-iterate statement is limited by the representation floor and by lucky stopping (a scheme can freeze at a grid point that happens to be close to μ), and a tight characterization is open (Appendix C).

Adaptive methods. Adam and Lion enter through the momentum propositions and through experiments, not through floor theorems. Our transformer and synthetic runs support the mechanism a stylized-Adam theorem would need: Adam's freeze threshold is independent of σ (measured drift 23% across a $16\times$ range of σ, where SGD's threshold would move by $16\times$), because the normalized update's dither scale is $\eta\sqrt{(1-\beta_1)/(1+\beta_1)}$, not $\eta\sigma$. Adam–SR floors at $\Theta(u)$ on our instance (exponent 0.98, CI $[0.95, 1.02]$). Across $(\beta_1, \beta_2, \epsilon)$ the freeze margin orders the RN penalty at 124M (Spearman 0.900, Appendix G.6), which is the mechanism such a theorem would have to reproduce, not a substitute for it.

Exponent range. We assume an unbounded exponent. Finite ranges add overflow and underflow failure modes, which can only strengthen a lower bound — except through the subnormal range, where the spacing stops tracking $|w|$ and SR's floor stops improving with it, which is measured on the MLP grid in Appendix G.5 and costs Table 2's $M{=}3$ SGD entry a factor of 5 against real E4M3.

Discrete gradient noise. Theorem 5.4's collapse is the Gaussian characteristic function; a gradient quantized before the update (an FP8 gradient) has discrete noise, and whether it dithers a stored weight depends on its lattice spacing relative to u. Sign-type updates are exactly invariant to gradient quantization (Proposition 4.4); the SGD–RN statement under discrete noise is open.

Corpus and budget. The 124M grid is reported on TinyStories and repeated on WikiText-103 (Appendix G.6), where every claim replicates and the effect sizes shrink slightly; we have not gone to web-scale text, where the gradient-noise scale that sets ditherability may differ again. The 10.8M SGD cells are not at a stationary floor and cannot be brought to one on this budget: at six times the published horizon the optimizer itself destabilizes at the tuned step, and at a stable smaller step both arms are still descending (Appendix G.6). The floors are measured in the converged MLP cells; the transformer runs measure the mechanism. The selective-compensation intervention is compared against blanket SR, blanket RN, an fp32 master and random tensor sets, not against error feedback (Richtárik et al., 2021; Nikdan et al., 2026) or block-wise quantized optimizer states (Dettmers et al., 2022), whose per-tensor costs differ from SR's.

G.4 Synthetic transformer and momentum experiments

Transformer. On a two-layer transformer (numpy, gradient-checked by central differences at $h = 10^{-5}$, worst relative error 8.5×10^{-7}) on an induction task with label noise, the excess-loss RN/SR ratios at $M{=}4$ are $2.6\times$ (SGD), $1.8\times$ (Adam), ≤ 1 (sign): sign is separated from both, but Adam and SGD are not separated from each other here, and in fact appear in the reverse order, the Adam-first ordering is a converged-MLP statement (Table 2), not a universal one; the binade wall is hit exactly (predicted/measured $0.125/0.125$, $0.5/0.5$, $1.0/1.0$) while SR ends at $2.75/3.63/9.5$, i.e. 7–$22\times$ past it. The flip side also appears: on the noiseless logit-growth task sign-SR can *underperform* sign-RN, as Corollary 6.5's limits allow (Appendix G).

Momentum. The absorption band of Proposition 7.1 holds in $9/9$ (β, M_m) cells; the amplification exponent fits at 0.99 (predicted 1). In the decay regime β_{crit} binds exactly: at $\beta = 0.999$ with bf16-like states RN momentum does not decay at all over $5{,}000$ steps ($|m| \equiv 1.0$, against 6.7×10^{-3} for fp and SR).

G.5 Robustness of the idealizations

Each theorem in this paper buys sharpness with an idealization. This section measures what each one costs.

The dither mechanism, measured directly. Theorem 5.4's explanation for dithered SGD–RN sitting *below* SGD–SR is a variance statement: a well-dithered RN residual is spread across the cell and carries variance $u^2/12$, while an SR residual sits on the two cell endpoints and carries up to $u^2/6$. Measuring the stationary residual $r_t = Q(y_t) - y_t$ at matched step size confirms it with no free parameter. At $\rho = \eta\sigma/u = 0.70$ the RN residual variance is $0.0831\,u^2$ against the uniform prediction $1/12 = 0.0833$ (a 0.2% gap), and the SR residual variance is $0.1664\,u^2$: a ratio of 2.001–2.002 across $M \in \{9, \ldots, 12\}$, against the predicted 2. It degrades only as the dither weakens ($0.0662\,u^2$, ratio 2.259, at $\rho = 0.30$, where RN is not yet fully dithered). The resulting

floor advantage is ρ-dependent (1.12–1.15× at $\rho = 0.7$, 1.63–1.70× at $\rho = 0.3$), because the floor also carries a bias term; the variance ratio is the part that is exactly 2.

The heterogeneous exponent under the prescribed protocol. The horizon requirement above applies to the heterogeneous panel as much as to the homogeneous one. It is tempting to keep $T = 1.5 \times 10^5$ there on the grounds that the noiseless coordinate relaxes in $e_{2,0}/u \leq 3.3 \times 10^3$ steps, true, and irrelevant to the noisy coordinate, which at $\eta \sim u$ gets 293 contraction times at $M{=}9$ but 9.2 at $M{=}14$. Since that is exactly the regime artifact (ii) lives in, and since the two arms were fitted over different mantissa windows, we remeasured *both* arms under the protocol this appendix prescribes: $M \in \{9, \ldots, 13\}$, per-cell $T = 120\tau/(1 - \text{burn}) = 240\tau$ with $\tau = 1/(\eta_{\min} L)$ for the *noisy* coordinate (floored at 1.5×10^5 steps, so the coarse cells get up to 820τ), $e_0 = e_{2,0} = 0.2$, 5 seeds, and the tail-drift check reported per cell.

The exponent survives and tightens: 0.6669 with seed-bootstrap 95% CI $[0.6644, 0.6693]$, against the predicted $2/3 = 0.66667$, with tail drifts in $[0.997, 1.007]$ and the constant $V/(u\sigma^2/L^2)^{2/3}$ at 0.7451–0.7498 (a 0.63% spread, against 1.19% before). This is the measurement the $2/3$ claim rests on, and two things about it should be said plainly.

The seed bootstrap is not the whole uncertainty. It resamples seeds within a fixed set of mantissa cells, so it captures seed noise and not the choice of window. Refitting the same five cells on three-point sub-windows gives 0.6622 (M 9–11), 0.6679 (10–12), and 0.6711 (11–13); leave-one-out gives 0.6652–0.6691. The first of these falls *outside* the quoted CI. The drift is monotone in M and it approaches $2/3$ from below, which is what a positive vanishing finite-precision correction looks like rather than a wrong exponent, the same shape as the $O(uL/\sigma)$ correction on the sign side, and equally unmeasurable at this budget. The honest summary is that the exponent is 0.662–0.671 over M-windows with a point estimate of 0.6669 on the full window, and that the *constant* (0.745–0.750, stable to 0.63%) is the better-determined quantity.

The anchoring does not manufacture it. EXP19's grid is anchored at the predicted balance point, so every grid point scales as $u^{2/3}$ by construction, and a reader is right to ask whether the fit could have come out any other way. It could: `exp26b` measures the same instance at $\sigma_2 = 0$ on a *geometric* grid running from $0.1u$ to $3\eta^\star$, which is anchored at nothing, and returns $0.6616\,[0.6612, 0.6620]$ on $M \in \{9, 10, 11\}$ against the anchored grid's 0.6622 on the same three cells — agreement to 6×10^{-4}, well inside the window drift above. The anchoring costs resolution in $\eta^\star$, not correctness in the exponent. On the non-anchored grid the tuned $\eta^\star/u$ reads 8.02, 9.65, 11.61 against the predicted $u^{-1/3}$ ladder 8.02, 10.10, 12.73: it climbs as predicted but with fitted slope -0.267 rather than $-1/3$, so $\eta^\star$ tracks the balance point to within 9% over this window and drifts systematically. Three cells is too few to call that a discrepancy or a correction, and we do not.

One protocol point the rerun exposed. Putting both arms on the *same* η grid, anchored at the RN balance point $\eta^\star \propto u^{2/3}$, is also wrong: SR's own optimum is at $\eta \sim u/\sigma$, far below, and measured on the RN grid its "floor" is the noiseless coordinate's stall, returning 0.674 instead of 1. Re-run on the same window $M \in \{9, \ldots, 13\}$ with the same horizons, seeds and replicas but a grid spanning *its own* valley ($\eta/u \in [0.12, 0.75]$), the SR arm gives exponent 0.9975 with seed-bootstrap CI $[0.9912, 1.0033]$ and constant 0.3948–$0.3987\,u\sigma/L$, tail drifts in $[0.986, 1.010]$. That constant is its *homogeneous* one (0.392–0.407 measured with no noiseless coordinate at all), which is what Theorem 5.3 says should happen: SR can always move, so the un-ditherable coordinate costs it nothing while it costs RN a divergent factor. Each mode needs a grid spanning its own valley *and* a common mantissa window; the two requirements are independent, and satisfying only one of them is enough to bias the fit.

The SR argmin is not identified, and that is the lemma, not a defect. On that grid the tuned $\eta^\star$ landed at the lower edge in one of the five cells ($M{=}10$), which we reported as a grid-boundary caveat. It is worth more than a caveat. Lemma 6.2 says SR's stationary law is invariant to η across a window, and inside a flat window an argmin is not a well-defined quantity, which end it lands on is noise. EXP19d (`exp19d_het_sr_wide.py`) re-runs all five cells with the grid extended to $0.06u$ and $0.085u$, below $u/8$ where the lemma's guaranteed window ends, at the doubled horizon the smaller η requires. The argmin moves, to $0.060, 0.085, 0.180, 0.180, 0.060$ for $M = 9 \ldots 13$, i.e. it wanders rather than converging, and it is interior in three cells and at the new lower edge in two. *The floor does not*: the constant reads 0.3921–$0.3954\,u\sigma/L$ against 0.3948–0.3987 on the narrower grid, and

at $M{=}10$ specifically 0.3943 against 0.3948, a 0.1% difference. And the profile is flat to 1.4–2.1% across $\eta/u \in [0.06, 0.38]$ (a $6.3\times$ range) against a per-cell seed spread of 0.8–3.9% of V. The variation over η is smaller than the noise on any single cell, so no grid can identify the argmin, and extending the grid buys a different winner rather than a better floor. The flat region reaches below $u/8$, so the lemma's window is sufficient and not necessary; and this is the opposite situation from SGD–RN, whose floor *is* grid-sensitive, which is why we quote a scaling there and a constant here. The practitioner's reading of Lemma 6.2 is worth stating in its own right, because it is surprising and cheap to check: under SR, once the step size is at or below the ulp, tuning it buys *nothing* at the floor — it only changes how fast the chain gets there. Learning-rate search in that regime is search over wall-clock, not over quality.

The ℓ_1 law, and what a shared step size costs (`exp32_ell1_law.py`). Corollary 6.6 is the paper's only d-dimensional statement and it was the last one left untested. On a separable instance whose coordinates carry different weight scales W_i (so different spacings $u_i = W_i 2^{-M}$) it predicts two things. With *per-coordinate* steps η_i in each coordinate's own invariance window the identities decouple and sum, giving $V = C^\star\,(\sigma/L)\,2^{-M}\sum_i W_i$; and a *common* η cannot reach that, because coordinates with $u_i \ll \eta$ floor at $\approx C^\star\eta\sigma/L$ regardless of W_i while coordinates with $u_i > 2\eta$ freeze. The falsifiable part is the second: the price of sharing should be 1 when the scales are equal and grow with their spread. We test it at $M{=}10$, $d{=}8$, $\sigma = L = 1$, scales geometric over $[1, \text{spread}]$, 5 seeds and 40 replicas, both arms at the horizon the smallest step size demands, with $\eta_i = u_i$ for the per-coordinate arm and a gap-free geometric grid of 14 points for the common one.

spread $W_{\max}/W_{\min}$	$1\times$	$2\times$	$4\times$	$8\times$	$16\times$
per-coordinate V / ℓ_1 prediction	1.001	1.003	1.003	1.003	1.005
price of a shared η	$0.99\times$	$1.00\times$	$1.12\times$	$1.35\times$	$1.71\times$
best common η, in units of $u_{\min}$	1.05	1.11	2.23	4.22	8.90

The ℓ_1 law is exact to 0.5% at every spread, with no fitted parameter: the prediction is $C^\star$ times a sum of scales. The price of sharing is 1 to within 1% when the scales are equal (as it must be, since there is then nothing to share badly), and grows monotonically with the spread, reaching $1.71\times$ over four binades. The mechanism is visible in the last row: the best common η starts at $u_{\min}$ and is dragged toward the geometric middle of the scales as they spread, which is exactly the trade-off the corollary describes, buying the large-scale coordinates a workable step at the price of running the small-scale ones far above their own windows. All optima are interior and all tail drifts lie in $[0.9989, 1.0020]$.

This is also the answer to the sharpest objection available to the $2/3$ separation of Theorem 5.6: that it is an artefact of forcing one step size on coordinates that want different ones. It is. That is the content rather than a defect. RN's excess cost is a cost of *step-size sharing*, which is why per-channel and per-tensor scaling help, and why SR, which removes the cost without needing to know the scales, is the cheaper fix. A reader who objects that per-coordinate steps repair the heterogeneous instance is agreeing with Corollary 6.6 and paying its price, measured above.

Why we quote a scaling and not a tuned constant for SGD–RN. Figure 4 shows $V(\eta)$ for homogeneous SGD–RN decreasing monotonically from the variance ramp down to the freeze cliff and jumping at it. The tuned minimum is therefore attained *at the cliff edge*, so its value depends on how finely the grid resolves the cliff: three of our own scans returned 0.151, 0.175 and 0.197 for the same quantity, with $\eta^\star$ pinned to the smallest non-frozen grid point every time. The exponent is unaffected (all scans give 1.00 within seed error), and so is the floor at any fixed ρ inside the valley, which is what we report.

Real FP8, not just a mantissa count. Table 2 maps $M{=}3$ to "the FP8-E4M3 mantissa budget," which is an arithmetic identity, not a measurement: E4M3 also has a 4-bit exponent and a finite range. Rerunning the floor experiment on the actual `ml_dtypes` `float8_e4m3fn` and `float8_e5m2` types, with rounding done on the exact 256-code table (253 and 247 finite codes), and stochastic rounding verified unbiased to 0.03% of a step — gives RN/SR ratios of 0.6012 (E4M3, SGD), and 0.5912 (E5M2, SGD) against 0.6012 and 0.5912 from our unbounded-exponent simulator at the matching mantissa: agreement to every digit reported, in all four cells, the sign-side

ratios are 0.7580 (E4M3), and 0.3516 (E5M2), matched to the same precision. On *this* instance the simulator is an exact stand-in and the finite exponent does not bind. On the MLP grid it does, and against us: see next.

Where the mantissa-matched simulator overstates the gap (`exp34_fp8_mlp_grid.py`). The synthetic instance above keeps its iterate near 1.5, so it never approaches a range limit. Network weights do: they initialise at $\Theta(1/\sqrt{\text{fan-in}})$ and the small ones keep shrinking, and E4M3's smallest normal is 2^{-6}. Rerunning the reference cell of Table 2 (`digits`, depth 2, width 32, the same lr grids, the same 4000 steps, the same 5 seeds and tail window) with storage on the exact code tables gives

	E4M3 vs. simulator $M{=}3$			E5M2 vs. simulator $M{=}2$		
RN/SR floor ratio	real	simulated	over	real	simulated	over
signSGD	3.57	3.77	1.06×	6.55	7.01	1.07×
SGD	4.24	21.9	5.16×	2.55	2.54	0.99×
Adam	115	202	1.77×	277	291	1.05×

The simulator column reproduces Table 2's $M{=}3$ entries to the digits reported, so the comparison is clean; and at E5M2 the two agree throughout. At E4M3 they do not, and the discrepancy sits entirely on the *SR* side: real-over-simulated tail losses are $0.98/1.19/1.01$ for RN against $1.04/6.14/1.79$ for SR. That is the theory's own prediction, not a counterexample to it. Both floors track the *local* spacing u, and below the smallest normal E4M3's spacing stops tracking $|w|$ and pins at 2^{-9}: at $|w| = 2^{-6}$ it is $2\times$ the simulator's, at 2^{-7} $4\times$, at 2^{-8} and below $8\times$. So on the 1.7–5.2% of weights that land in the subnormal range (measured per cell; E5M2 has none at this weight scale, its smallest normal being 6.1×10^{-5}) SR's $\Theta(u\sigma/L)$ floor stops improving, while RN, which had already frozen there, is unaffected. The tuned step size moves with it, from 1.0 to 0.3 for SGD–SR and 0.01 to 0.003 for Adam–SR.

Two things follow for how Table 2 should be read. Its $M{=}3$ column is a faithful *mantissa* statement and we keep it as one, but it is not a measurement of FP8-E4M3: for SGD it overstates the RN/SR gap by $5\times$. The direction is informative too. A finite exponent does not make RN worse, it makes SR less good, so the ordering the paper claims survives while the magnitude does not.

Truncated noise. Theorem 5.4 assumes exactly Gaussian noise and argues truncation at $c\sigma$ perturbs each harmonic by $O(e^{-c^2/2})$. Sweeping c confirms the shape: for $\rho \le 0.35$ the measured bias is within 0.8% across $c \in \{2, 2.5, 3, 4, 8\}$ (identical to three digits at $\rho = 0.35$, to two at $\rho = 0.2$), and truncation only sets a floor on how far the collapse can go, at $\rho = 0.8$ the measured bias bottoms out at $2.0 \times 10^{-2}u$ ($c{=}2$), $1.0 \times 10^{-3}u$ ($c{=}2.5$), $5.2 \times 10^{-4}u$ ($c{=}3$), and $2.0 \times 10^{-5}u$ ($c{=}4$); at $c{=}8$ it is $2.9 \times 10^{-5}u$, so the sweep flattens an order of magnitude above the untruncated $1.0 \times 10^{-6}u$ rather than reaching it. Our chain experiments use $c = 3$: over the measured range the bias falls from $9.8 \times 10^{-2}u$ to $5.2 \times 10^{-4}u$, a span of 2.3 orders of magnitude. (No measurement of this residual can span five: it is bounded by $u/2$. The five-order figure quoted elsewhere is the untruncated sweep of Appendix B.5.)

Mid-cell placement. The hard instances put μ exactly at a cell midpoint, which is worst-case for the *representation* floor (Theorem 5.1) by construction. It is not worst-case for the *dynamic* floors. Drawing μ uniformly inside a cell (24 draws, $M{=}10$): sign-RN is indifferent (mid-cell 6.02×10^{-4} against a random mean of 6.04×10^{-4}; 54% of draws land above mid-cell), while SGD–RN is *better* at mid-cell than at every random placement (mid-cell 1.68×10^{-4} against a random mean of 2.35×10^{-4}, and 100% of draws above it). Generic placements are $1.39\times$ worse on average and $1.59\times$ at worst, so the mid-cell choice understates the dynamic floor rather than inflating it.

Finite exponent range. Clipping the grid at 2^4 and 2^5 leaves the sign-RN floor unchanged to four digits ($0.6150\,u\sigma/L$ in all three settings), as expected for an interior optimum where the iterate never approaches the range limit. The exponent range binds in growth phases instead, which is where Corollary 4.3 already places it.

Lion: the mode equivalence does not survive momentum. Lion's update is $\mathrm{sign}(\beta_1 m + (1-\beta_1)g)$, a sign, so one might expect it to inherit Corollary 6.5's mode equivalence. It does not. Both modes floor at the universal scale (fitted exponents 0.95 $[0.92, 0.98]$ (RN), and 0.99 $[0.97, 1.00]$ (SR) over $M \in [8, 12]$), but the constants are 2.90–3.35 for RN against 1.03–1.08 for SR, a ratio of 3.0. The lattice reduction behind Lemmas 6.1–6.2 needs the sign to be a function of the current state plus *fresh* noise; Lion's EMA gives the sign a memory whose timescale interacts with SR's random holding times, and the two chains stop coinciding. So the correct statement is that mode equivalence is a property of *memoryless* sign updates, and a Lion floor theorem is genuinely open rather than a corollary. Decoupled weight decay at 10^{-2} and 10^{-1} left the floor unchanged at these scales.

The price of dropping movement weights. Remark 6.4 notes that the exact constant is proved for the movement-weighted average and that a scheme can freeze at a lucky grid point. Solving the lattice chain exactly — average-cost dynamic programming over half-integer error levels $e = (k + \frac{1}{2})u$, since μ is a cell midpoint and $|e| \geq u/2$ always, makes the price exact. The optimal predictable schedule for the *plain* average achieves precisely the representation floor $u^2/4$, by moving to the nearest grid point and then freezing, so

$$\frac{\text{optimal plain average}}{C^\star u\sigma/L} = \frac{uL}{4C^\star\sigma},$$

which the DP reproduces to six digits at $M = 8, 10, 12, 14$ (1.558×10^{-3}, 3.90×10^{-4}, 9.7×10^{-5}, 2.4×10^{-5}). Movement weighting is therefore not a technical convenience: without it the bound is false by a factor $\Theta(uL/\sigma)$, and the gap is exactly the ratio of the representation floor to the noisy floor. (Integer error levels would place a state at $e = 0$ that the mid-cell instance does not have; the DP would then park there and report an optimal average of zero, which is why the levels are half-integer.)

Do the real-data ratios grow with the horizon? Appendix G concedes that some cells of the 440-run grid are still transient, so their RN/SR ratios understate the floors. Reusing EXP5's model, training code and tuning protocol, but at a reduced configuration (the reference architecture only, three seeds), so its $1\times$ column is *not* Table 2, whose numbers average four architectures and five seeds, and raising only the step budget within that configuration confirms the prediction for the two optimizers whose updates are gradient-scaled: at M=7 the SGD ratio goes $4.3 \to 16.3 \to 32.0$ at $1\times$ /$4\times$ /$8\times$ its own budget and the Adam ratio $15.0 \to 348.6 \to 480$, and similarly at M=5 ($13.7 \to 60.7 \to 118.9$ and $86.6 \to 849.0 \to 1.1 \times 10^6$, the last Adam cell already in the runaway regime described next). The published numbers were transient- limited, as claimed. Neither has plateaued by $8\times$, and at the coarsest mantissa the Adam ratio runs away ($10^3 \to 10^8$ at M=3) because its RN arm is by then frozen while SR keeps descending, past that point the ratio compares a frozen run to a converging one rather than two floors, so we read the trend and not the endpoint. signSGD moves the other way ($1.3 \to 0.08$ at M=7) because on this task RN's freeze is an *advantage* over SR's slowed clock once the budget is long, the flip side of the wall that Corollary 6.5 already scopes, and a reminder that its equivalence statement is about the stationary floor, not about equal progress at a fixed budget.

G.6 Transformer training with bf16 weight storage

Setup. A 10.8M-parameter GPT (6 layers, 6 heads, width 384, context 256) on character-level text, batch 32, 6000 steps. Weights are `torch.bfloat16` parameters; each step computes the update in fp32 and writes back $w \leftarrow Q(w + \Delta)$ with Q round-to-nearest (the hardware cast), or stochastic rounding. This is *not* mixed-precision autocast, which keeps an fp32 master copy; that recipe is our control arm, where the fp32 master is updated and cast to bf16 only for the forward pass. SR from fp32 to bf16 adds a uniform 16-bit value to the discarded low half of the word and truncates; incrementing the integer view steps to the next float away from zero, so the rule is unbiased in magnitude for both signs. Validated at 2×10^5 samples on a value one quarter of the way into a cell: $P(\text{up}) = 0.2490$ against 0.25, mean bias -7.6×10^{-6} (that is 0.1% of an ulp), and the same for the negative reflection; the run is logged in `results/bf16_sr_selftest.json`.

Grid. Three optimizers (AdamW, SGD, signSGD) $\times$ three storage modes (RN, SR, fp32-master) $\times$ three step sizes (27 runs of 6000 steps), plus a β_1 sweep $\{0.99, 0.997, 0.999, 0.9995\} \times \{\text{bf16, fp32}\}$ momentum state at $\beta_2 = 0.9999$ (8 runs of 3000 steps): 35 runs and $27 \times 30 + 8 \times 15 = 930$ logged

checkpoints. Four further runs from a first β_1 sweep are excluded and reported below. Diagnostics are logged every 200 steps: per-coordinate freeze margin, the fraction of stored words whose bits are unchanged since the previous step, the momentum buffer's equivalent, and per-tensor $\max|w|$.

Freeze margin predicts freezing. Over the 930 checkpoints of the 35 valid runs the fit is frozen $= 1.01\,\phi_{<1} + 0.03$ with $r = 0.991$ and mean absolute error 0.039; per optimizer, $r = 0.986$ (Adam), and $r = 0.978$ (SGD). signSGD is the exception at $r = 0.563$, but only because it has almost no variance to explain: predicted 0.009 against measured 0.017, both essentially zero, because at these step sizes sign updates clear the threshold everywhere, which is what Theorem 4.2 says happens above it. Note $\phi < 1$ is sufficient but not necessary for a single step to be absorbed, so the diagnostic is expected to under-predict slightly, and does ($+0.03$ intercept).

Seeds at 10.8M, and where the fragility ordering does not hold. The nine tuned cells were also re-run at seeds 1338 and 1339 at this scale (`exp25_nano_seed_analysis.py`). One reading confirms the account and one qualifies it.

The held-out reversal is real and its mechanism is visible. This corpus is small enough that the model memorizes it: with the fp32 master, held-out loss ends $9.7\times$ the training loss for Adam and $4.4\times$ for sign, but only $1.05\times$ for SGD, which never fits the task at this budget. The held-out RN/SR ratio splits on exactly that line — 0.864 ± 0.008 (Adam), and 0.952 ± 0.023 (sign), RN winning in $3/3$ seeds and beating the master in $3/3$; against 1.176 ± 0.046 (SGD), RN losing in $3/3$. So RN's freezing is a regularizer precisely where there is something to regularize, and pure damage where there is not. On training loss RN is worse in every one of the nine cells: 1.174 ± 0.012 (Adam), 1.247 ± 0.063 (SGD), 1.063 ± 0.034 (sign).

The Adam-over-SGD half of the ordering does not reproduce here. On training loss at 10.8M the order is SGD $>$ Adam $>$ sign, not Adam $>$ SGD $>$ sign. We think the reason is that SGD is not near a floor at this scale (with SR it still sits at a training loss of 1.73 against Adam's 0.32), so its RN/SR gap measures how much rounding slows an optimizer that has not converged, not the ratio of two stationary floors. That is the same caveat we record for the synthetic transformer, and it is why the ordering claim is stated for converged cells and tuned budgets; six times the budget does not remove it (see below). In the converged MLP cells Adam $>$ SGD $>$ sign holds in all 15; at the 124M tuned budget it holds in every seed on both losses. The strict Adam-over-SGD half is a tuned-budget statement there, not a floor statement: at three times the budget the two tie on training loss (1.0337 each), and SGD edges ahead on held-out within two-seed noise (the budget table below), consistent with SGD still descending toward its floor while Adam has stopped. A reader who wants the ordering to be a scale-free law should read these exceptions as saying it is not one: it is a statement about ditherability at the floor, and it needs the floor.

Can the 10.8M SGD cells be run to a floor? (`exp35_nano_sgd_converge.py`, `exp38_nano_lr_control.py`). The caveat above, that SGD's gap here measures a slowed transient rather than a ratio of floors, invites the obvious experiment: give it more budget. We gave it six times, 36000 steps, all three storage arms, three seeds, at the $\eta = 0.3$ that won the 6000-step grid. It does not work, in a way worth recording.

At the tuned step the optimizer itself comes apart, and not because of rounding. Past roughly 25000 steps the runs start spiking: over the last 97 logged checkpoints, the fraction above $1.5\times$ the running median is $8/97$, $34/97$ and $16/97$ for SR — and $0/97$, $30/97$, $32/97$ for the *fp32-master* arm, which has no storage rounding at all and is the one that reaches `NaN`, at step 30000. RN alone is stable ($0/97$ in all three seeds), because it has already frozen most coordinates and never reaches the regime where the instability bites. This is plain SGD at the top of a grid tuned for a sixth of the horizon, not a precision effect. (One tempting mis-reading is worth naming: SR seed 1338 ends with $\max|w| = 19.4$ against 4.2–4.8 elsewhere, which looks like the binade wall letting SR run away. That reading does not survive the checkpoints: $\max|w| = 4.25$ is the embedding initialisation, present from step 0 in every arm, and the jump to 19.25 happens in the same checkpoint as that seed's loss spike, so it follows the spike rather than causing it.)

At a stable step, both arms are still descending. Rerunning all three arms at $\eta = 0.1$ for the same 36000 steps gives $0/97$ spikes in all six runs, no `NaN`, and no weight growth — and RN's loss is still falling at the end. So the answer to "run it to a floor" is that 10.8M will not get there on this budget either way, and the transient caveat stands, now with a bound on it: $6\times$ is not enough.

What the pair does establish is the mechanism, and cleanly. Holding the budget at 36000 steps and cutting the step size by $3\times$ moves the freeze margin and the penalty together, exactly as Theorem 4.2 says it should:

η	RN median ϕ	RN frozen fraction	RN/SR
0.3	0.813	0.635	1.170 ± 0.151
0.1	0.195	0.903	1.699 ± 0.001

A smaller step is a smaller ϕ, a larger frozen fraction, and a larger RN penalty — and the seed spread collapses from ± 0.151 to ± 0.001, because at $\eta = 0.1$ the dynamics are stable and which coordinates freeze is essentially deterministic. At that step SGD–SR equals the fp32 master to four significant figures (1.0000 ± 0.0000 on the late-median training loss) while RN sits 70% above both: the compensation recovers the master exactly, and the thing it is compensating is freezing. That is a sharper statement than the floor measurement we set out to make, and it does not need the runs to have converged.

The same measurements at GPT-2-small scale. The 124M configuration (12 layers, 12 heads, width 768, context 1024, 50257 BPE tokens, tied embeddings, GPT-2 initialization, warmup and cosine decay, gradient clipping at 1) consumes 25M tokens over 3000 steps at batch $8 \times 1024 = 8192$ tokens, under half of the 58.8M-token training split. The corpus is **TinyStories** (Eldan & Li, 2023), GPT-2 BPE, and naming it matters for reading the losses: it is synthetic, small-vocabulary and highly redundant, which is why a 124M model reaches ≈ 2 nats held-out on 25M tokens where the same model on web text would not. It was chosen so that three optimizers $\times$ three storage modes $\times$ three step sizes $\times$ three seeds would fit the compute available and still leave every arm past its transient; nothing in the claims depends on the corpus being hard, and the small batch is worth flagging in the other direction — it is the regime in which gradient noise, and therefore self-dithering, is largest. The grid is again three optimizers $\times$ three storage modes $\times$ three step sizes (27 runs), plus four β_1 runs, with diagnostics logged every 100 steps at this scale (31 runs $\times$ 30 checkpoints $= 930$). Results: $r = 0.987$ over the 930 checkpoints with frozen $= 0.94\,\phi_{<1} + 0.07$ and MAE 0.033; per optimizer $r = 0.979$ (Adam), 0.980 (SGD), 0.966 (sign), at this scale signSGD has enough spread in $\phi_{<1}$ for the correlation to be meaningful, which it did not at 10.8M. The wall reads cleanly not by luck of scale but because the tensors it binds on are the LayerNorm gains, which every transformer initializes at exactly 1.0, a binade boundary, with bf16 ulp 2^{-7} above it and 2^{-8} below, so moving a gain *up* costs twice what moving it down does. Over the 25 gain tensors, RN leaves all 25 at $\max|w| = 1.0000$ exactly in 7 of the 9 runs, and $24/25$ and $8/25$ in the two largest-step SGD runs; SR pins none in any run, reaching 1.234–1.656; the fp32-master control pins all 25 in none, reaching 1.133–1.664, though its smallest-step SGD run pins 18 because the master itself has barely moved by step 3000. At the middle step size of each grid the whole-model $\max|w|$ reads 1.000, 1.008 and 1.000 under RN for Adam, SGD and sign, against 1.508, 1.609 and 1.438 under SR and 1.492, 1.414 and 1.312 under the master control. Momentum: with a bf16 buffer the per-step frozen fraction of m is 0.0168 at β_1=0.99 and 0.0893 at β_1=0.999, against 0.0000 for the fp32 control at both. **Seed robustness.** Every number above is from seed 1337. We re-ran the nine tuned cells at two further seeds (1338, 1339), keeping the learning rate that won each cell's grid on 1337 rather than re-tuning, so that seed variance is not confounded with grid variance (`exp23_seed_analysis.py`). All three claims survive, and two of them sharpen into measurements:

The wall. RN leaves all 25 LayerNorm gains at $\max|w| = 1.0000$ in 8 of the 9 (optimizer, seed) cells and $24/25$ in the ninth; SR pins not one gain in any of the nine; the fp32-master control pins all 25 in none.

The diagnostic. $r(\phi_{<1}, \text{frozen})$ is 0.9928, 0.9927 and 0.9928 on the three seeds, identical to three significant figures.

The held-out estimator. `exp12` logged `final_val` as a *single* validation batch ($8 \times 1024 = 8192$ tokens) at the last logged step, while the training number averaged the last 20% of logged steps. That asymmetry is not defensible at these effect sizes: the spread of the last six val readings within one run is 0.16–0.18 nats, against ratios of 1–3%. We therefore quote the same tail average for held-out loss as for train (the mean over the last 20% of logged steps, six points) recomputed from the committed logs by `exp29_tailval_analysis.py`. The estimator is stable for two reasons. Every arm at a given seed draws its batches from one generator seeded at `cfg["seed"]+2` and

consumes them in the same order, so RN, SR and the master are scored on *the same* held-out batches and the batch noise cancels in the ratio; and the tail average uses six of them rather than one. This is why the across-seed spread of a ratio is 0.001–0.004 and not 0.08. The single-batch numbers are reported alongside in the JSON; they shift the ratios by at most 0.012 and change no conclusion.

The ratios. On held-out loss, RN/SR is 1.0245 ± 0.0011 (Adam), 1.0113 ± 0.0034 (SGD), and 0.9929 ± 0.0036 (sign), i.e. 22.7, 3.4 and 2.0 seed standard deviations from 1; on train, 1.0267 ± 0.0054, 1.0127 ± 0.0030 and 0.9935 ± 0.0041. The fragility ordering Adam $>$ SGD $>$ sign holds in every seed on both. Sign's ratio is the one that does not clear 1%, and it sits slightly *below* 1 — mode indifference at the floor (Theorem 6.3) plus the small weight-growth advantage the wall gives RN, which is the direction Corollary 6.5 already flags.

Stochastic rounding recovers the master copy. SR/master is 0.9989 ± 0.0024 (Adam), 0.9950 ± 0.0035 (SGD), and 1.0015 ± 0.0018 (sign) on held-out loss — each within 1.5 seed standard deviations of 1, and below 1 in five of the nine (optimizer, seed) cells. At this scale and budget the fp32 master buys nothing the rounding rule does not, at twice the weight memory. That is the practical form of Theorem 5.3: SR's residual is conditionally mean-zero, so it never freezes, and freezing is what the master copy was there to prevent.

The tuned RN/SR loss ratios at seed 1337 are 1.025 (Adam), 1.010 (SGD), and 0.991 (sign): storage rounding costs little at this budget. Our reading has been that 3000 steps on 124M parameters is deep in the transient and a longer run would separate the arms, as it does on the MLP (Appendix G.5). We tested that here and it did not happen.

The budget dependence is not one story (`exp24_horizon_analysis.py`). Re-running the same cells at 9000 steps moves the three optimizers in three different directions. Both budgets carry seeds (three at 3000, two at 9000), and the ratios go

RN/SR	3000 steps (3 seeds)	9000 steps (2 seeds)	change
training loss			
Adam	1.0267 ± 0.0054	1.0337 ± 0.0065	$+1.3$ seed sd
SGD	1.0127 ± 0.0030	1.0337 ± 0.0050	$+6.9$ seed sd
sign	0.9935 ± 0.0041	0.9264 ± 0.0027	-16.3 seed sd
held-out loss			
Adam	1.0245 ± 0.0011	1.0223 ± 0.0079	-2.1 seed sd
SGD	1.0113 ± 0.0034	1.0317 ± 0.0073	$+6.1$ seed sd
sign	0.9929 ± 0.0036	0.9265 ± 0.0017	-18.4 seed sd

So SGD's ratio grows with the budget on both losses, as the transient reading predicts and as it does on the MLP; Adam's does not grow at all, moving $+1.3$ sd on train and -2.1 on held-out; and sign's moves sharply the *other* way, RN ending 7.4% *better* than SR, identically on both losses and with a two-seed spread of 0.004 on train and 0.002 on held-out. Any account that explains only the first of these is not an account.

The sign master-weight control at 9000 steps completes the nine-cell grid: train 2.4715, held-out 2.6544. One caveat is not removable by more seeds: 9000 steps is 73.7M tokens against a 58.8M-token training split, so these runs make 1.25 passes and end with held-out above train by 1.07–$1.13\times$ on the master arms, against 1.02–$1.05\times$ at 3000 steps. They are a different regime, not the 3000-step runs continued, and we do not read them as one.

The quantity that does organize them is the wall, and it is measured rather than posited. **RN holds the whole model's** $\max|w|$ **at** 1.000 **in five of the six RN runs across both budgets**, the sixth (SGD at 3000 steps) sitting one bf16 ulp above at 1.008, while the compensated arms grow: $\max|w|$ reaches 1.31–1.61 at 3000 steps and 1.92–2.20 at 9000. The wall's constraint is fixed; what it constrains keeps growing. So the wall's effect on the comparison strengthens with the budget, in the direction that *helps* RN, while freezing (the floor mechanism) keeps hurting it. The net is a race between the two, and the three optimizers sit at different points of it. Sign descent is the clean case: Theorem 6.3 makes its floor rounding-mode-indifferent, so the wall is the *only* thing left to separate RN from SR, and at 9000 steps SR has let $\max|w|$ run to 2.20, the largest of any run here, at a cost of 7% in training loss. That is Corollary 6.5's weight-growth caveat, and this experiment says it grows with the budget rather than washing out.

Adam's held-out ratio falling while its training ratio holds still has the memorization explanation: 9000 steps is 73.7M tokens against a 58.8M-token training split, so the longer runs make 1.25 passes and the master arm ends 1.13× above its training loss (SGD 1.08×, sign 1.07×) against 1.05× (SGD 1.02×) at 3000 steps — held-out sits above train at both budgets, and what grows is how far above. By the mechanism the 10.8M seeds establish, RN's held-out ratio should fall once memorization starts, most for whoever memorizes most, and Adam memorizes most. Adam's flat *training* ratio we do not explain.

One limit stands. 1.25 passes is not the same regime as one, so "more budget" here also means "some memorization," which is why we separate the training and held-out readings above; no number of seeds removes it.

Two protocol notes. First, the β_1 sweep requires $\beta_2 > \beta_1$: with $\beta_2 = 0.95 < \beta_1 = 0.999$ the runs *diverge* (weights to 2.1×10^6, then NaN) for reasons that have nothing to do with precision, and the frozen fraction is undefined on a diverged run. The four such runs are excluded. Pooling them in would have *lowered* the reported correlation from $r = 0.991$ to $r = 0.938$. The exclusion is therefore not flattering the diagnostic; it removes runs on which the measured quantity is undefined. Second, the RN/SR loss ratios here (1.07–1.18× at 10.8M and 0.99–1.03× at 124M) are far smaller than the 0.84–322× span of Table 2; the MLP cells there are run to stationarity, whereas a few thousand steps of language modelling is transient, where floors have not yet bound. This is the same caveat we record for the full-MNIST cells.

The same grid on a corpus that cannot be memorized (`exp33_wt103_124m.py`). TinyStories is synthetic, small-vocabulary and highly redundant, and this appendix blames the 10.8M held-out reversal on exactly that. So the 124M measurements are open to the objection that they inherit the confound. We repeated the whole grid on WikiText-103, holding everything else fixed: the same model, the same $3000 \times 8 \times 1024 = 25$M-token budget, the same 58.8M-token training split (the corpus truncated to a 60.0M-token file and split 98/2, so tokens-per-pass is identical), the same lr grids, warmup, cosine decay and clipping, and the same three seeds on the tuned cells. Only the corpus changes. The manipulation does what it should: the fp32-master arm's held-out/train gap falls from 1.049/1.022/1.03 to 1.029/1.018/1.010 (Adam/SGD/sign), i.e. there is measurably less to memorize.

	training loss		held-out loss	
RN/SR, 3 seeds	TinyStories	WikiText-103	TinyStories	WikiText-103
Adam	1.0267 (5.0)	1.0256 (6.5)	1.0245 (22.7)	1.0221 (6.2)
SGD	1.0127 (4.2)	1.0076 (9.6)	1.0113 (3.4)	1.0079 (13.5)
sign	0.9935 (1.6)	0.9963 (3.6)	0.9929 (2.0)	0.9965 (2.3)

Parenthesised numbers are seed standard deviations from 1. Every claim the 124M runs carry survives: the ordering Adam > SGD > sign holds in *every* seed on both losses; RN leaves 25 of the 148 tensors at $\max|w| = 1.0000$ exactly (Adam and sign; 24 for SGD) while SR pins *none* in any of the three and the master pins 0–5; and SR still recovers the master, at 1.0009 ± 0.0040 (Adam), 0.9994 ± 0.0017 (SGD) and 1.0038 ± 0.0008 (sign) on training loss. The tuned step sizes come out identical to TinyStories' (6×10^{-4}, 1.0, 3×10^{-4}), so the comparison is not confounded by re-tuning.

Two things are worth stating against ourselves. Memorization *was* inflating the effect, slightly: SGD's ratio falls by about 40% of its excess over 1 and Adam's by 4%, and sign moves toward 1 — the direction the regularizer reading predicts. And the reversal itself does not appear at this scale on either corpus, which is consistent with what we already report, that it is a 10.8M phenomenon. What the higher-entropy corpus buys is resolution: the seed spreads shrink, and sign's floor sitting *below* 1 goes from 1.6 to 3.6 seed standard deviations, i.e. from not clearing 1% to a measurement. The weight-growth advantage Corollary 6.5 allows for RN is easier to see when the corpus is not doing regularization of its own.

What the diagnostic is competing with: error feedback and 8-bit optimizer states (`exp36_compensation_baselines.py`). Section 9 recommends compensating the flagged tensors and preferring SR, and Appendix G.2 compares that against blanket SR, blanket RN, an fp32

master and random tensor sets. It does not compare it against the two things practitioners actually reach for when low-precision training stalls: error feedback — carrying the discarded low half forward, $v = w + \Delta + r$, $w' = \mathrm{RN}(v)$, $r \leftarrow v - w'$, which is the compensated-summation recipe of Zamirai et al. (2020) and the weight-storage analogue of Richtárik et al. (2021); Nikdan et al. (2026) — and block-wise 8-bit optimizer states (Dettmers et al., 2022). Both are run here on WikiText-103 at the lr EXP33 tuned for Adam–RN, three seeds, so the RN, SR and master arms of that grid are the controls and the whole table sits on one corpus.

storage scheme	B/param	train	held-out	train, relative to RN
RN (uncompensated)	2	4.8312 ± 0.0663	4.9575 ± 0.0490	1.0000
SR	2	4.7104 ± 0.0465	4.8501 ± 0.0381	0.9750
fp32 master	6	4.7064 ± 0.0568	4.8427 ± 0.0508	0.9742
error feedback	4	4.7113 ± 0.0476	4.8482 ± 0.0405	0.9752
8-bit state, RN weights	2	4.8256 ± 0.0670	4.9539 ± 0.0428	0.9988
SR on the flagged 25	2	4.7666 ± 0.0561	4.9004 ± 0.0422	0.9866

Paired per-seed differences against all-RN: SR -0.1208 ± 0.0199 (6.1 sd), master -0.1248 ± 0.0225, error feedback -0.1199 ± 0.0273, flagged 25 -0.0646 ± 0.0122, 8-bit state -0.0056 ± 0.0224 (0.2 sd). Three readings.

Error feedback recovers what SR recovers, and costs twice as much to do it. The two are indistinguishable at three seeds and both sit on the master. That is the expected result, since both remove the bias RN injects, but the memory column is the point. The carried residual is a second buffer per parameter, so at bf16 error feedback costs 4 B/param against the master recipe's 6 and SR's 2: SR costs an RNG draw and no memory at all. For *weight storage* specifically, the compensation the paper recommends is the cheap one.

8-bit optimizer states are orthogonal, not competing. They move the loss by 0.2 seed standard deviations, because they attack a different memory: quantizing the moments does nothing about the rounding of the weight on write-back, which is what freezes a coordinate. They compose with SR rather than replacing it, and a practitioner short of memory should expect to want both. The second moment needs a logarithmic code, which is what dynamic quantization buys. A linear absmax code cannot be used for it: the small entries of a block go to zero, the update then divides by ϵ, and the run diverges (loss $\sim 3 \times 10^3$ at 2800 steps).

The selective arm is efficient but does not beat blanket SR here. It recovers 53% of the RN–SR gap while touching 0.0154% of the parameters. On TinyStories (Appendix G.2) the same protocol recovered *twice* what blanket SR did; that comparison does not replicate on this corpus, and we report the discrepancy rather than the more flattering number. What does replicate exactly is the diagnostic itself: run on WikiText-103 it selects the same 25 LayerNorm gains, 25/25, with $\phi < 1$ fraction 1.0000 against 0.547 for the next tensor down — the cliff is in the same place on both corpora, so what is corpus-dependent is how much compensating those tensors buys, not which tensors to compensate.

Adam has no floor theorem, but the freeze margin still orders the damage (`exp37_adam_hparam_floor.py`). Appendix G.3 records that Adam enters only through the momentum propositions and the experiments. What a floor theorem for it would have to reproduce is the mechanism sketched there: Adam's normalized update has dither scale $\eta\sqrt{(1-\beta_1)/(1+\beta_1)}$ rather than $\eta\sigma$, so its exposure should be set by its own hyperparameters and not by the gradient noise. That is testable without the theorem, because the freeze margin ϕ is computable for any optimizer: vary the three hyperparameters that set the update scale, at fixed lr and corpus, and ask whether the cells with more exposure carry the larger RN penalty. On WikiText-103 at the lr EXP33 tuned for Adam–RN:

β_1	β_2	ϵ	median ϕ	$\phi < 1$ fraction	RN/SR (train)
0.9	0.999	10^{-8}	0.528	0.752	1.0133
0.9	0.95	10^{-8}	0.386	0.808	1.0259
0.9	0.95	10^{-5}	0.367	0.832	1.0238
0.95	0.95	10^{-8}	0.275	0.860	1.0432
0.99	0.95	10^{-8}	0.110	0.956	1.0485

Ranked by exposure the penalty is monotone but for one adjacent pair (the ϵ cell): Spearman is $+0.900$ between the $\phi < 1$ fraction and the ratio, and -0.900 between median ϕ and the ratio. The excess over 1 moves by $3.65\times$ across the grid (0.0133 to 0.0485) with the learning rate, the corpus and the budget all held fixed. Raising β_1 shrinks the update relative to the weight and costs the most; raising β_2, which lets the denominator adapt more slowly and leaves larger normalized steps, is the one cell that is *less* exposed than the default and it carries the smallest penalty.

The $(\beta_1, \beta_2) = (0.999, 0.95)$ cell is excluded: it ends at 52.5 (RN) and 25.1 (SR) nats against a chance loss of 10.82 for this vocabulary, the $\beta_1 > \beta_2$ divergence this appendix already documents, and a ratio taken between two diverged runs measures nothing. Its exposure is the highest in the grid ($\phi < 1$ at 0.981), so including it would have strengthened the correlation; we drop it because the runs are invalid, not because of what it would do to the number.

This does not substitute for a floor theorem, and we do not claim it does. It establishes the weaker thing the theorem would have to explain: across Adam's own hyperparameters the diagnostic, not the gradient noise, is what predicts which configuration storage rounding hurts.

Why the $O(uL/\sigma)$ correction is not measurable here. Theorem 6.3's upper bound carries a relative correction $O(uL/\sigma)$. Over the grid of Figure 3 ($M \in [9, 14]$, $\sigma \in \{0.25, 0.5, 1\}$) that quantity runs from 7.8×10^{-3} down to 6.1×10^{-5}, so the predicted *absolute* shift in $V/(u\sigma/L)$ spans 4.9×10^{-3} to 3.8×10^{-5}, against a mean seed spread of 5.0×10^{-3}. In only 5 of the 36 cells, the coarsest mantissas at the smallest σ, does the predicted signal clear its own seed spread. A log-log fit of $|V/(u\sigma/L) - C^\star|$ against uL/σ returns slope 0.22, CI $[-0.10, 0.54]$, $r^2 = 0.05$: the fit is measuring seed noise, and neither confirms nor refutes the predicted slope of 1. Resolving it needs roughly $10^2\times$ the replica budget at the fine end, which we did not spend because the constant, not the rate, is the claim. (One bookkeeping note for anyone reading the released logs: the `predicted_slope` field in `exp6_constant.json` records 0.5, which is the mislabelling this correction fixes. We left the emitted file as the run produced it and corrected the script, the figure and the text; the theorem predicts 1.)

Two protocol artifacts that both bias the exponent below one. Measuring a *stationary* floor by tuning over a step-size grid is more delicate than it looks, and we hit two distinct traps that push the fitted V-vs-u exponent below the true value of 1. Both are worth recording because both are invisible in the fitted line and only show up when the protocol is varied.

(i) Freeze at initialization. With an initial error e_0 only a few ulps from the optimum, the smallest step size in the grid freezes the run bitwise (Theorem 4.2), and the tail average equals e_0^2, a number the tuner happily reports as the best "floor." Reproduced under the old protocol (`exp20_protocol_artifacts.py`): at M=7, e_0=0.02 gives $e_0/u = 2.6$, the tuner selects the grid's smallest step $0.15u$ (just under the predicted freeze cliff $u/(2(Le_0 + 3\sigma)) = 0.166u$), and the cell is bitwise frozen with $V/e_0^2 = 0.95$: frozen at its starting point, not at a floor. The same run gives $0.61, 1.37, 0.95, 1.15$ at $M = 5, 6, 7, 8$, all with no bitwise movement, and the artifact switches off at M=9 where the tuner first moves to $0.20u$. A grid that simply starts above the cliff never sees it, which is why the failure is invisible in a fitted line. The artifact decays as e_0/u grows with M, manufacturing a spurious rising trend in $V/(u\sigma)$. Fix: initialize at $e_0 = 0.2$, so $e_0/u \geq 25$ across the range and a frozen run scores far above any real floor; we also report a per-cell bitwise-movement check.

(ii) Un-converged tails. The contraction time $1/(\eta L)$ grows like $1/u$, so at a *fixed* horizon the fine-M cells are still in the transient. Worse, the tuner then *avoids* the small-η columns for exactly that reason, and the reported $\eta^\star$ drifts upward with M. In the same reproduction, at $T = 1.5 \times 10^5$ the slowest column gets 176 contraction times at M=7 but only 2.7 at M=13, $\eta^\star$ drifts $0.20u \to 0.20u \to 0.30u \to 0.45u$ over $M \in \{7, 9, 11, 13\}$, and the fitted exponent comes out at 0.839 instead of 1, from this artifact alone, with the freeze-proof e_0 already in place. Fix: set T to a fixed number of contraction times of the slowest η in the grid (120 here), and report the ratio of the second half of the averaging window to the first, which must be ≈ 1.

With both fixed, the four homogeneous exponents are $0.99/0.99/1.02/1.01$ with every 95% seed-bootstrap CI containing 1 and tail drifts in $[0.93, 1.22]$; an independent fixed-η check ($\eta = 0.3u$, horizons up to 320 contraction times) gives $V/(u\sigma) = 0.24$ flat in both M and T. Under the earlier

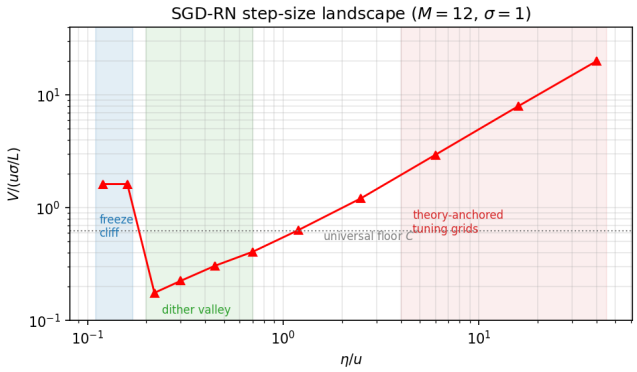

Figure 4: Homogeneous SGD–RN at $M{=}12$, $\sigma{=}1$: freeze cliff, dither valley, variance ramp. The shaded right region marks theory-anchored tuning grids that miss the valley.

fixed-horizon protocol the same measurement returned exponents of 0.66–0.90 depending on the M range, which is what a reader re-running an under-specified protocol would get.

The SGD–RN step-size landscape and the tuning-protocol pitfall. Figure 4 shows the tuned-floor landscape of homogeneous SGD–RN at $M{=}12$, $\sigma{=}1$: the freeze cliff at the predicted $\eta = 0.166u$, the dither valley ($V \approx 0.2\,u\sigma/L$ around $\eta \in [0.2, 0.7]u$), and the variance ramp. Grids anchored at the $2/3$-balance point (16–$40u$ here) miss the valley by more than an order of magnitude in η and inflate the measured floor by 20–$114\times$; an early version of our own protocol did exactly this. Every tuned-floor claim in this paper uses grids spanning the valley.