WHERE, WHEN, AND HOW MANY BITS? SCALING LAWS FOR SGD WITH STOCHASTIC ROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-precision training with stochastic rounding has become popular for large language models (LLMs). However, there is currently no theory to guide how much precision each matrix product needs or how it should be scheduled over training. We study one-pass SGD on random-feature regression with power-law spectra, where the features in the forward propagation and the weight-gradient products are stochastically rounded with time-varying precision. We derive an exact recursion for the rounded loss curve and scaling laws for the extra loss due to rounding. For the final weights of a training run, performed at a constant learning rate, these laws answer three questions. Where: rounding in the forward pass limits accuracy, while rounding in the weight-gradient product is hidden by the noise of SGD. When: the bits are best spent late in training. How much: the number of bits needed grows only logarithmically with compute budget, whereas a fixed precision lowers the exponent of the scaling law. Iterate averaging of the output and learning-rate decay remove the noise of SGD and thereby expose the rounding error in the weight-gradient product. This indicates that the learning-rate and precision schedules must be designed together.

1 INTRODUCTION

Training large machine learning models in low precision is now standard practice to reduce memory usage and computational costs. For example, mixed-precision training computes matrix products of the forward and backward passes in 16-bit arithmetic, while keeping a high-precision copy of the weights (Micikevicius et al., 2018). Recently, advanced LLMs are even pretrained in 8-bit and 4-bit floating-point formats (Micikevicius et al., 2022; DeepSeek-AI, 2024; Tseng et al., 2025; Castro et al., 2025; Chmiel et al., 2025; Panferov et al., 2026). Meanwhile, quantization-aware training (QAT) also employs low-precision arithmetic during training to ensure that the trained model can run in low precision at inference (Jacob et al., 2018; Esser et al., 2020; Liu et al., 2024). A third setting, activation-compressed training, keeps the forward pass in full precision but stores the activations needed for the backward pass in two to four bits with stochastic rounding (Chen et al., 2021; Liu et al., 2022); Chen et al. (2020) analyze the variance that quantized gradients add to training.

In this paper, we focus on the first setting, in which the arithmetic of training itself is rounded. In this case, stochastic rounding (SR), which rounds a number up or down randomly, has become a popular tool. This is because the value rounded via SR is unbiased, keeping training accurate at precisions where round-to-nearest stalls (Courbariaux et al., 2015; Gupta et al., 2015; Zhou et al., 2016; Zhang et al., 2017; De Sa et al., 2018; Zhao et al., 2025; Tseng et al., 2025; Ozkara et al., 2025). Existing theory on SR provides probabilistic bounds on rounding errors in numerical computations (Higham & Mary, 2019; 2020; Connolly et al., 2021; Croci et al., 2022; El Arar et al., 2023; Drineas & Ipsen, 2024; El Arar et al., 2026) and convergence guarantees for low-precision SGD (Li et al., 2017; De Sa et al., 2018; Ozkara et al., 2025). It also shows that SR can increase the small singular values of a matrix (Boutsikas et al., 2024; Dexter et al., 2025; Ma et al., 2026). However, these numerical guarantees do *not* by themselves determine how to allocate precision across operations and training steps to control the loss. In this paper, we aim to bridge this gap and answer the following question:

Where, when, and with how many bits should we stochastically round during training?

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.

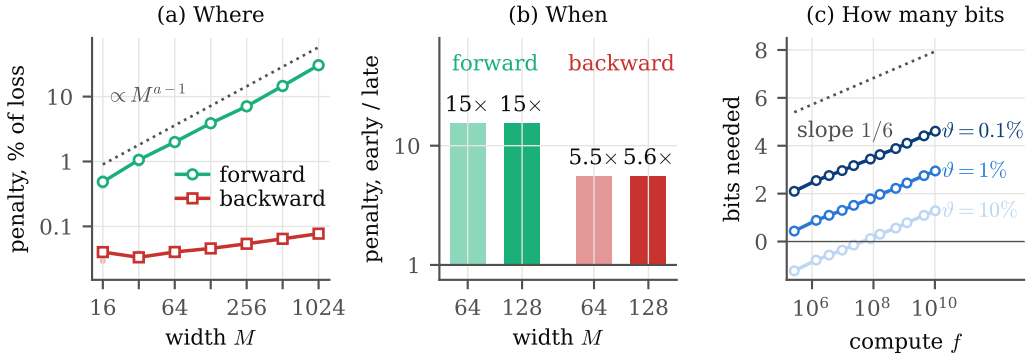


Figure 1: *Where* (a): rounding the forward copy of the features to one bit costs the last iterate a share of its loss that grows with the width like M^{a-1}; rounding the backward copy costs less than 0.1% at every width (exact recursion of Theorem 1, $(a, b) = (2, \frac{3}{2})$, horizon $\tau_T = M^a$; Appendix G). *When* (b): the same bits are worth more late in the run (last iterate, $M = 64$ and 128): the schedule "three bits, then one" costs about 15 times (forward copy) and 5.5 times (backward copy) the penalty of "one bit, then three". *How many bits* (c): the mantissa bits that keep the penalty below a fraction ϑ of the loss grow like the logarithm of the compute f, one bit per six doublings at every ϑ (Theorem F.8(i)).

In a parallel line of work, neural scaling laws describe how the loss of neural networks decreases as a power of the model size, the amount of data, and the amount of compute (Kaplan et al., 2020; Hoffmann et al., 2022; Subramanian et al., 2026); and these laws can be used to determine how compute is spent before an LLM is trained. Recent empirical work extends these laws to low-precision training and model quantization (Kumar et al., 2025; Sun et al., 2025; Frantar et al., 2025; Chen et al., 2025), treating precision as an additional axis of a fitted loss surface. These findings motivate a theoretical study of how the rounding mechanism gives rise to precision-dependent loss scaling. We develop such an analysis of the scaling laws for SR to answer our target question, complementing recent theoretical work on training dynamics under quantization (Zhang et al., 2025; 2026).

Another line of work aims to explain neural scaling laws theoretically in simple solvable models (Bahri et al., 2024; Maloney et al., 2022; Bordelon et al., 2024) and derive compute-optimal scaling laws for SGD (Paquette et al., 2024; 2025; Lin et al., 2024; 2025; Bordelon & Mori, 2026; Chen et al., 2026), all in exact arithmetic. One such solvable model (Paquette et al., 2024; Lin et al., 2024; Bordelon & Mori, 2026) is a one-pass mini-batch SGD on random-feature (sketched) linear regression with power-law spectrum. In this model, the eigenvalues of the data covariance and the coefficients of the target decay as powers, and a frozen random feature map sets the width of the model. We use this model and apply SR at every SGD step, with a precision that may differ between rounding locations and that may change over training. In our analysis, we round the two copies of the feature vector at each SGD step: the copy in the forward product, which produces the prediction, and the backward-pass copy in the weight-gradient product. In a network, these are the activations that enter the forward matrix product of a linear layer and the weight-gradient product in backpropagation; activation-compressed training rounds exactly the second copy (Chen et al., 2021; Liu et al., 2022).

Our contributions answer the question above as follows.

- **Exact dynamics and scaling laws for rounded SGD** (§3). We derive an exact finite-dimensional recursion for the expected loss of rounded SGD under arbitrary width, batch size, and deterministic learning-rate and precision schedules. Using this recursion, we obtain precision-dependent scaling laws for the additional loss caused by rounding scales with model width and training time.
- **Where to round** (§4.1). We show that forward and backward SR affect training differently. The rounding location that limits accuracy depends on the output rule and the level of SGD noise.
- **When to round** (§4.2). We characterize how the effect of rounding depends on when it occurs during training and derive time-varying precision allocations for different outputs.
- **How much precision is needed.** (§4.3). We derive the number of bits required as a function of model width, training time, and compute. The required precision grows only logarithmically with these scales, while fixed precision can change the compute-scaling exponent.
- **Precision and learning-rate schedules interact** (§4.4). A cooldown learning rate schedule removes SGD noise but only a bounded fraction of rounding noise. We characterize this interaction and show how lower precision changes the optimal learning-rate schedule.

Figure 1 previews the three answers. Appendix A reviews related work; the proofs, the experimental details and the extensions beyond the model are in Appendices B–H.

2 Model, stochastic rounding, and rounded SGD

Data, target, and random features. Let $d \geq M$. M is the number of random features, the width. The inputs are Gaussian, $x \sim N(0, \Sigma_d)$ with $\Sigma_d = \mathrm{diag}(\lambda_j)$ and $\lambda_j \asymp j^{-a}$ for an exponent $a > 1$. The target $\theta_* \sim N(0, C_\theta)$, $C_\theta = \mathrm{diag}(\vartheta_j)$, is independent of the data and satisfies $\lambda_j \vartheta_j \asymp j^{-b}$ for an exponent $b > 1$. These are the capacity and source conditions. The labels are $y = x^\top \theta_* + \epsilon$ with label noise $\epsilon \sim N(0, \sigma^2)$. A Gaussian feature map $S \in \mathbb{R}^{M \times d}$ with i.i.d. $N(0, 1/M)$ entries is drawn once and frozen, the features are $z = Sx \in \mathbb{R}^M$, and the predictor is $f_v(x) = v^\top z$ with trainable weights $v \in \mathbb{R}^M$. The covariance of the features and its eigendecomposition are

$$G = S\Sigma_d S^\top = \sum_{j=1}^{M} g_j \phi_j \phi_j^\top, \quad g_1 \geq \cdots \geq g_M, \tag{1}$$

with orthonormal eigenvectors $\phi_1, \ldots, \phi_M$. Define ϕ_j the j-th *mode* of the features. The eigenvalues inherit the power law of the input, $g_j \asymp j^{-a}$ for $j \leq M/C$ (Lin et al., 2024); this can be motivated by prior work (Martin & Mahoney, 2019; 2020; Martin et al., 2021).

We write the best predictor on the features, its approximation error and the effective noise level as

$$v_*^S = G^{-1} S \Sigma_d \theta_*, \qquad A_S = \tfrac{1}{2} \theta_*^\top \big(\Sigma_d - \Sigma_d S^\top G^{-1} S \Sigma_d\big) \theta_*, \qquad \sigma_{\mathrm{eff}}^2 = \sigma^2 + 2A_S. \tag{2}$$

Given (S, θ_*), $y = z^\top v_*^S + \tilde{\epsilon}$ with $\tilde{\epsilon}$ independent of z and of variance σ_{eff}^2. So, training on the features is a well-specified linear regression in $\mathbb{R}^M$. Our metric throughout is the population excess risk of a predictor on exact inputs,

$$\mathcal{E}(v) = \mathbb{E}_{x,\epsilon} \tfrac{1}{2} (v^\top S x - y)^2 - \tfrac{\sigma^2}{2} = A_S + \tfrac{1}{2} \|v - v_*^S\|_G^2. \tag{3}$$

The first term is the approximation error of the feature map, which no choice of v removes. The second term is the error of estimating v_*^S, measured in the geometry of the features.

Stochastic rounding (SR). Let $\mathbb{F}_p = \{0\} \cup \{\pm 2^e (1 + k 2^{-p}) : e \in \mathbb{Z},\ 0 \leq k < 2^p\}$ be the binary floating-point grid with p mantissa bits and an unbounded exponent, e.g., FP8-E4M3 has $p = 3$ and $p = \infty$ is exact arithmetic. For $x \notin \mathbb{F}_p$ let $\ell_p(x) < x < u_p(x)$ be its two neighbors on the grid. The stochastic rounding $Q_p(x)$ equals $u_p(x)$ with probability $(x - \ell_p(x))/(u_p(x) - \ell_p(x))$ and $\ell_p(x)$ otherwise (the *rounding decision*), and it leaves a point of the grid unchanged; in particular $\mathbb{E}[Q_p(x)] = x$. The variance of the rounding error is

$$\nu_p(x) = \mathrm{Var}[Q_p(x)] = (x - \ell_p(x))(u_p(x) - x) \leq \tfrac{1}{4}\, 4^{-p} x^2, \tag{4}$$

which vanishes on the grid. 4^{-p} is an envelope, not a constant relative variance. A vector is rounded entrywise, with independent rounding decisions for its entries.

The rounded SGD training dynamics. We train f_v by one-pass mini-batch SGD with deterministic step sizes η_t, starting from $v_0 = 0$: at step t we draw a fresh batch of B samples (x_ℓ, y_ℓ), $\ell = 1, \ldots, B$, and form their features $z_\ell = S x_\ell$. Two copies of each feature vector are rounded independently,

$$\begin{aligned} a_\ell &= Q_{p_{\mathrm{fw},t}}(z_\ell) \qquad \text{(the } \textit{forward copy}\text{, which enters the prediction)}, \\ b_\ell &= Q_{p_{\mathrm{bw},t}}(z_\ell) \qquad \text{(the } \textit{backward copy}\text{, which enters the gradient)}. \end{aligned} \tag{5}$$

The rounding decisions are independent across entries, samples, steps and the two copies (Zhang et al., 2017). The two copies model low-precision neural-network training, where the activations used in the forward matrix product and the weight-gradient product are rounded separately. The update is

$$v_{t+1} = v_t - \frac{\eta_t}{B} \sum_{\ell=1}^{B} b_\ell \big(a_\ell^\top v_t - y_\ell\big), \qquad t = 0, \ldots, T-1, \qquad N = BT, \tag{6}$$

A *precision schedule* is a deterministic sequence $\{(p_{\mathrm{fw},t}, p_{\mathrm{bw},t})\}_{t=0}^{T-1}$. We write $q = 4^{-p}$ for the relative-variance scale of p bits, with q_{fw} and q_{bw} at the forward and backward precisions, and $\varrho = \eta/B$ for the noise scale of SGD at step size η.

Outputs and the rounding penalty. For a learning-rate schedule $\{\eta_t\}_{t=0}^{T}$ we evaluate two outputs, the last iterate v_T and the uniform average $\bar{v}_T = \frac{1}{T}\sum_{t\le T} v_t$. Averaging is the classical variance-reduction device of Polyak & Juditsky (1992), used in low-precision training as stochastic weight averaging (Yang et al., 2019) and as an exponential moving average (Morales-Brotons et al., 2024). We also evaluate the last iterate after a learning-rate *cooldown*: $\eta_s = \eta_{\text{peak}}$, for $s \le T_1$, and then

$$\eta_s = \eta_{\text{peak}}\Big[1 - (1-\zeta)\,\frac{s-T_1}{T_c}\Big], \qquad T_1 < s \le T, \qquad \tau_c := \sum_{s>T_1} \eta_s = \phi\,\tau_1. \tag{7}$$

We call ϕ the relative length of the cooldown and ζ its final ratio, and we write v_{out} for the output, $\text{out} \in \{\text{last}, \text{avg}, \text{cooled}\}$. The quantities we study are excess risks for rounded runs and *rounding penalty* defined below, respectively,

$$\mathcal{E}^{\text{out}}(t) = \mathbb{E}\big[\mathcal{E}(v_{\text{out}}) \mid S, \theta_*\big] - A_S, \qquad \Delta^{\text{out}}(t) = \mathcal{E}^{\text{out}}(t) - \mathcal{E}_0^{\text{out}}(t), \tag{8}$$

where $\mathcal{E}_0^{\text{out}}$ denotes the same output of the exact-arithmetic run, $p = \infty$, with the same schedules.

Rounding the input. The input x_ℓ may also be stored at p_x bits before the feature map, with one rounded copy shared by the two products or one copy per product in (5). Appendix C considers these two input rounding strategies. The last row in Table 1 of §4 also studies the input rounding.

The regime and SGD noise. We focus on the regime $1 < b < a$, in which the norm of the best predictor grows with the width, $\|v_*^S\|^2 \asymp M^{a-b+1}$ while $\|v_*^S\|_G^2 \asymp 1$. After the optimization time $\tau_t = \sum_{s<t}\eta_s$, SGD has fitted the target on the first $k_t = \min(M, \tau_t^{1/a})$ modes of the features, denoted by the *effective dimension*. Notice that $\tau_T \asymp M^a$ at which $k_T = M$. At a constant step size the excess risk of the last iterate stops at the *SGD-noise floor* $\asymp \varrho\sigma_{\text{eff}}^2$, which averaging or a learning-rate cooldown removes on the learned modes. Define the *noise ratio* $\text{byn} := \varrho\sigma_{\text{eff}}^2/M^{1-b}$.

3 From the exact dynamics to scaling laws with precision

In this section, we derive the exact dynamics of the rounded SGD, establish an approximated dynamics, and use these to eventually conclude scaling laws with precision.

The exact dynamics of the rounded algorithm. The expected loss curve of rounded SGD obeys an exact recursion, for any width, batch size, step-size schedule and precision schedule. We now show that SR adds three terms to its SGD noise. Let $\mathbb{E}$ average over the fresh inputs, the label noise and the SR randomness, conditional on (S, θ_*); $\delta_t = v_t - v_*^S$ is the error of the iterate, $\tilde{\mu}_t = \mathbb{E}\delta_t$, $C_t = \mathbb{E}[\delta_t\delta_t^\top]$ and $W_t = \mathbb{E}[v_t v_t^\top]$ are its mean, its second moment and the second moment of the iterate, and $H_t = I - \eta_t G$ is one step of exact-arithmetic gradient descent on the error. SR induces two scalars, the mean-squared residual $\mathsf{R}_t = \sigma_{\text{eff}}^2 + \text{tr}(GC_t)$ and the forward rounding variance F_t, and four diagonal matrices $M_{r,t}$, $A_{\text{fw},t}$, $\Lambda_{\text{bw},t}(\cdot)$, $\Delta_{\text{fwbw},t}$, $r \in \{\text{fw}, \text{bw}\}$, built from three Gaussian moments of the SR error variance at the feature variances G_{kk}, (25) and (27) in Appendix B.

Theorem 1 (exact second-moment recursion)**.** *Consider the rounded SGD loop* (6) *with deterministic schedules* $(\eta_t, p_{\text{fw},t}, p_{\text{bw},t})$ *and* $v_0 = 0$*. Conditional on* (S, θ_*)*, the mean error follows exact-arithmetic gradient descent,* $\tilde{\mu}_{t+1} = H_t\tilde{\mu}_t$*, and the second moment satisfies*

$$\begin{aligned} C_{t+1} = H_t C_t H_t + \frac{\eta_t^2}{B}\Big[&\underbrace{\mathsf{R}_t G + G C_t G}_{\textit{SGD noise}} + \underbrace{\mathsf{R}_t M_{\text{bw},t} + 2\Lambda_{\text{bw},t}(GC_tG)}_{\textit{backward rounding}} \\ &+ \underbrace{\mathsf{F}_t G + 2GA_{\text{fw},t}G}_{\textit{forward rounding}} + \underbrace{\mathsf{F}_t M_{\text{bw},t} + \Delta_{\text{fwbw},t}}_{\textit{interaction}}\Big], \end{aligned} \tag{9}$$

closed by $W_t = C_t + v_*^S v_*^{S\top} + v_*^S\tilde{\mu}_t^\top + \tilde{\mu}_t v_*^{S\top}$ *(only* $\text{diag}\, W_t$ *enters), with* $C_0 = v_*^S v_*^{S\top}$*. The expected excess risk is* $\mathbb{E}[\mathcal{E}(v_t) \mid S, \theta_*] = A_S + \frac{1}{2}\text{tr}(GC_t)$.

We prove Theorem 1 in Appendix B.3. The mean follows exact-arithmetic gradient descent whatever the precisions, so SR adds no bias. Raising either precision at any step can only lower the expected risk of every fixed-weight output, with no condition on the step sizes (Corollary B.5). The forward term is driven by the size of the parameters, the backward term by the size of the residual, and the interaction term is present only when both copies are rounded.

The deterministic-equivalent (DE) dynamics. We reduce the $M \times M$ recursion (9) to one scalar recursion per mode. Fix constant *controls* $(\eta, B, p_{\rm fw}, p_{\rm bw})$ and average the recursion over the target, $\bar{C}_t = \mathbb{E}_{\theta_*} C_t$, $\bar{W}_t = \mathbb{E}_{\theta_*} W_t$, $\bar{\sigma}^2_{\rm eff} = \mathbb{E}_{\theta_*}\sigma^2_{\rm eff}$. In the eigenbasis of G, $e_j(t) = \phi_j^\top \bar{C}_t \phi_j$ is the error energy in mode j and $w_j = e_j(0)$ the target energy there. The SGD-noise term of (9) is diagonal in this basis, whereas the rounding matrices are diagonal in the feature coordinates; replacing their projections on mode j by their averages over the coordinates (Lemma B.2) leaves three *rounding coefficients* $\iota_r, \bar{\alpha}_r, \gamma_{\rm fwbw}$. These are the coordinate averages of the mean SR error variance $m_{p_r}(G_{kk})$, of its second Hermite coefficient $\alpha_{p_r}(G_{kk})/G_{kk}$ and of the cross-covariance of the two copies' error variances ((26), Appendix B). From now on a plain symbol refers to the rounded algorithm, the subscript DE to the deterministic-equivalent dynamics, and the subscript 0 to exact arithmetic; a bar marks the state of the DE dynamics, whose mode energies $\bar{e}_j(t)$ approximate the $e_j(t)$. The *DE dynamics* is one scalar recursion per mode coupled through three sums, with $\bar{e}_j(0) = w_j$,

$$\begin{aligned}
\bar{e}_j(t{+}1) &= (1{-}\eta g_j)^2 \bar{e}_j(t) + \frac{\eta^2}{B}\Big[\big(\bar{\mathsf{R}}_t{+}\iota_{\rm fw}\mathsf{U}_t\big)\big(g_j{+}\iota_{\rm bw}\big){+}g_j^2\bar{e}_j(t){+}\frac{2g_j^2\bar{\alpha}_{\rm fw}}{M}\mathsf{U}_t{+}2\bar{\alpha}_{\rm bw}\bar{\mathsf{S}}_t{+}\frac{\gamma_{\rm fwbw}}{M}\mathsf{U}_t\Big] \\
&\bar{\mathsf{R}}_t = \bar{\sigma}^2_{\rm eff} + \sum_l g_l \bar{e}_l(t) \ \text{(mean-squared residual)}, \qquad \bar{\mathsf{S}}_t = \frac{1}{M}\sum_l g_l^2 \bar{e}_l(t), \\
&\mathsf{U}_t = \sum_l \big[w_l\big(1-2(1-\eta g_l)^t\big) + \bar{e}_l(t)\big] \ \text{(mean squared parameter norm)}.
\end{aligned} \tag{10}$$

Without rounding, (10) is the target-averaged exact-arithmetic recursion of Bordelon & Mori (2026). For each output rule out, $\mathcal{E}^{\rm out}_{\rm DE}(t)$ is the excess risk that (10) predicts for the output ($\mathcal{E}^{\rm last}_{\rm DE}(t) = \frac{1}{2}\sum_j g_j \bar{e}_j(t)$; the average and the cooled last iterate are defined in Appendix D), $\Delta^{\rm out}_{\rm DE}(t) = \mathcal{E}^{\rm out}_{\rm DE}(t) - \mathcal{E}^{\rm out}_0(t)$ the predicted penalty, $\mathcal{E}^{\rm out}(t)$ and $\Delta^{\rm out}(t)$ the risk and the penalty of the algorithm (8), and $\mathcal{E}^{\rm out}_0(t)$ exact arithmetic. Let σ^2 stand for the target-averaged effective noise $\bar{\sigma}^2_{\rm eff} = \sigma^2 + 2\mathbb{E}_{\theta_*} A_S$.

Conjecture 2 (DE dynamics predict the penalty)**.** *Let S be the Gaussian feature map, $1 < b < a$, $d = \lceil c_d M \rceil$ with $c_d > 1$, $\sigma^2 > 0$ fixed, and let the controls $\eta, B, p_{\rm fw}, p_{\rm bw}$ be constant in time but allowed to depend on M, with $\eta g_1 \le c_0 < 1$. Then, for* out $\in \{\text{last}, \text{avg}\}$*, the relative error* $\sup_{M^a/10 \le \tau_t \le M^a} |\Delta^{\rm out}(t) - \Delta^{\rm out}_{\rm DE}(t)|/\Delta^{\rm out}(t)$ *tends to* 0 *in probability over S as $M \to \infty$.*

The cooled last iterate is not covered since the controls of the conjecture are constant in time. For only backward rounding, Theorem E.12 proves every schedule. Under a small-gain condition the transfer from (9) to (10) is Theorem D.6 . This relative error is checked empirically in Figure 3(b).

Assumption 3 (standing hypotheses for the scaling laws)**.** Consider $1 < b < a$. (a) $cj^{-a} \le g_j \le Cj^{-a}$ and $cj^{a-b} \le w_j \le Cj^{a-b}$ for $j \le M$. (b) $B \ge 1$, $0 < \sigma^2 \le \Sigma_{\max}$, $\eta g_1 \le c_0 < 1$. (c) $\iota_r \asymp q_r \,\mathrm{tr}\,\Sigma_d/M$, $\bar{\alpha}_r \asymp q_r$, $0 \le \gamma_{\rm fwbw} \lesssim q_{\rm fw} q_{\rm bw} (\mathrm{tr}\,\Sigma_d)^2/M$, with constants independent of M and of the precisions. (d) $\beta_0 := \varrho \sum_j g_j/\big(2 - \eta(1 + B^{-1})g_j\big) \le 1 - \kappa_0$. (e) $\Gamma(T) \le \gamma_* < 1$, with $\Gamma(T)$ defined in (68) of Appendix E. Constants below depend only on $(a, b, c, C, c_0, \mathrm{tr}\,\Sigma_d, \Sigma_{\max}, \kappa_0, \gamma_*)$; τ_* is a fixed burn-in time and the *window* is $\tau_* \le \tau_t \le M^a$.

Scaling laws with precision. Without SR, the dynamics obeys the known scaling law (Lin et al., 2024; Bordelon & Mori, 2026; Chen et al., 2026): under Assumption 3(a)(b)(d), $\mathsf{U}^0_t \asymp k_t^{a-b+1}$,

$$\mathcal{E}^{\rm last}_0(t) \asymp k_t^{1-b} + \varrho\,\sigma^2, \quad \mathcal{E}^{\rm avg}_0(t) \asymp k_t^{1-b} + \varrho\,\sigma^2\, k_t/\tau_t, \quad \mathsf{R}^0_t \asymp r_t := \sigma^2 + k_t^{1-b}, \tag{11}$$

with R^0_t, U^0_t the residual and parameter-norm sums without rounding (Lemma E.1). The risk of the last iterate is the target energy not yet fitted, k_t^{1-b}, plus the SGD-noise floor $\varrho\sigma^2$, which averaging replaces by the decreasing $\varrho\sigma^2 k_t/\tau_t$ and a cooldown removes on the learned modes. The residual scale r_t and the parameter-norm scale k_t^{a-b+1} drive the backward and the forward rounding noise. Below we prove the scaling law with precision is (11) plus a penalty made of a forward term, a backward term and, for the average, an interaction term.

Theorem 4 (scaling law with precision)**.** *Under Assumption 3, for $t \le T$ on the window,*

$$\mathcal{E}^{\rm out}_{\rm DE}(t) = \mathcal{E}^{\rm out}_0(t) + \Delta^{\rm out}_{\rm DE}(t), \qquad \begin{aligned}
\Delta^{\rm last}_{\rm DE}(t) &\asymp \varrho\Big[\underbrace{q_{\rm fw}\frac{k_t^{a-b+1}}{M}}_{\textit{forward}} + \underbrace{q_{\rm bw}\frac{r_t k_t}{M}}_{\textit{backward}}\Big], \\
\Delta^{\rm avg}_{\rm DE}(t) &\asymp \varrho\Big[\underbrace{q_{\rm fw}\frac{k_t^{2-b}}{M}}_{\textit{forward}} + \underbrace{q_{\rm bw}\frac{r_t k_t}{M}}_{\textit{backward}} + \underbrace{q_{\rm fw}q_{\rm bw}\frac{k_t^{a-b+2}}{M^2}}_{\textit{interaction}}\Big],
\end{aligned} \tag{12}$$

Table 1: Bits needed to keep (13) with ϑ and $\tau_T = M^a$ (Corollary F.1). A negative value means that the smallest mantissa precision suffices. $c(\phi, \zeta)$ is defined in (14), and the last row is from Proposition C.1. "theory": the formula at $M = 1024$, $\varrho = 1/64$, $\sigma^2 = 0.1$, $\vartheta = 0.1$, $(\phi, \zeta) = (1, 0.02)$; "measured": the exact recursion at the same parameters (cooled row at $M = 512$; Appendix G).

output	location	formula for p^ϑ	theory	measured
last	forward	$\frac{1}{2}\log_2\big(\varrho M^{a-1}/(\vartheta(1+\mathsf{n}))\big)$	1.9	1.8
last	backward	$\frac{1}{2}\log_2\big((\mathsf{n}+\varrho)/(\vartheta(1+\mathsf{n}))\big)$	-2.1	-2.5
cooled	backward	$\frac{1}{2}\log_2\big(c(\phi,\zeta)\,\mathsf{n}/\vartheta\big)$	-2.2 (0.3 at $B=1$)	-3.3
average	forward	$\frac{1}{2}\log_2(\varrho/\vartheta)$	-3.1	-4.5
average	backward	$\frac{1}{2}\log_2\big((\mathsf{n}+\varrho)/\vartheta\big)$	-2.0	-3.5
average	joint, $p_{\mathrm{fw}}+p_{\mathrm{bw}}$	$\frac{1}{2}\log_2(\varrho M^{a-1}/\vartheta)$	0.2	-1.7
any	input, shared copy	$\frac{1}{4}\log_2(M^{b-1}/\vartheta)$	0.3	–

with matching upper and lower bounds; the lower bounds need only Assumption 3(a)–(d).

We prove Theorem 4 in Appendix E. The forward term is aligned with G and grows with the parameter norm k_t^{a-b+1}; the backward term is isotropic, driven by the residual scale r_t; the interaction term is seen only by the average and carries $q_{\mathrm{fw}}q_{\mathrm{bw}}$. Under Conjecture 2, (12) implies neural scaling law for the actual algorithm, $\mathcal{E}^{\mathrm{out}}(t) = \mathcal{E}_0^{\mathrm{out}}(t) + (1 + o(1))\Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$ (Corollary D.7, Appendix D.8). This is the law that §4 turns into rules. Theorem E.12 in Appendix E.8 proves this law for backward rounding alone algorithm, for every schedule and both output rules, up to logarithmic factors in the width. What a learning-rate cooldown does to the backward term is the cooled law (14) of §4.1.

4 Where, when and how much to round

In this section, we turn the scaling laws of §3 into rules for training: where to apply SR, when low precision can be used, how much precision a limit on the width, the training time or the compute allows, and how precision and learning-rate schedules interact. Every rule is a consequence of Theorem 4, Conjecture 2 for the actual algorithm, or Theorem E.12 for backward-only SR. Below we state most of results informally with full statements and proofs in Appendix F.

Rules for training with stochastic rounding.

- A constant-rate last iterate is forward-limited: spend the bits on the forward copy (Proposition 5(i)).
- The appropriate precision depends on the output rule: an average, EMA, or a cooldown removes SGD noise, then needs more backward bits (Proposition 5(ii) and the cooled law (14)).
- Spend the bits late: for a last iterate both copies want more bits at the end of run ((15); Corollary F.7).
- Bits grow with the logarithm of compute; a fixed precision bends the compute-optimal law (Theorem F.8).
- Add another bit only while the reduction in rounding error is worth more than the training progress lost because that bit makes each step more expensive (Corollary F.9).
- Tune the learning-rate schedule together with precision: when aggressive low precision starts hurting, do not try only "more bits." Also try a smaller peak learning rate or an earlier, gentler decay. (Proposition F.10).

4.1 Where: which rounding location limits accuracy

We choose enough precision that SR adds only an acceptable fraction of the error,

$$\Delta_{\mathrm{DE}}^{\mathrm{out}}(T) \le \vartheta\, \mathcal{E}_0^{\mathrm{out}}(T), \tag{13}$$

at the final step T of a compute-optimal run, $\tau_T = M^a$, where ϑ is the *tolerance*, the constraint of Table 1 and of the compute law in §4.3. Corollary F.9 shows how to choose it. Table 1 lists all bit threshold formulas p^ϑ for each output rule and SR location: Corollary F.1 in Appendix F.1 proves (13) holds if $p \ge p^\vartheta + c_1$ for every requirement of the output rule and fails if $p \le p^\vartheta - c_2$ for one of them, with c_1, c_2 depending only on the constants of Assumption 3. As the width increases, the joint bits of an average and the forward bits of a last iterate grow like $\frac{a-1}{2}\log_2 M$, the backward bits of a cooled last iterate like $\frac{b-1}{2}\log_2 M$, and the other two stay bounded.

Below we show the location with the larger entry p^ϑ is the one whose term *dominates* the penalty.

Proposition 5 (which rounding location limits accuracy: the regime map)**.** *Let* $\mathsf{n} := \varrho\sigma^2/M^{1-b}$ *be the* noise ratio*. Compare the terms of* (12) *at* $\tau_T = M^a$ *and equal bits* $p_{\mathrm{fw}} = p_{\mathrm{bw}} = p$*.*

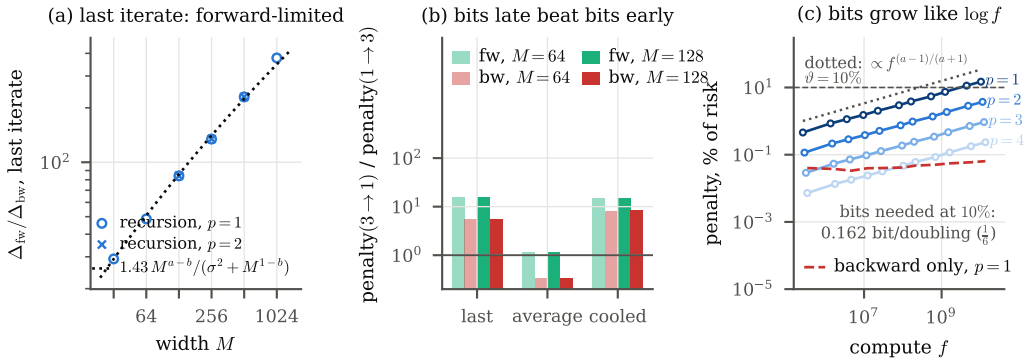


Figure 2: The precision rules for $a=2$, $b=\frac{3}{2}$, see settings in Appendix G. (a) Where: forward over backward penalty of the last iterate at one and two bits, against $1.43\,M^{a-b}/(\sigma^2+M^{1-b})$ of Proposition 5(i). (b) When: the penalty of "3 bits, then 1" over that of "1 bit, then 3" for each copy, output rule and width; above one, bits are worth more late. (c) How much: penalty of the last iterate as a fraction of the risk against compute, both copies at $p=1,\dots,4$ and the backward copy alone; the bits needed at tolerance 10% grow by 0.162 per doubling, against $\frac{1}{6}$. The allocation, the cooldown and the learning-rate schedule are in Figure 5 (Appendix G).

(i) Constant-learning-rate last iterate. *The forward term exceeds the backward term in* (12) *by the factor $M^{a-b}/(\sigma^2+M^{1-b})$, for every* n*: the last iterate is forward-limited. Moreover the backward penalty is at most $\approx 0.09\cdot 4^{-p_{\rm bw}}$ of the SGD-noise floor $\varrho\sigma^2$ of the exact-arithmetic law* (11)*. Thus backward rounding is hidden by SGD noise while the rate is constant.*

(ii) Average iterate. *The backward term is at least of the order of the forward term, the two are comparable iff $\sigma^2 \asymp M^{1-b}$, and the interaction term overtakes the backward term iff $\sigma^2+M^{1-b} \lesssim 4^{-p}\,\mathrm{tr}\,\Sigma_d\,M^{a-b}$, so that the average is then limited by the joint requirement on $q_{\rm fw}q_{\rm bw}$.*

Proposition 5 is proved in Appendix F.2. The last iterate requires enough forward precision so that rounding does not become the dominant source of error (at FP8-E4M3, $p=3$, the backward penalty of (i) is 0.15% of the SGD noise). Averaging and cooldown make backward SR more important relative to the remaining SGD noise, so a constant-η ablation need not predict the sensitivity after a cooldown. Figures 1(a), 2(a) and 4(c) test (i) and (ii) on the exact recursion.

The cooled row of Table 1 is the law of the cooled last iterate under backward SR alone, at the rate $\eta_{\rm peak}$ for optimization time $\tau_1=M^a$ and then along the cooldown (7) of relative length $\phi\geq 1$, with $\varrho_{\rm peak}=\eta_{\rm peak}/B$ and $\asymp_\sigma$. Proposition E.10 of Appendix E.7 proves that

$$\mathcal{E}_{\rm DE}^{\rm cooled}(T)=\mathcal{E}_0^{\rm cooled}(T)+\Delta_{\rm DE}^{\rm cooled}(T),$$
$$\Delta_{\rm DE}^{\rm cooled}(T)\asymp_\sigma \varrho_{\rm peak}\,q_{\rm bw}\,\sigma^2\,c(\phi,\zeta),\qquad c(\phi,\zeta)=\sqrt{\zeta^2+\frac{1-\zeta^2}{\phi}}\,. \tag{14}$$

The cooldown (7) removes the SGD-noise floor of $\mathcal{E}_0^{\rm cooled}(T)$ on the learned modes (Bordelon & Mori, 2026) but retains the fraction $\Upsilon(\phi)=(1+\phi)^{1/a}-\phi^{1/a}$ of the backward rounding noise added before it (0.41 at $\phi=1$, $a=2$; Figure 5(b)), so the backward bits for training with constant learning rate does not need become necessary once a cooldown removes the SGD noise (§4.4).

4.2 When: distributing the bits over a run

When during training can low precision be used? The answer is to spend the bits at the steps where rounding does the most damage to the output. Define $S_2(\chi_s)\asymp\chi_s^{1/a-2}$ as aligned and $S_1(\chi_s)\asymp\chi_s^{1/a-1}$ as isotropic noise (Lemma E.3). The *response weight* $W(s)=\eta_s^2\mathsf{U}_s S_2(\chi_s)$ for the forward copy and the last iterate. General definition of $W(s)$ is in (113) of Appendix F.

For a rounding location and an output rule, the penalty of a precision schedule $(p_s)_{s<T}$ is then $\sum_s W(s)\,4^{-p_s}$, and among precision schedules with mean $\bar p$ bits fixed, it is minimized by

$$p_s=\bar p+\tfrac{1}{2}\log_2\big(W(s)/\,\mathrm{GM}(W)\big),\qquad \frac{\text{penalty}(p_s)}{\text{penalty}(\bar p)}=\frac{\mathrm{GM}(W)}{\mathrm{mean}(W)} \tag{15}$$

(Proposition F.6, Appendix F.3), with $\mathrm{GM}(W)$ and $\mathrm{mean}(W)$ the geometric and arithmetic means of the weights over the steps. For constant η, the allocation is Corollary F.7. The last iterate wants the bits of both copies at the end of the run, since 99% of the forward weight sits in the last 10% of the steps, whereas an average iterate wants its backward bits early and its forward precision nearly constant. Figures 1(b), 2(b) and 5(a) test both on the exact recursion (§5).

4.3 How much: bits against width, training time and compute

Against the width the bits needed are the entries of Table 1, and against the training time at fixed width they grow at most like $\log_2 \tau_t$ (Corollary F.1). We focus on compute. Define $w(p) = (p+p_0)/(p_{\mathrm{ref}}+p_0)$, with p_0 the sign and exponent bits and p_{ref} the mantissa bits of a reference format (a copy that is not rounded is charged $w=1$). Define compute with precision as $f = BT\,M\,[w(p_{\mathrm{fw}})+w(p_{\mathrm{bw}})]$. At given precisions the compute-optimal width is $M_*(f) \asymp (f/\bar\kappa)^{1/(a+1)}$, $\bar\kappa = [w(p_{\mathrm{fw}})+w(p_{\mathrm{bw}})]/\varrho$, with $\tau_T = M_*^a$, and without rounding the compute-optimal loss is $\mathcal{E}_*(f) \asymp f^{-(b-1)/(a+1)}$ (Paquette et al., 2024). The compute scaling law for precision is Theorem F.8. Keeping tolerance ϑ in (13) requires $p^*_{\mathrm{fw}} = \frac{a-1}{2(a+1)}\log_2 f + O(1)$ forward bits for the last iterate (the same count for $p_{\mathrm{fw}}+p_{\mathrm{bw}}$ of the average), $p^*_{\mathrm{bw}} = \frac{b-1}{2(a+1)}\log_2 f + O(1)$ backward bits. The compute-optimal loss is $\asymp (f/\log f)^{-(b-1)/(a+1)}$: the exponent of $\mathcal{E}_*(f)$ is kept, and the logarithm is the cost of the growing bits. At fixed precisions, instead, the loss of the average iterate at matched width saturates at the floor $\Phi(p) = 16^{-p_x} + \varrho\, q_{\mathrm{fw}} q_{\mathrm{bw}} \sigma^2 + (\varrho\, q_{\mathrm{fw}} q_{\mathrm{bw}})^{(b-1)/(a-1)}$. With exact input, once f exceeds a threshold that grows with the precision, the best loss over widths decays at the exponent $(b-1)/(a+b)$ when only the backward copy is rounded and $(b-1)/(2a)$ when both are. At $(a,b) = (2, \frac{3}{2})$ this is one forward bit per six doublings of compute and one backward bit per twelve (Figures 1(c) and 2(c)), and a fixed precision changes the slope of the compute-optimal loss on a log–log plot, not only its offset, unlike the parallel shift of the fitted N_{eff} law of Kumar et al. (2025).

Under a wall-clock budget the tolerance follows from the trade-off between bits and steps: at the precision p^* that minimizes the total excess risk the remaining penalty is $\vartheta^* \approx e_0/(\ln 4\,(p^*+p_0))$ of the excess risk, with e_0 the local loss–time exponent of the exact-arithmetic risk, about 1.5% at $(a,b) = (2,\frac{3}{2})$, $p^* = 3$, $p_0 = 9$ (Corollary F.9, Appendix F.5). Table 1, the allocation (15) and Theorem F.8 combine into a explicit precision schedule in Algorithm 1 of Appendix F.7.

4.4 Precision and the learning-rate schedule

How should the learning-rate schedule change when rounding is present? In the optimal-control formulation of Bordelon & Mori (2026), the last-iterate loss of a learning-rate schedule is a bias term plus the noise added at every step weighted by a kernel of the remaining time χ, $S_2(\chi)$ in exact arithmetic; backward rounding adds the slower term $\iota_{\mathrm{bw}} S_1(\chi)$, which dominates the kernel once the remaining time exceeds the crossover $\chi_\times \asymp M/(q_{\mathrm{bw}}\,\mathrm{tr}\,\Sigma_d)$.

For only the backward copy rounded, Proposition F.10 proves a learning-rate schedule that minimizes the last-iterate loss decays as $(T-t)^{(a-1)/(a+1)}$ while the remaining time exceeds $\chi_\times$ and as the exact-arithmetic $(T-t)^{2a-1}$ when below $\chi_\times$: a gentler and longer cooldown. When rounding noise dominates SGD noise, one bit less lowers the optimal peak rate at fixed width and number of steps by the factor $4^{a/(a+b)}$. At fixed width and bits the best loss over schedules decreases with training time at the rate $(b-1)/(a+b)$ instead of $(b-1)/a$, which bits grow as $p_{\mathrm{bw}} \geq \frac{b}{2a}\log_2 \tau_T - \frac{1}{2}\log_2 M + O(1)$. Learning rate and precision must therefore be chosen together, because a rate that is effective without rounding amplifies the rounding noise. Full statement adds how the precision should change along the cooldown in Appendix F.6 and empirical validation is in Figure 5(c).

5 Empirical evaluation

In this section, we verify the theory and test the rules of the previous sections, in three settings: the exact recursion (9); the deterministic-equivalent dynamics (10) with the coefficients of the same feature map; and a bit-exact simulation of rounded SGD (6). Unless stated otherwise, $a=2$, $b=\frac{3}{2}$, $\sigma^2 = 0.1$, $B = 32$, the step size is constant and the run is compute-optimal, $\tau_T = M^a$. Here, we summarize our main findings. Additional details and further figures can be found in Appendix G.

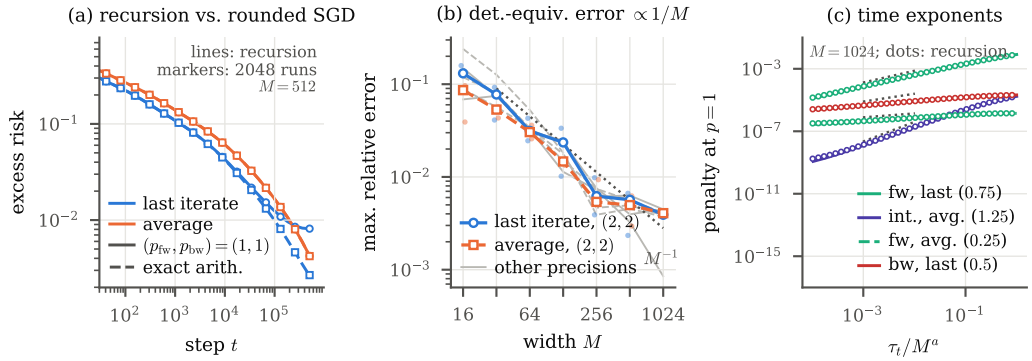


Figure 3: (a) Excess risk along training at $M = 512$: exact recursion (lines) against 2048 runs of rounded SGD with both copies at one bit. (b) Largest relative error of the penalty predicted by DE dynamics over the compute-optimal range, against the width (dots: feature maps; gray: the other precisions). (c) Penalty of each term against time at $M = 1024$ and one bit: dynamics (lines), recursion (dots), predicted exponents (dotted). The joint term and the interaction are in Figure 4.

Theory validation. Figure 3(a) shows the exact recursion agrees with the mean of 2048 runs of rounded SGD at $M = 512$. Figure 3(b) shows the relative error of the penalty predicted by the DE dynamics, which Conjecture 2 asserts to vanish, falls from 13% at $M = 16$ to 0.4% at $M = 1024$. Figure 3(c) indicates the time exponents of Theorem 4 hold at $M = 1024$. The interaction makes the penalty of the average 3.4 times the sum of the single-copy penalties at $M = 1024$, against 1.03 for the last iterate, and grows like M^{a-b} at three pairs (a, b) from Figure 4 in Appendix G.

The precision rules in §4. The allocation (15) at a mean of two bits removes 99.8% of the forward penalty of the last iterate, and its three other ratios are within 0.09 of the prediction (Figures 2(b) and 5(a)). Figure 5(b) shows the retained fraction of the backward penalty after a cooldown. In Figure 2(c), the bits that keep the penalty of the last iterate below a fixed fraction of loss grow by 0.162 bit per doubling of compute, against the $\frac{1}{6}$ of Theorem F.8. Figure 5(c) shows that optimal learning-rate schedule decays earlier and more gently with fewer backward bits (Proposition F.10).

Practical setting. We also trained a two-layer MLP on CIFAR-10 and a six-layer transformer with 30M parameters, with every linear layer rounding its forward copy, its backward copy or both at two bits under a warmup–stable–decay schedule (Figure 7). Backward SR has no measurable effect on the final validation loss. The validation loss uses the rounded forward pass, so the forward gap includes the SR at inference, and at these batch sizes the linear model predicts a backward penalty below the resolution of these runs even without a cooldown, so they test neither the cooldown mechanism nor the timing rules in networks.

6 Conclusion and Limitations

We have derived an exact recursion for the loss curve of one-pass SGD with SR, and a scaling laws for its rounding penalty. The constant-learning-rate last iterate is mainly limited by forward rounding, whereas and backward rounding can become visible once SGD noise is removed, such as an average iterate or a cooldown learning rate. Bits are best spent late in training for a last iterate, their number grows only logarithmically with compute, and a fixed precision lowers the exponent of the compute-optimal loss. Moreover, learning-rate and precision schedules must be designed together. Precision in general should thus be treated as a training schedule, not as one global format: the right precision depends on which matrix product is rounded, when we apply SR, what model is returned at the end, and whether an extra bit is worth more than the training steps it costs.

One-pass SGD makes the recursion of Theorem 1 exact, but it excludes training for several epochs. Since the rounding decisions are drawn afresh at every access, we expect Theorem 4 to hold below the data reuse threshold of Lin et al. (2025) (Appendix H; Figure 8). Rounding the gradient or the stored weights has a state-dependent quantity, and the recursion no longer closes; a heuristic in Appendix H predicts, e.g., that rounding bf16 master weights adds a penalty that a cooldown neither removes nor amplifies. Our model trains a linear random features, and so it excludes nonlinear activations, trained features with heavy-tailed spectra (Hodgkinson et al., 2025), and feature learning.

REPRODUCIBILITY STATEMENT

Every number and figure in the paper is produced by one of three deterministic programs, described in Appendix G: the exact recursion of Theorem 1 and its input-rounding version (Appendix C), the deterministic-equivalent dynamics (10) with the coefficients of the same realised feature map, and a bit-exact simulation of rounded SGD in which the only randomness is the data and the random rounding decisions, with fixed seeds. The same appendix lists the configuration of every experiment, including the network runs on CIFAR-10 and FineWeb, maps each figure panel to its run (paragraph "Panels of the main-text figures"), and states how the rounding-error moments are computed and to what accuracy; the constants of Lemma B.1 are certified by interval arithmetic. The complete proofs are in Appendices B–F, each result with its assumptions. We will release the code, with the command line and the seeds of every run, upon publication.

AI-USE STATEMENT

In this work we used generative AI tools for the following tasks that require disclosure: formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, interpret results. We did not use generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses,assist with translation, clean and reformat dataset, support qualitative and thematic data analysis. We also used generative AI tools to suggest experimental parameters, summarize or analyse existing literature, edit a research paper to improve readability, identify relevant literature, and format references. We have reviewed all AI-assisted work: we checked the decisive step of every proof, state as proved only results whose proofs they accepted, and label as a conjecture (Conjecture 2). We take full responsibility for the content of this paper.

REFERENCES

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023. URL https://arxiv.org/abs/2303.01372.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121.

Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model, 2026. URL https://arxiv.org/abs/2602.04774. Version 2.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Christos Boutsikas, Petros Drineas, and Ilse C. F. Ipsen. Small singular values can increase in lower precision. *SIAM Journal on Matrix Analysis and Applications*, 45(3):1518–1540, 2024. doi: 10.1137/23M1557209.

Roberto L. Castro, Andrei Panferov, Soroush Tabesh, Oliver Sieberling, Jiale Chen, Mahdi Nikdan, Saleh Ashkboos, and Dan Alistarh. Quartet: Native FP4 training can be optimal for large language models, 2025. URL https://arxiv.org/abs/2505.14669. Version 4, revised January 2026.

Jianfei Chen, Yu Gai, Zhewei Yao, Michael W. Mahoney, and Joseph E. Gonzalez. A statistical framework for low-bitwidth training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jianfei Chen, Lianmin Zheng, Zhewei Yao, Dequan Wang, Ion Stoica, Michael W. Mahoney, and Joseph E. Gonzalez. ActNN: Reducing training memory footprint via 2-bit activation compressed training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1803–1813, 2021.

Mengzhao Chen, Chaoyi Zhang, Jing Liu, Yutao Zeng, Zeyue Xue, Zhiheng Liu, Yunshui Li, Jin Ma, Jie Huang, Xun Zhou, and Ping Luo. Scaling law for quantization-aware training, 2025. URL https://arxiv.org/abs/2505.14302.

Ziyan Chen, Zhongzhu Zhou, and Ding-Xuan Zhou. Scaling laws for dynamic mini-batch SGD in sketched linear regression, 2026. URL https://arxiv.org/abs/2605.24316v3. Version 3. Earlier versions titled From One-Pass SGD to Data Reuse: Mini-Batch Scaling Laws in Sketched Linear Regression.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6): 2879–2912, 2024. URL https://arxiv.org/abs/2210.08571.

Brian Chmiel, Maxim Fishman, Ron Banner, and Daniel Soudry. FP4 all the way: Fully quantized training of LLMs, 2025. URL https://arxiv.org/abs/2505.19115.

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. doi: 10.1137/20M1334796. URL https://epubs.siam.org/doi/10.1137/20M1334796.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3123–3131, 2015.

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. doi: 10.1098/rsos.211631. URL https://doi.org/10.1098/rsos.211631.

Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R. Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training, 2018. URL https://arxiv.org/abs/1803.03383.

DeepSeek-AI. DeepSeek-V3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bd18189308a4c45c7d71ca83acf3deaa-Abstract-Conference.html.

Gregory Dexter, Christos Boutsikas, Linkai Ma, Ilse C. F. Ipsen, and Petros Drineas. Stochastic rounding implicitly regularizes tall-and-thin matrices. *SIAM Journal on Matrix Analysis and Applications*, 46(1):341–369, 2025. doi: 10.1137/24M1647679.

Petros Drineas and Ilse C. F. Ipsen. Stochastic rounding 2.0, with a view towards complexity analysis, 2024. URL https://arxiv.org/abs/2410.10517.

El-Mehdi El Arar, Devan Sohier, Pablo de Oliveira Castro, and Eric Petit. Stochastic rounding variance and probabilistic bounds: A new approach. *SIAM Journal on Scientific Computing*, 45(5): C255–C275, 2023. doi: 10.1137/22M1510819. URL https://epubs.siam.org/doi/10.1137/22M1510819.

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, 2026. URL https://arxiv.org/abs/2603.06060v2. Version 2.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Andrew Fitzgibbon and Stephen Felix. On stochastic rounding with few random bits. In *2025 IEEE 32nd Symposium on Computer Arithmetic (ARITH)*, pp. 133–140, 2025. doi: 10.1109/ARITH64983.2025.00029. URL https://arxiv.org/abs/2504.20634.

Elias Frantar, Utku Evci, Wonpyo Park, Neil Houlsby, and Dan Alistarh. Compression scaling laws: Unifying sparsity and quantization, 2025. URL https://arxiv.org/abs/2502.16440.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1737–1746. PMLR, 2015. URL https://proceedings.mlr.press/v37/gupta15.html.

Nicholas J. Higham and Theo Mary. A new approach to probabilistic rounding error analysis. *SIAM Journal on Scientific Computing*, 41(5):A2815–A2835, 2019. doi: 10.1137/18M1226312.

Nicholas J. Higham and Theo Mary. Sharper probabilistic backward error analysis for basic linear algebra kernels with random data. *SIAM Journal on Scientific Computing*, 42(5):A3427–A3446, 2020. doi: 10.1137/20M1314355.

Liam Hodgkinson, Zhichao Wang, and Michael W Mahoney. Models of heavy-tailed mechanistic universality. *arXiv preprint arXiv:2506.03470*, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b2cac94f82928a85055987d9fd44753f-Abstract-Conference.html.

Daniel LeJeune, Pratik Patil, Hamid Javadi, Richard G. Baraniuk, and Ryan J. Tibshirani. Asymptotics of the sketched pseudoinverse, 2022. URL https://arxiv.org/abs/2211.03751. Version 4, revised October 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*, volume 30, pp. 5811–5821, 2017.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6fcb1afcc1e9c2c82c8ddddf03bcf0f6-Abstract-Conference.html.

Licong Lin, Jingfeng Wu, and Peter L. Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.08415.

Xiaoxuan Liu, Lianmin Zheng, Dequan Wang, Yukuo Cen, Weize Chen, Xu Han, Jianfei Chen, Zhiyuan Liu, Jie Tang, Joey Gonzalez, Michael Mahoney, and Alvin Cheung. GACT: Activation compressed training for generic network architectures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14139–14152. PMLR, 2022. URL https://proceedings.mlr.press/v162/liu22v.html.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-acl.26/.

Linkai Ma, Tingzhou Yu, and Petros Drineas. Stochastic rounding increases small singular values, 2026. URL https://arxiv.org/abs/2606.00312.

Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws, 2022. URL https://arxiv.org/abs/2210.16859.

C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4284–4293, 2019.

C. H. Martin and M. W. Mahoney. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 20th SIAM International Conference on Data Mining*, 2020.

C. H. Martin, T. S. Peng, and M. W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(4122): 1–13, 2021.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008. URL https://onlinelibrary.wiley.com/doi/10.1002/cpa.22008.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. FP8 formats for deep learning, 2022. URL https://arxiv.org/abs/2209.05433.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *arXiv preprint arXiv:2411.18704*, 2024.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4402–4410. PMLR, 2025. URL https://proceedings.mlr.press/v258/ozkara25b.html.

Andrei Panferov, Erik Schultheis, Soroush Tabesh, and Dan Alistarh. Quartet II: Accurate LLM pre-training in NVFP4 by improved unbiased gradient estimation, 2026. URL https://arxiv.org/abs/2601.22813. Version 2.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: exact dynamics and generalization properties. *Mathematical Programming*, 214:1–90, 2025. doi: 10.1007/s10107-024-02171-3. URL https://link.springer.com/article/10.1007/s10107-024-02171-3.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1dccfc3ee01871d05e33457c61037d59-Abstract-Conference.html.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Shashank Subramanian, Alexander Kiefer, Arnur Nigmetov, Amir Gholami, Dmitriy Morozov, and Michael W. Mahoney. On neural scaling laws for weather emulation through continual training. *arXiv preprint arXiv:2603.25687*, 2026.

Xingwu Sun, Shuaipeng Li, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, Yu Cheng, Yangyu Tao, Zhanhui Kang, Chengzhong Xu, Di Wang, and Jie Jiang. Scaling laws for floating point quantization training, 2025. URL https://arxiv.org/abs/2501.02423.

Albert Tseng, Tao Yu, and Youngsuk Park. Training LLMs with MXFP4. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1630–1638. PMLR, 2025. URL https://proceedings.mlr.press/v258/tseng25a.html.

Cameron R. Wolfe and Anastasios Kyrillidis. Better schedules for low precision training of deep neural networks. *Machine Learning*, 2024. URL https://arxiv.org/abs/2403.02243.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. SWALP: Stochastic weight averaging in low precision training. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7015–7024. PMLR, 2019. URL https://proceedings.mlr.press/v97/yang19d.html.

Zhongzhi Yu, Yonggan Fu, Shang Wu, Mengquan Li, Haoran You, and Yingyan Lin. LDP: Learnable dynamic precision for efficient deep neural network training and inference, 2022. URL https://arxiv.org/abs/2203.07713.

Dechen Zhang, Junwei Su, and Difan Zou. Learning under quantization for high-dimensional linear regression, 2025. URL https://arxiv.org/abs/2510.18259. Version 3, revised February 2026.

Dechen Zhang, Xuan Tang, Yingyu Liang, and Difan Zou. Scaling laws for precision in high-dimensional linear regression, 2026. URL https://arxiv.org/abs/2602.19241. Version 2.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4035–4043. PMLR, 2017. URL https://proceedings.mlr.press/v70/zhang17e.html.

Kaiyan Zhao, Tsuguchika Tabaru, Kenichi Kobayashi, Takumi Honda, Masafumi Yamazaki, and Yoshimasa Tsuruoka. Direct quantized training of language models with stochastic rounding. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 1150–1165. PMLR, 2025.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients, 2016. URL https://arxiv.org/abs/1606.06160.

CONTENTS

1 Introduction **1**

2 Model, stochastic rounding, and rounded SGD **3**

3 From the exact dynamics to scaling laws with precision **4**

4 Where, when and how much to round **6**

4.1 Where: which rounding location limits accuracy 6

4.2 When: distributing the bits over a run 7

4.3 How much: bits against width, training time and compute 8

4.4 Precision and the learning-rate schedule 8

5 Empirical evaluation **8**

6 Conclusion and limitations **9**

A Related work and Notations **17**

B Proofs for Section 3 **18**

B.1 The cell-variance lemma 18

B.1.1 Scaling, regularity, derivative identity (item 1) 19

B.1.2 Two-sided bounds (item 2) 19

B.1.3 The phase function (item 3) 19

B.1.4 The second-chaos coefficient (item 4) 20

B.1.5 Pair moments (item 5) 21

B.1.6 Certified constants 21

B.2 Lemma B.2: the diagonal of the sketched covariance 22

B.3 Proof of Theorem 1 22

B.4 Averaging over the target: Corollary B.4 25

B.5 Corollary B.5 and its proof 25

B.6 Two-time moments: Corollary B.8 27

C Rounding the input: Proposition C.1 and its proofs **28**

C.1 The nine terms of the P3 gradient second moment 29

C.2 The exact P3 recursion (independent copies) 30

C.3 The Gaussian-smoothing oracle A' 31

C.4 The exact P2 recursion (shared copy) 32

D The deterministic-equivalent dynamics and Conjecture 2 **33**

D.1 The projection identities (exact) 34

D.2 The limit system, positivity and the Neumann series . . . 34
D.3 Two obstructions . . . 35
D.4 The transfer contract (proved) . . . 36
D.5 A static defect of the forward coefficient . . . 37
D.6 An explicit admissible family . . . 37
D.7 What a proof of Conjecture 2 requires . . . 37
D.8 The law for the algorithm under Conjecture 2 . . . 37

E Proofs of the scaling laws: Theorem 4, Proposition E.10 and Theorem E.12 38
E.1 Clean solution . . . 38
E.2 Spectral sums and envelopes . . . 38
E.3 First insertion . . . 39
E.4 The forcing-relative gain . . . 40
E.5 Proof of Theorem 4 . . . 45
E.6 The natural gain and the fourth term . . . 45
E.7 Cooldown: proof of Proposition E.10 . . . 45
E.8 Proof of Theorem E.12: backward rounding, comparison for the exact recursion . . . 48
E.8.1 Definitions, conditioning, and the rounding-error moments . . . 49
E.8.2 Exact innovations, positivity, and a local sandwich . . . 50
E.8.3 The actual algorithm is bracketed, also after averaging . . . 51
E.8.4 Gaussian sketch: proved logarithmic coefficient bounds . . . 52
E.8.5 Optional fixed-width stationary consequences, with full proofs . . . 53
E.8.6 Proof of the law (85) of Theorem E.12 . . . 54

F Full statements and proofs for Section 4 55
F.1 Bit requirements . . . 55
F.2 The regime map: derivation of the ratios behind §4.1 . . . 57
F.3 When: the response weights, the allocation (15) and the closed-form schedules . . . 58
F.4 The compute scaling law: derivation of (116) and of the fixed-precision frontier . . . 60
F.5 The tolerance under a wall-clock budget: derivation behind §4.3 . . . 64
F.6 Precision and the learning-rate schedule: the lemmas and the derivation behind §4.4 66
F.7 Algorithm 1: a closed-form precision schedule . . . 70
F.8 Joint optimisation of learning-rate and precision schedules . . . 70

G Experimental details 71

H Predictions beyond the model: other rounding sites and data reuse 77

A Related work and Notations

Quantization-noise classes. Zhang et al. (2026) (and the earlier Zhang et al., 2025) derive precision scaling laws for high-dimensional linear regression under two classes of quantization noise: a multiplicative class whose covariance is sandwiched between rank-one multiples of the feature covariance, and an isotropic additive class. Ideal binary SR belongs to neither: its error covariance $D_p(z) = \mathrm{diag}_k\, \nu_p(z_k)$ is diagonal in the feature coordinates, vanishes on the grid and depends on the position of each coordinate inside its rounding interval, so no pointwise lower bound of either class holds; after averaging over Gaussian features its effect is the three moments m_p, α_p, c_{ab} of Lemma B.1, and the relation $\alpha_p \approx m_p$ alone does not distinguish it from a multiplicative model with variance cqx^2 (for which $\alpha = m$ exactly) — the distinguishing information is the log-periodic profile and the joint moments. What we import from their work is the setting (power-law regression, one-pass SGD, deterministic equivalents) and the idea of a precision-dependent scaling exponent. Kumar et al. (2025) fit a loss surface in which quantized training acts as a multiplicative reduction of the effective parameter count, a parallel shift of the frontier; our additive, exponentially decaying penalty predicts a bend instead (Section 4.3); the two are fits and calculations for different objects, not contradictory theorems.

Low-precision training, quantization-aware training and post-training quantization. Three settings are often grouped under quantization. In quantization-aware training (QAT) (Jacob et al., 2018), the rounding of weights and activations is simulated during training so that the trained model can run in low-precision (for instance integer) arithmetic at inference, and the object of study is the accuracy of the quantized model. In post-training quantization, a model trained in high precision is rounded after training. In low-precision training, the setting of this paper, the arithmetic of training itself is rounded: here the two copies of the features that enter the forward and the backward products are rounded stochastically, and the object of study is the excess risk of the trained predictor. Kumar et al. (2025) fit precision-aware scaling laws for both low-precision training and post-training quantization, and Chen et al. (2025) fit one for 4-bit quantization-aware training; for low-precision training in our model, Theorem F.8 derives how the compute-optimal number of bits grows with compute, and Appendix F.4 contrasts the change of exponent at fixed precision with the parallel shift of their fitted law.

Low-precision training and precision schedules. Stochastic quantization inside the forward and backward passes, with a high-precision copy of the weights that accumulates the updates, goes back to BinaryConnect (Courbariaux et al., 2015), which can binarize the weights stochastically during both passes, and DoReFa-Net (Zhou et al., 2016), which quantizes the gradients stochastically; Li et al. (2017) compare such schemes with SR applied to the stored weights and analyze their convergence, and Zhao et al. (2025) train language models whose low-precision weights are updated directly by SR, without a high-precision copy. Unbiased independent rounding of the two copies of an operand is the double-sampling protocol of Zhang et al. (2017), who also give end-to-end guarantees for linear models; averaging low-precision SGD iterates is the mechanism of SWALP (Yang et al., 2019); learnable or scheduled precision over training is studied by Yu et al. (2022) and Wolfe & Kyrillidis (2024), and low-precision activation storage by Liu et al. (2022). Our contribution relative to these is not the protocol or the idea of a schedule but the exact dynamics behind them: the location-specific geometry, the joint budget, the output-dependent time allocation and the interaction with the learning-rate schedule are consequences of Theorems 1, 4 and E.12, and Wolfe & Kyrillidis (2024) is the natural baseline for the schedules of Section 4.2. Backward error analyses of SR (Connolly et al., 2021; Croci et al., 2022; El Arar et al., 2023; 2026), which build on the probabilistic rounding error analysis of Higham & Mary (2019; 2020), bound the errors of single products and kernels, in the worst case and with high probability, and are the source of the constant-relative-variance rule of thumb that Lemma B.1 refines. Beyond single products, SR can increase the small singular values of a matrix: it keeps the smallest singular value of a tall-and-thin matrix away from zero (Dexter et al., 2025) and lifts clusters of small singular values (Ma et al., 2026), refining perturbation bounds for rounding to lower precision (Boutsikas et al., 2024); Drineas & Ipsen (2024) argue that SR makes rounding errors cancel rather than accumulate and propose it as a tool for complexity analysis. Here, in contrast, the rounding errors are random variables whose second moments are propagated through the training dynamics, and the low-precision training systems and precision schedules above motivate the rounding locations and the schedules that we study. Deterministic equivalents for sketched

regression (Mei & Montanari, 2022; Bach, 2023; LeJeune et al., 2022; Cheng & Montanari, 2024; Defilippis et al., 2024; Ba et al., 2022) and SGD scaling laws (Paquette et al., 2024; 2025; Lin et al., 2024; 2025; Bordelon & Mori, 2026; Chen et al., 2026) supply the exact-arithmetic dynamics, the compute-optimal horizon $\tau_T \asymp M^a$, and (Bordelon & Mori, 2026) the optimal-control formulation of Section 4.4. Activation-compressed training keeps the forward pass in full precision and stores the activations needed for the weight-gradient product in two to four bits with stochastic rounding (Chen et al., 2021; Liu et al., 2022); in our model this is the backward copy rounded and the forward copy exact. Chen et al. (2020) give a variance analysis of quantized gradients in low-bitwidth training. Hardware implementations of SR draw only a few random bits for each rounding decision, which can bias the rounding (Fitzgibbon & Felix, 2025); the analysis of this paper takes the ideal unbiased rule.

Notation. For $u \in \mathbb{R}^M$, $\|u\|$ is the Euclidean norm and $\|u\|_A^2 = u^\top A u$ for a positive semidefinite matrix A; $A \preceq B$ means that $B - A$ is positive semidefinite, tr is the trace, and $\mathrm{diag}(\cdot)$ is the diagonal matrix with the given entries. Throughout, c and C denote positive constants that do not depend on M, t or the precisions, $f \asymp g$ means $cg \le f \le Cg$, and $f \lesssim g$ means $f \le Cg$.

B Proofs for Section 3

B.1 The cell-variance lemma

Throughout, $p \ge 0$ is an integer, $q = 4^{-p}$, $r = 2^{-p}$ (so $q = r^2$), $g > 0$, $\zeta \sim N(0, g)$, $Z = \zeta/\sqrt{g}$, $\varphi_g(x) = (2\pi g)^{-1/2} e^{-x^2/(2g)}$, and $\mathfrak{b}(x) = 2^{\lfloor \log_2 |x| \rfloor}$ is the base of the binade $[\mathfrak{b}, 2\mathfrak{b})$ containing $|x| \ne 0$. A positive cell of $\mathbb{F}_p$ is $I = [\ell, \ell + \Delta]$ with $\ell = \mathfrak{b}(1 + jr)$, $0 \le j < 2^p$, $\Delta = r\mathfrak{b}$; in the local coordinate $\xi = (x - \ell)/\Delta \in [0, 1]$, $\nu_p(x) = \Delta^2 \xi(1 - \xi)$, so $0 \le \nu_p(x) \le \Delta^2/4 = q\mathfrak{b}(x)^2/4 \le qx^2/4$; ν_p is even and vanishes on the grid. We write

$$h_p(g) = \frac{4^p m_p(g)}{g}, \qquad A_p(g) = \frac{4^p \alpha_p(g)}{g}, \qquad h(g) = \frac{\mathbb{E}[\mathfrak{b}(\zeta)^2]}{6g}, \qquad c_p = \min_{g>0} h_p(g),$$

and $a_n(p, g) = \mathbb{E}[\nu_p(\zeta)\,\mathrm{He}_n(Z)]/n!$ with the probabilists' Hermite polynomials He_n (generating function $e^{tx - t^2/2}$), so $a_0 = m_p$, $a_2 = \alpha_p$, $a_n = 0$ for odd n; c_p is the constant of Proposition C.1.

Lemma B.1 (cell variance)**.** *Let $p \ge 0$, $g > 0$, and let m_p, α_p, c_{ab} be as in* (25)*. Statements marked* [a] *are analytic; those marked* [c] *are computer-assisted with certified (outward-rounded interval) error control, §B.1.6.*

1. (scaling, regularity) [a] $\nu_p(2x) = 4\nu_p(x)$*; hence* $m_p(4g) = 4m_p(g)$, $\alpha_p(4g) = 4\alpha_p(g)$*, and* h_p, A_p *are continuous and* $\log_4$*-periodic, so* c_p *is attained. Moreover* $m_p \in C^\infty(0, \infty)$ *and* $\alpha_p(g) = g\, m_p'(g)$*, equivalently* $A_p = h_p + g\, h_p'$.

2. (two-sided bounds) [a] $\dfrac{q\,g}{24(1+r)^3} \le m_p(g) \le \dfrac{q\,g}{4}$*, so* $c_p \ge 1/192$. [c] $c_p \ge 0.0814$ *for every* p *(certified constant* 0.0814848*; the sharp value is* $\min_p c_p = c_0 = 0.081487\ldots$*).*

3. (phase function) [a] h *is smooth,* $\log_4$*-periodic and nonconstant,* $\frac{1}{24} \le h \le \frac{1}{6}$*, its log-scale average* $\int_0^1 h(4^t)\,dt = \frac{1}{8\ln 4} = 0.09017\ldots$ *lies strictly between* $\min h$ *and* $\max h$*, and* $\big|h_p(g) - h(g)\big| \le \frac{3}{32}\,4^{-p}$ *for all* $g > 0$*. The order is optimal:* $\sup_g |h_p - h| \ge 4^{-p}/(960 \ln 2)$. [c] $0.08976 \le h \le 0.09058$ *and* $\sup_g |g\,h'(g)| \le 0.00185$.

4. ($\alpha_p \approx m_p$) [a] $0 < 0.9\, m_p(g) \le \alpha_p(g) \le 1.1\, m_p(g)$ *for all* $p \ge 0$, $g > 0$*;*

$$\sup_{g>0} \big|A_p - (h + g\,h')\big| \le \tfrac{17}{32}\,4^{-p}, \qquad \sup_{g>0} \Big|\frac{\alpha_p - m_p}{m_p} - \frac{g\,h'}{h}\Big| \le 8 \cdot 4^{-p} \tag{16}$$

   and gh'/h *is a continuous* $\log_4$*-periodic function that is not identically zero, so* $\sup_g |\alpha_p - m_p|/m_p$ *does not tend to zero as* $p \to \infty$. [c] $0.97\, m_p \le \alpha_p \le 1.03\, m_p$ *for all* p, g.

5. (pair moments) [a] *For jointly Gaussian* (ζ, ζ') *with variances* (g, g') *and correlation* $\rho \in [-1, 1]$*,* $c_{ab}(g, g', \rho) = \sum_{n \ge 2\ even} n!\, a_n(p_a, g)\, a_n(p_b, g')\, \rho^n$*, absolutely and uniformly convergent on* $[-1, 1]$*; at* $\rho = \pm 1$ *it equals* $\mathbb{E}[\nu_{p_a}(\zeta)\nu_{p_b}(\sqrt{g'/g}\,\zeta)] - m_{p_a}(g) m_{p_b}(g')$ *(for* $g' = g$*:* $\nu_{p_b}(\pm\zeta)$*). Further* $\sum_{n \ge 1} n\, n!\, a_n(p, g)^2 \le 4^{-p} g^2$ *and* $\big|c_{ab} - 2\alpha_{p_a}(g)\alpha_{p_b}(g')\rho^2\big| \le \frac{3}{16}\,4^{-(p_a + p_b)} g g' \rho^4$.

The universal bounds $0.0814 \le h_p \le \frac{1}{4}$ are a corollary of items 3–4 and the certified constants, not a replacement for the phase function: h_p oscillates with g at every precision, and the "relative variance $4^{-p}/(8\ln 4)$" of Section 2 is its log-scale average.

B.1.1 Scaling, regularity, derivative identity (item 1)

Multiplication by 2 maps $\mathbb{F}_p$ bijectively onto itself and preserves adjacency (a grid point strictly between $2\ell_p(x)$ and $2u_p(x)$ would halve to one strictly between $\ell_p(x)$ and $u_p(x)$), so $\nu_p(2x) = 4\nu_p(x)$; ν_p is even by reflection. Substituting $\sqrt{4g}\,Z = 2\sqrt{g}\,Z$ in (25) gives the scalings, hence the periodicity. For smoothness, $\partial_g^j \varphi_g(x) = g^{-j} P_j(x^2/g)\varphi_g(x)$ with polynomials $P_0 = 1$, $P_{j+1}(v) = (v/2 - j - \frac{1}{2})P_j(v) - vP_j'(v) = \sum_k p_{jk} v^k$, and on $[g_-, g_+] \subset (0,\infty)$ the envelope $\nu_p \le qx^2/4$ gives the dominator

$$|\nu_p(x)\,\partial_g^j \varphi_g(x)| \le \frac{q\,x^2}{4\sqrt{2\pi g_-}\,g_-^{\,j}} \Big(\sum_{k=0}^{j} |p_{jk}|\,\frac{x^{2k}}{g_-^{\,k}}\Big) e^{-x^2/(2g_+)}, \tag{17}$$

a polynomial times a Gaussian. The mean-value theorem bounds the order-j difference quotients on $[g_-/2, 2g_+]$ by the order-$(j+1)$ dominator, so dominated convergence gives $m_p \in C^\infty$ with $m_p^{(j)} = \int \nu_p\,\partial_g^j \varphi_g$. With $j = 1$, $\partial_g \varphi_g = \frac{1}{2g}(x^2/g - 1)\varphi_g$, so $m_p'(g) = \frac{1}{2g}\mathbb{E}[\nu_p(\zeta)(\zeta^2/g - 1)] = \alpha_p(g)/g$; the second form is the product rule for $h_p = m_p/(qg)$. □

B.1.2 Two-sided bounds (item 2)

Upper bound: integrate $\nu_p(\zeta) \le q\zeta^2/4$. *Analytic lower bound.* On a positive cell $I = [\ell, \ell + \Delta] \subset [\mathfrak{b}, 2\mathfrak{b})$, φ_g is decreasing, so

$$\int_I \nu_p \varphi_g \;\ge\; \varphi_g(\ell + \Delta)\,\Delta^3 \int_0^1 \xi(1-\xi)\,d\xi = \tfrac{1}{6}\Delta^3 \varphi_g(\ell + \Delta). \tag{18}$$

For $x \in I$, $(1+r)x \ge (1+r)\ell \ge \ell + r\mathfrak{b} = \ell + \Delta$, hence $\int_I \varphi_g((1+r)x)\,dx \le \Delta\varphi_g(\ell+\Delta)$ and $\int_I \nu_p\varphi_g \ge \frac{q\mathfrak{b}^2}{6}\int_I \varphi_g((1+r)x)\,dx$. Summing over all cells (Tonelli), reflecting, and using $\mathfrak{b}(x)^2 \ge x^2/4$ and $y = (1+r)x$: $m_p(g) \ge \frac{q}{6}\int_{\mathbb{R}} \mathfrak{b}^2 \varphi_g((1+r)x)\,dx \ge \frac{q}{24}\int_{\mathbb{R}} x^2 \varphi_g((1+r)x)\,dx = \frac{q\,g}{24(1+r)^3} \ge \frac{q\,g}{192}$.

Phase-dependent lower bound. Let $P_I = \int_I \varphi_g \le \Delta\varphi_g(\ell)$ and $\varphi_g(\ell+\Delta)/\varphi_g(\ell) = e^{-\theta_I}$ with $\theta_I = (\ell\Delta + \Delta^2/2)/g \le c_e := \mathfrak{b}^2 r(2 + r/2)/g$ (since $\ell < 2\mathfrak{b}$). As $e^{-\theta} \ge 1 - \theta$ for all $\theta \ge 0$ (no smallness is needed; when $1 - c_e < 0$ the bound is weak, not false), (18) gives $\int_I \nu_p\varphi_g \ge \frac{\Delta^2}{6} e^{-\theta_I} P_I \ge \frac{q\mathfrak{b}^2}{6}(1 - c_e)P_I$. Sum over the 2^p cells of the binade $[\mathfrak{b}, 2\mathfrak{b})$ and both signs, then over the binades $\mathfrak{b} = 2^e$: with $P_e = \Pr(\mathfrak{b}(\zeta) = 2^e)$, $\sum_e 4^e P_e = \mathbb{E}\mathfrak{b}(\zeta)^2 = 6gh(g)$ and $\sum_e 16^e P_e = \mathbb{E}\mathfrak{b}(\zeta)^4 \le \mathbb{E}\zeta^4 = 3g^2$, both finite, so the possibly negative terms may be summed, and

$$m_p(g) \ge \frac{q}{6}\Big[\mathbb{E}\mathfrak{b}(\zeta)^2 - \frac{r(2+r/2)}{g}\mathbb{E}\mathfrak{b}(\zeta)^4\Big] \ge \frac{q}{6}\big[6gh(g) - 3g\,r(2+r/2)\big],$$

$$\text{i.e.} \qquad h_p(g) \ge h(g) - 2^{-p}\big(1 + 2^{-p-2}\big). \tag{19}$$

For $p \ge 8$ this and item 3 give $h_p \ge 0.08976 - 0.00392 > 0.0858$; the certified 0.0814848 is obtained for every p at once in §B.1.4, without any phase grid. □

B.1.3 The phase function (item 3)

Kernel identity and rate. For a cell I with local coordinate ξ and $w \in C^2$, let $K(\xi) = -\xi^2(1-\xi)^2/12$: then $K''(\xi) = \xi(1-\xi) - \frac{1}{6}$, $K = K' = 0$ at $\xi \in \{0, 1\}$, and $|K| \le \frac{1}{192}$, so two integrations by parts give

$$\int_I \big[\xi(1-\xi) - \tfrac{1}{6}\big] w(x)\,dx = \Delta^2 \int_I K(\xi)\,w''(x)\,dx, \qquad \Big|\int_I \big[\xi(1-\xi) - \tfrac{1}{6}\big] w\Big| \le \frac{\Delta^2}{192}\int_I |w''| : \tag{20}$$

the value and the linear Taylor term of w at the cell centre both cancel against the symmetric weight. On each cell $4^p \nu_p(x) = \mathfrak{b}(x)^2 \xi(1-\xi)$ and $\Delta^2 = q\mathfrak{b}^2$, so summing (20) with $w = \varphi_g$ over all cells,

with $\varphi_g'' = (x^2/g^2 - 1/g)\varphi_g$, $\mathfrak{b}^4 \le x^4$ and the Gaussian moments $3g^2, 15g^3$,

$$h_p(g) - h(g) = \frac{q}{g}\int_{\mathbb{R}} \mathfrak{b}(x)^4 K(\xi(x))\,\varphi_g''(x)\,dx,$$
$$|h_p(g) - h(g)| \le \frac{q}{192g}\int_{\mathbb{R}} x^4\Big(\frac{x^2}{g^2} + \frac{1}{g}\Big)\varphi_g\,dx = \frac{15+3}{192}\,q = \frac{3}{32}\,q; \tag{21}$$

the integrable majorant justifies the sum over cells and the uniformity in g; keeping $|x^2/g - 1|$ gives $C_m q$ with $C_m = \mathbb{E}[Z^4|Z^2-1|]/192 < 0.0629$. Smoothness of h follows from (17) with $\mathfrak{b}^2$ in place of ν_p; $h(4g) = h(g)$ from $\mathfrak{b}(2x) = 2\mathfrak{b}(x)$; and $x^2/4 \le \mathfrak{b}^2 \le x^2$ gives $\frac{1}{24} \le h \le \frac{1}{6}$.

Log-scale average. Write $g = 4^t$ and $L = \log_2|Z|$ (finite a.s.). Then $\mathfrak{b}(2^tZ)^2 = 4^{\lfloor t+L\rfloor} = Z^2\,4^{-\{t+L\}}$, so $h(4^t) = \frac{1}{6}\mathbb{E}\big[Z^2 4^{-\{t+L\}}\big]$. The integrand $(t,\omega) \mapsto Z^2 4^{-\{t+L\}}$ is nonnegative and jointly measurable, so by Tonelli and translation invariance modulo one,

$$\int_0^1 h(4^t)\,dt = \frac{1}{6}\,\mathbb{E}\Big[Z^2\int_0^1 4^{-\{t+L\}}\,dt\Big] = \frac{\mathbb{E}Z^2}{6}\int_0^1 4^{-u}\,du = \frac{1}{6}\cdot\frac{1-1/4}{\ln 4} = \frac{1}{8\ln 4}.$$

Fourier profile; nonconstancy. Let $V_p(u) = 4^p\nu_p(2^u)/4^u$ (one-periodic, $0 \le V_p \le \frac{1}{4}$), $v_{p,k} = \int_0^1 V_p(u)e^{-2\pi iku}du$, $\mu_p = v_{p,0} > 0$, $\tau_k = 2\pi k/\ln 2$, and

$$M_k = \mathbb{E}|Z|^{2+i\tau_k} = 2^{i\tau_k/2}\,\frac{\Gamma\big(\frac{3+i\tau_k}{2}\big)}{\Gamma\big(\frac{3}{2}\big)}, \qquad b_k = |M_k| = \Big(\frac{1+\tau_k^2}{\cosh(\pi\tau_k/2)}\Big)^{1/2} \tag{22}$$

(by $|\Gamma(\frac{1}{2}+iy)|^2 = \pi/\cosh\pi y$). Since $h_p(4^t) = \mathbb{E}[Z^2V_p(t+\log_2|Z|)]$, Fubini gives $h_p(4^t) = \sum_{k\in\mathbb{Z}} v_{p,k}M_k e^{2\pi ikt}$; the profile of h is $V_\infty(u) = 4^{-u}/6$ with $v_{\infty,k} = [8(\ln 4 + 2\pi ik)]^{-1}$, so $h(4^t) = \sum_k v_{\infty,k}M_k e^{2\pi ikt}$. The three elementary constants $\tau_1 < 9.1$, $e^{-\pi^2/(2\ln 2)} < R := 0.00081$, $\sqrt{2(1+\tau_1^2)} < 13$ and $\cosh y \ge e^y/2$ give $b_k \le 13kR^k$ ($k \ge 1$), so these series and their termwise derivatives converge absolutely and uniformly, and Fourier uniqueness identifies them with the continuous functions $h_p(4^t)$, $h(4^t)$. As $M_1 \ne 0$, h is nonconstant, so its minimum lies strictly below and its maximum strictly above its average.

Optimality of the order. By the logarithmic change of variable, $\mu_p = \int_0^1 h_p(4^t)dt = \frac{1}{\ln 2}\sum_j \int_{I_j} \xi(1-\xi)x^{-3}dx$ over the cells I_j tiling $[1,2]$, and $\mu_\infty = 1/(8\ln 4)$. Apply (20) with $w = x^{-3}$: $w'' = 12x^{-5} \ge \frac{3}{8}$ on $[1,2]$, $-K \ge 0$ and $\int_0^1 K = -\frac{1}{360}$, so $\mu_\infty - \mu_p \ge \frac{q}{\ln 2}\cdot\frac{3}{8}\cdot\frac{1}{360} = \frac{q}{960\ln 2}$; since $|\int_0^1 (h_p - h)(4^t)dt| \le \sup_g|h_p - h|$, no uniform $o(4^{-p})$ rate exists. (In fact $h_p = h - q\,k + O(q^2)$ with $k(g) = \mathbb{E}[\mathfrak{b}(\zeta)^4(\zeta^2/g - 1)]/(360g^2) > 0$; not used.)

Elementary range of h and gh' [c]. $|h - \mu_\infty| \le 2\sum_{k\ge1}|v_{\infty,k}|b_k$ with $|v_{\infty,k}| \le (16\pi k)^{-1}$. An elementary interval evaluation of the closed form (22) (real exponentials only) gives $b_1 \in [0.0104369508, 0.0104369509]$, hence $2|v_{\infty,1}|b_1 < 0.000406$, and the tail is $\le \frac{13}{8\pi}\frac{R^2}{1-R} < 4\cdot 10^{-7}$. Therefore $|h - 1/(8\ln 4)| < 0.000407$, i.e. $h \in [0.08976, 0.09058]$; likewise $gh'(g) = \frac{d}{dt}h(4^t)/\ln 4$ gives $\sup|gh'| \le \frac{2\pi}{\ln 4}\big(2|v_{\infty,1}|b_1 + \sum_{k\ge2} 2k|v_{\infty,k}|b_k\big) < 0.00185$. □

B.1.4 THE SECOND-CHAOS COEFFICIENT (ITEM 4)

Weighted limit. Apply (20) with $w_g = \frac{1}{2}(x^2/g - 1)\varphi_g$, for which $w_g'' = \frac{1}{2g}(x^4/g^2 - 6x^2/g + 3)\varphi_g = \frac{1}{2g}\,\mathrm{He}_4(x/\sqrt{g})\varphi_g$; the cell summation of (21) gives $\big|A_p(g) - \frac{1}{6g}\int_{\mathbb{R}}\mathfrak{b}^2 w_g\big| \le \frac{q}{192g}\int_{\mathbb{R}}\mathfrak{b}^4|w_g''| \le \frac{q}{384}\mathbb{E}[Z^4(Z^4 + 6Z^2 + 3)] = \frac{105+90+9}{384}\,q = \frac{17}{32}\,q$. With $J(g) = \int \mathfrak{b}^2\varphi_g = 6gh(g)$, the dominator (17) (with $\mathfrak{b}^2$ for ν_p) gives $gJ'(g) = \int \mathfrak{b}^2 w_g$, so the limit is $J'(g)/6 = h + gh'$: no derivative of a limit is taken.

Positivity and the ten-percent bound without any scan. By item 1 and termwise differentiation of the Fourier series,

$$A_p(4^t) = h_p(4^t) + \frac{1}{\ln 4}\frac{d}{dt}h_p(4^t) = \sum_{k\in\mathbb{Z}}\big(1 + \tfrac{i\tau_k}{2}\big)v_{p,k}M_k e^{2\pi ikt}.$$

The decisive observation is that $V_p \geq 0$, so $|v_{p,k}| \leq \mu_p$ for every k, with no dependence on p. With $b_k \leq 13kR^k$, $\sum_{k\geq 1} kR^k = R/(1-R)^2$, $\sum_{k\geq 1} k^2R^k = R(1+R)/(1-R)^3$ and $\tau_k/2 = \pi k/\ln 2$ ($26\pi/\ln 2 < 1183/10$),

$$\begin{aligned} |h_p(4^t) - \mu_p| &\leq D_0\mu_p, & D_0 &= 26R/(1-R)^2 < 0.022, \\ |A_p(4^t) - h_p(4^t)| &\leq D_1\mu_p, & D_1 &= \tfrac{1183}{10}R(1+R)/(1-R)^3 < 0.097. \end{aligned} \tag{23}$$

Hence $h_p \geq (1-D_0)\mu_p > 0$ and $|A_p - h_p|/h_p \leq D_1/(1-D_0) < 97/978 < 1/10$, i.e. $0.9m_p \leq \alpha_p \leq 1.1m_p$ and $\alpha_p > 0$ for every $p \geq 0$ (including $p = 0$) and $g > 0$; ν_p itself is not monotone.

Three percent and the constant 0.0814848 [c]. Separate $k = \pm 1$ in (23). Cell-wise antiderivatives on $[1, 2]$ give closed forms, $\mu_p = \frac{rU_{p,0}/2 - \ln 2}{r^2 \ln 2}$ and $v_{p,k} = \frac{U_{p,k}}{r \ln 2\,(1+i\tau_k)(2+i\tau_k)}$ ($k \neq 0$) with $U_{p,k} = \frac{3}{2} + 2\sum_{j=1}^{2^p-1}(1+jr)^{-1-i\tau_k}$. Elementary interval evaluation gives $|v_{p,1}|/\mu_p \leq \varrho_1 := 0.31167$ and $\mu_p \geq 0.08202128$ for $p \leq 2$; for $p \geq 3$, (20) applied to $x^{-3-i\tau_1}$ and to x^{-3} gives $|v_{p,1} - v_{\infty,1}| < 0.00262$ and $\frac{1}{16\ln 2} - \frac{15q}{1024\ln 2} \leq \mu_p \leq \frac{1}{16\ln 2}$, whence $\mu_p > 0.0897$ and $|v_{p,1}|/\mu_p < 0.246$. With $\beta = 0.010437 > b_1$, $T = 9.065 > \tau_1$ and the rational $k \geq 2$ tails $E_0 = 26\big(R/(1-R)^2 - R\big)$, $E_1 = \frac{1183}{10}\big(R(1+R)/(1-R)^3 - R\big)$,

$$\begin{aligned} \frac{|A_p - h_p|}{h_p} &\leq \frac{T\varrho_1\beta + E_1}{1 - 2\varrho_1\beta - E_0} < 0.029995 < \frac{3}{100}, \\ h_p &\geq (1 - 2\varrho_1\beta - E_0)\mu_p > 0.993460 \cdot 0.08202128 > 0.0814848. \end{aligned} \tag{24}$$

Neither uses a phase grid or Stirling's formula. ***Relative limit.*** Taking limits in the ten-percent bound gives $|h + gh'| \leq 1.1h$; with $c = 0.0814848 \leq h_p$, $|A_p/h_p - (h+gh')/h| \leq \big(|A_p - (h+gh')| + 1.1|h_p - h|\big)/c \leq (17/32 + 1.1 \cdot 3/32)q/c < 8q$ (with the analytic $c = 1/192$ the constant becomes 122). Since $h' \not\equiv 0$ and $h > 0$, gh'/h is a continuous nonzero periodic function, and uniform convergence carries the suprema along. □

B.1.5 Pair moments (item 5)

Let $f_a(x) = \nu_{p_a}(\sqrt{g}\,x)$, $f_b(x) = \nu_{p_b}(\sqrt{g'}\,x)$ and (Z, Z') standard Gaussian with correlation ρ. From $\mathbb{E}[e^{sZ - s^2/2}e^{tZ' - t^2/2}] = e^{\rho st}$, $\mathbb{E}[\mathrm{He}_n(Z)\,\mathrm{He}_m(Z')] = n!\,\rho^n\,\mathbf{1}_{n=m}$, also for $\rho = \pm 1$. Polynomials are dense in $L^2(\gamma)$, so $f_a = \sum_n a_n(p_a, g)\,\mathrm{He}_n$ in L^2 with $\sum_n n!a_n^2 = \mathbb{E}f_a^2 \leq \frac{3}{16}4^{-2p_a}g^2$ (from $\nu_p \leq qx^2/4$, $\mathbb{E}Z^4 = 3$); odd coefficients vanish by evenness, $a_0 = m$, $a_2 = \alpha$. The pair formula holds for finite truncations and passes to the limit by Cauchy–Schwarz under the joint law (the marginals are unchanged); on the coefficient side, $\sum_{n\geq 1} n!\,|a_n(p_a, g)a_n(p_b, g')| \leq \sqrt{\mathrm{Var}\,f_a(Z)\,\mathrm{Var}\,f_b(Z)} \leq \frac{3}{16}4^{-(p_a+p_b)}gg'$, so the series converges absolutely and uniformly on $[-1, 1]$; factoring ρ^4 from the terms $n \geq 4$ gives the quartic remainder bound. At $\rho = \pm 1$, $\mathbb{E}(Z' \mp Z)^2 = 0$ gives $Z' = \pm Z$ a.s., hence the endpoint formula with the factor $\sqrt{g'/g}$ (which cannot be dropped when $g \neq g'$). For the H^1 bound, ν_p is locally absolutely continuous with $\nu_p'(x) = \ell_p(x) + u_p(x) - 2x$ a.e., so $|\nu_p'(x)| \leq \Delta \leq r|x|$ and $\mathbb{E}f_p'(Z)^2 \leq qg^2$ for $f_p(x) = \nu_p(\sqrt{g}\,x)$; since $(\mathrm{He}_{n-1}\,\varphi_1)' = -\,\mathrm{He}_n\,\varphi_1$, integration by parts (boundary terms cancel by continuity) gives $\mathbb{E}[f_p'\,\mathrm{He}_{n-1}] = n!\,a_n$, so f_p' has Hermite coefficients na_n and Parseval yields $\sum_{n\geq 1}(n-1)!\,(na_n)^2 = \sum_{n\geq 1} n\,n!\,a_n^2 = \mathbb{E}f_p'(Z)^2 \leq 4^{-p}g^2$. The higher coefficients are not sign-definite ($a_4(1, 1.7) < 0$ [c]). □

B.1.6 Certified constants

Every [c] constant is produced by a script in the released code using only outward-rounded interval arithmetic with exact rational endpoints and elementary functions evaluated by Taylor polynomials with explicit remainders; every assertion compares interval endpoints. ***Elementary route*** (no phase grid, no Stirling formula): the closed forms for μ_p, $v_{p,1}$ ($p \leq 2$), b_1, C_m and the three constants $\tau_1 < 9.1$, $e^{-\pi^2/(2\ln 2)} < R$, $\sqrt{2(1+\tau_1^2)} < 13$; it certifies 0.0814848, (24), $h \in [0.08976, 0.09058]$ and $\sup|gh'| \leq 0.00185$, i.e. every [c] number in Lemma B.1. ***Phase-grid route***: h_p, A_p, h, $h + gh'$ are evaluated from their Fourier series with modes $|k| \leq 8$ (tails from $b_k \leq 13kR^k$; the gamma factor in (22) by the finite Stirling expansion with the complex remainder bound of DLMF §5.11(ii), the

only external input) at the 2000 points $g_j = 4^{j/2000}$ of one period, and the grid extrema are extended to the period by the curvature bound $\min_t f \geq \min_j f(j/N) - \|f''\|_\infty/(8N^2)$, $\|f''\|_\infty$ from the differentiated series. The grid values are reproduced independently by the closed-form cell integrals $\int_I \nu_p\varphi_g = -I_2 + (\ell+\ell')I_1 - \ell\ell' I_0$ ($I_n = \int_\ell^{\ell'} x^n\varphi_g$, $\ell' = \ell + \Delta$) in multiprecision, where the a-priori Lipschitz bound $|d(m_p/g)/d\ln g| = |\alpha_p - m_p|/g \leq \frac{3}{4}q$ (from $|\alpha_p| \leq \frac{q}{2}g$ and item 2) alone certifies the grid minima of h_p to within $\frac{3}{4}\ln 4/2000 = 0.00052$. The phase-grid route gives:

p	0	1	2	3	4	5	6	7
$c_p \geq$	0.08148775	0.08709218	0.08903888	0.08957839	0.08971679	0.08975161	0.08976033	0.08976252
$4^p \sup_g \lvert h_p - h\rvert \leq$	0.00879	0.01102	0.01175	0.01195	0.01201	0.01202	0.01202	0.01202

and $h \in [0.0897632, 0.0905743]$, $\sup_g |gh'| \leq 0.0018385$, $h + gh' \in [0.0882867, 0.0920511]$, $\alpha_p/m_p \in [0.970536, 1.029510]$ for $0 \leq p \leq 8$ (for $p \geq 9$, (16) gives $|\alpha_p/m_p - 1| \leq 0.0018385/0.0897632 + 8\cdot 4^{-9} < 0.0206$). The c_p row (lower endpoints rounded down) and (19) with $\min h > 0.0897632$ give the sharp headline $c_p \geq 0.08148775$ for all p; the paper quotes 0.0814. The second row increases to $0.0120 = \max_g k(g)$, as the optimality statement of item 3 predicts.

Throughout, S (of full row rank, an event of probability one) and θ_* are fixed, $\mathbb{E}$ is the expectation given (S, θ_*), $H_t = I - \eta_t G$, and $\mathcal{F}_t$ is the σ-field generated by (S, θ_*) and the inputs, label noises and rounding decisions of steps $0, \ldots, t-1$. Since $|Q_p(u)| \leq 2|u|$ for every p, (6) gives $\|v_{t+1}\| \leq U_t\|v_t\| + V_t$ with U_t, V_t polynomial in the batch, whence by induction $\mathbb{E}\|v_t\|^2 \leq L_t(S)(1 + \|\theta_*\|^2)$ for a finite $L_t(S)$: all second moments below are finite and, as $\operatorname{tr} C_\theta < \infty$, teacher-integrable at fixed S; the envelope (4) bounds every cell integrand by a Gaussian polynomial. This justifies all interchanges of expectations below.

B.2 Lemma B.2: the diagonal of the sketched covariance

Lemma B.2 (the sketch basis has a flat diagonal but not a vanishing one)**.** *For the Gaussian sketch of Section 2, $M\,G_{kk} \overset{d}{=} \sum_{j\leq d}\lambda_j w_j^2$ with w_j i.i.d. $N(0,1)$, independently across the rows k of S; hence $\mathbb{E}G_{kk} = \operatorname{tr}\Sigma_d/M$ and $\operatorname{Var}(G_{kk})/(\mathbb{E}G_{kk})^2 = 2\operatorname{tr}(\Sigma_d^2)/(\operatorname{tr}\Sigma_d)^2$, which stays bounded away from zero as $d \to \infty$ when $a > 1$ (the effective rank $(\operatorname{tr}\Sigma_d)^2/\operatorname{tr}\Sigma_d^2$ is bounded), while $\max_k G_{kk} \lesssim (\operatorname{tr}\Sigma_d/M)\log M$ with probability $1 - o(1)$. Moreover $OGO^\top \overset{d}{=} G$ for every fixed orthogonal O, so the eigenvectors of G are Haar-distributed in the sketch basis.*

Proof. The rows of $\sqrt{M}S$ are i.i.d. standard Gaussian vectors, which gives the representation and the variance ratio; the maximum is a chi-square tail bound; the invariance follows from $S \overset{d}{=} OS$. □

B.3 Proof of Theorem 1

Notation. In this appendix the subscripts a and b on the rounding matrices and on the cross-covariance abbreviate the forward and backward copies of the main text: $M_{b,t} = M_{\mathrm{bw},t}$, $A_{a,t} = A_{\mathrm{fw},t}$, $\Lambda_{b,t} = \Lambda_{\mathrm{bw},t}$, $\Delta_{ab,t} = \Delta_{\mathrm{fwbw},t}$ and $c_{ab} = c_{p_{\mathrm{fw}}p_{\mathrm{bw}}}$ in the notation of (27) below.

The rounding coefficients. Rounding enters the dynamics only through three Gaussian moments of the SR error variance: for $\zeta \sim N(0,g)$, (ζ, ζ') jointly Gaussian with variances (g, g') and correlation ρ, and $r \in \{\mathrm{fw}, \mathrm{bw}\}$,

$$m_p(g) = \mathbb{E}\,\nu_p(\zeta), \quad \alpha_p(g) = \tfrac{1}{2}\mathbb{E}\big[\nu_p(\zeta)(\zeta^2/g - 1)\big], \quad c_{ab}(g, g', \rho) = \operatorname{Cov}\big(\nu_{p_a}(\zeta), \nu_{p_b}(\zeta')\big), \tag{25}$$

$$\iota_r = \sum_k \frac{m_{p_r}(G_{kk})}{M}, \quad \bar{\alpha}_r = \sum_k \frac{\alpha_{p_r}(G_{kk})}{MG_{kk}}, \quad \gamma_{\mathrm{fwbw}} = \frac{1}{M}\sum_{k,m} c_{p_{\mathrm{fw}}p_{\mathrm{bw}}}(G_{kk}, G_{mm}, \rho_{km}), \tag{26}$$

with $\rho_{km} = G_{km}/\sqrt{G_{kk}G_{mm}}$ the correlation of the features k and m. Here m_p is the mean error variance, α_p its second Hermite coefficient (the part of $\nu_p(\zeta)$ correlated with ζ^2) and c_{ab} the cross-covariance of two error variances, the last evaluated at the precisions $(p_a, p_b) = (p_{\mathrm{fw},t}, p_{\mathrm{bw},t})$ of step t and written $c_{p_{\mathrm{fw}}p_{\mathrm{bw}}}$ below; ι_r, $\bar{\alpha}_r$ and γ_{fwbw} are their averages over the feature coordinates, the coefficients of the deterministic-equivalent dynamics (10) (§3), whose sizes for the Gaussian feature map are Assumption 3(c) (Lemmas B.1 and B.2). We index the two copies by $r \in \{\mathrm{fw}, \mathrm{bw}\}$, so that $p_{r,t}$ is the precision of copy r at step t, and we write $\rho_{km} = G_{km}/\sqrt{G_{kk}G_{mm}}$ for the correlation of the features k and m. For a sequence $(x_k)_{k\le M}$, $\mathrm{diag}_k(x_k)$ is the $M \times M$ diagonal matrix whose (k,k) entry is x_k. The map $\Lambda_{\mathrm{bw},t}(X)$ below is defined for symmetric matrices X and uses only the diagonal entries X_{kk}. With this notation, the rounding noise enters through two scalars and four diagonal matrices:

$$\begin{gathered}
\mathsf{R}_t = \sigma^2_{\mathrm{eff}} + \mathrm{tr}(GC_t), \quad \mathsf{F}_t = \mathrm{tr}(M_{\mathrm{fw},t}W_t), \\
M_{r,t} = \mathrm{diag}_k\, m_{p_{r,t}}(G_{kk}), \qquad A_{\mathrm{fw},t} = \mathrm{diag}_k\, \frac{\alpha_{p_{\mathrm{fw},t}}(G_{kk})(W_t)_{kk}}{G_{kk}}, \\
\Lambda_{\mathrm{bw},t}(X) = \mathrm{diag}_k\, \frac{\alpha_{p_{\mathrm{bw},t}}(G_{kk})X_{kk}}{G_{kk}}, \quad \Delta_{\mathrm{fwbw},t} = \mathrm{diag}_m \textstyle\sum_k (W_t)_{kk}\, c_{p_{\mathrm{fw}}p_{\mathrm{bw}}}(G_{kk}, G_{mm}, \rho_{km}),
\end{gathered} \tag{27}$$

where the sum over k in $\Delta_{\mathrm{fwbw},t}$ includes $k = m$; its entries are of order $q_{\mathrm{fw}}q_{\mathrm{bw}}\ \mathrm{tr}(W_t)\,(\mathrm{tr}\,\Sigma_d)^2/M^2$, that is, a factor q_{bw} below the entries of the forward matrix $\mathsf{F}_t G$, and of the same order in every coordinate (Appendix B). Here R_t is the mean-squared residual of the prediction on a fresh sample: by (3), $\mathrm{tr}(GC_t)$ is twice the expected estimation error, and σ^2_{eff} is the variance of the label residual. F_t is the forward rounding variance, the total error variance that rounding the forward copy adds to the prediction $a_\ell^\top v_t$. $M_{r,t}$ collects the mean error variance of copy r in each feature coordinate. $A_{\mathrm{fw},t}$ and $\Lambda_{\mathrm{bw},t}$ carry the second Hermite coefficient α_p, the part of the error variance that is correlated with the square of the feature (§3), weighted by the diagonal of W_t and of X respectively. $\Delta_{\mathrm{fwbw},t}$ collects the cross-covariance of the error variances of the two copies. All four matrices are diagonal in the feature coordinates, and W_t enters them only through its diagonal.

(a) Conditional Gaussian structure. For $w \neq 0$, $w^\top G w = \|\Sigma_d^{1/2} S^\top w\|^2 > 0$, so $G \succ 0$. Put $\tilde{\epsilon} = x^\top(\theta_* - S^\top v_*^S) + \epsilon$, so that $y = z^\top v_*^S + \tilde{\epsilon}$. Given (S, θ_*), $(z, \tilde{\epsilon})$ is jointly Gaussian with $\mathrm{Cov}(z, \tilde{\epsilon}) = S\Sigma_d\theta_* - G v_*^S = 0$ and $\mathrm{Var}\,\tilde{\epsilon} = \sigma^2 + \theta_*^\top \Sigma_d \theta_* - 2 v_*^{S\top} S \Sigma_d \theta_* + v_*^{S\top} G v_*^S = \sigma^2 + \theta_*^\top \Sigma_{\perp S} \theta_* = \sigma^2_{\mathrm{eff}}$; hence $\tilde{\epsilon} \perp z$. With the orthogonal projector $\Pi_S = \Sigma_d^{1/2} S^\top G^{-1} S \Sigma_d^{1/2}$ one has $\Sigma_{\perp S} = \Sigma_d^{1/2}(I - \Pi_S)\Sigma_d^{1/2} \succeq 0$, so $A_S \geq 0$. The samples of a batch are i.i.d. given (S, θ_*).

(b) The weighted Gaussian identity.

Lemma B.3. *Let $z \sim N(0, G)$, fix k, write $g_k = G_{kk}$, and let ν be measurable with $\mathbb{E}[|\nu(z_k)|(1 + z_k^2)] < \infty$ (e.g. $\nu = \nu_p$, by (4)). With $m = \mathbb{E}\nu(z_k)$ and $\alpha = \frac{1}{2}\mathbb{E}[\nu(z_k)(z_k^2/g_k - 1)]$ (for $\nu = \nu_p$: $m_p(g_k)$, $\alpha_p(g_k)$ of (25)),*

$$\mathbb{E}[\nu(z_k) z_m z_n] = m\, G_{mn} + \frac{2\alpha}{g_k}\, G_{km} G_{kn} \qquad \textit{for all } m, n. \tag{28}$$

Proof. Write $z_m = (G_{km}/g_k) z_k + \xi_m$. Then $\xi = (\xi_m)_m$ is Gaussian and uncorrelated with, hence independent of, z_k, with $\mathbb{E}\xi = 0$ and $\mathbb{E}[\xi_m \xi_n] = G_{mn} - G_{km}G_{kn}/g_k$ (and $\xi_k = 0$). Expanding $z_m z_n$, the two terms linear in ξ have zero expectation because $\mathbb{E}\xi = 0$ and $\xi \perp z_k$ (no evenness of ν is needed), so $\mathbb{E}[\nu(z_k) z_m z_n] = \frac{G_{km}G_{kn}}{g_k^2}\mathbb{E}[\nu(z_k) z_k^2] + m\big(G_{mn} - \frac{G_{km}G_{kn}}{g_k}\big)$. By definition of α, $\mathbb{E}[\nu(z_k) z_k^2] = g_k(m + 2\alpha)$; substitution cancels the two terms $\pm m G_{km} G_{kn}/g_k$. □

(c) The rounding-decision identity. Fix one sample and a deterministic v; put $\delta = v - v_*^S$, $a = z + \xi$, $b = z + \zeta$, where $\xi, \zeta \in \mathbb{R}^M$ are the two rounding errors: given z they are independent and centred with independent entries, $\mathbb{E}[\xi\xi^\top \mid z] = D_{p_{\mathrm{fw}}}(z)$, $\mathbb{E}[\zeta\zeta^\top \mid z] = D_{p_{\mathrm{bw}}}(z)$. With $r = z^\top v - y = z^\top \delta - \tilde{\epsilon}$ the one-sample gradient is $\gamma = b(a^\top v - y) = (z + \zeta)(r + \xi^\top v)$, and independence of the two families of rounding decisions gives

$$\mathbb{E}[\gamma \mid z, y, v] = z r, \qquad \mathbb{E}[\gamma\gamma^\top \mid z, y, v] = \big(zz^\top + D_{p_{\mathrm{bw}}}(z)\big)\big(r^2 + v^\top D_{p_{\mathrm{fw}}}(z) v\big); \tag{29}$$

integrating $\tilde{\epsilon}$ (independent of z) replaces r^2 by $(z^\top\delta)^2 + \sigma_{\text{eff}}^2$. Averaging the fresh z in the first identity gives $\mathbb{E}[\gamma \mid v] = G\delta$.

(d) The four Gaussian moments. Given $\mathcal{F}_t$, $v = v_t$ is fixed and independent of the fresh $(z, \tilde{\epsilon})$; we integrate the sample first and then v_t (so $\delta\delta^\top \to C_t$, $vv^\top \to W_t$), with $D_r = D_{p_{r,t}}$. Expanding (29) yields four terms.

(M1) $\mathbb{E}[zz^\top r^2] = \mathsf{R}_t G + 2GC_tG$: by Isserlis, $\mathbb{E}_z[zz^\top(z^\top\delta)^2] = (\delta^\top G\delta)G + 2G\delta\delta^\top G$; the term linear in $\tilde{\epsilon}$ vanishes and $\tilde{\epsilon}^2$ contributes $\sigma_{\text{eff}}^2 G$.

(M2) $\mathbb{E}[zz^\top v^\top D_{\text{fw}}(z)v] = \sum_k (W_t)_{kk}\mathbb{E}[\nu_{p_{\text{fw},t}}(z_k)zz^\top] = \mathsf{F}_t G + 2GA_{a,t}G$, by (28) with $\nu = \nu_{p_{\text{fw},t}}$, since $\sum_k (W_t)_{kk}\frac{2\alpha_{p_{\text{fw},t}}(G_{kk})}{G_{kk}} Ge_ke_k^\top G = 2GA_{a,t}G$.

(M3) $\mathbb{E}[D_{\text{bw}}(z)r^2] = \mathsf{R}_t M_{b,t} + 2\Lambda_{b,t}(GC_tG)$: the m-th diagonal entry is $\mathbb{E}[\nu_{p_{\text{bw},t}}(z_m)(z^\top\delta)^2] + \sigma_{\text{eff}}^2 m_{p_{\text{bw},t}}(G_{mm})$, and (28) (weight index m) contracted with $\delta\delta^\top$ gives $m_{p_{\text{bw},t}}(G_{mm})\,\delta^\top G\delta + \frac{2\alpha_{p_{\text{bw},t}}(G_{mm})}{G_{mm}}(e_m^\top G\delta)^2$; then $\delta^\top G\delta \to \operatorname{tr}(GC_t)$, $(e_m^\top G\delta)^2 \to (GC_tG)_{mm}$.

(M4) $\mathbb{E}[D_{\text{bw}}(z)\, v^\top D_{\text{fw}}(z)v] = \mathsf{F}_t M_{b,t} + \Delta_{ab,t}$: the matrix is diagonal with m-th entry $\sum_k (W_t)_{kk}\mathbb{E}[\nu_{p_{\text{fw},t}}(z_k)\nu_{p_{\text{bw},t}}(z_m)] = \sum_k (W_t)_{kk}\big[m_{p_{\text{fw},t}}(G_{kk})m_{p_{\text{bw},t}}(G_{mm}) + c_{ab}(G_{kk}, G_{mm}, \rho_{km})\big]$ by the definition (25) of c_{ab}, the term $k = m$ (correlation $\rho_{mm} = 1$) included.

Summing, the one-sample second moment is

$$\mathbb{E}[\gamma_t\gamma_t^\top] = (\mathsf{R}_t + \mathsf{F}_t)(G + M_{b,t}) + 2GC_tG + 2GA_{a,t}G + 2\Lambda_{b,t}(GC_tG) + \Delta_{ab,t}. \tag{30}$$

(e) The mini-batch step. Let $\gamma_{t\ell} = b_\ell(a_\ell^\top v_t - y_\ell)$ and $\hat{g}_t = \frac{1}{B}\sum_{\ell \le B}\gamma_{t\ell}$. Given $\mathcal{F}_t$ the $\gamma_{t\ell}$ are i.i.d. (fresh samples, own rounding decisions) with mean $G\delta_t$ by (c), so $\mathbb{E}[\delta_{t+1} \mid \mathcal{F}_t] = H_t\delta_t$, which gives the mean recursion, $\tilde{\mu}_t = P_t(v_0 - v_*^S)$ and $\mathbb{E}[v_t \mid S, \theta_*] = v_*^S + P_t(v_0 - v_*^S)$. In $\mathbb{E}[\hat{g}_t\hat{g}_t^\top \mid \mathcal{F}_t]$ the B same-sample pairs contribute $\mathcal{S}_t(v_t) = \mathbb{E}[\gamma_{t1}\gamma_{t1}^\top \mid \mathcal{F}_t]$ each and the $B(B-1)$ distinct pairs contribute $G\delta_t\delta_t^\top G$ each, so $\mathbb{E}[\hat{g}_t\hat{g}_t^\top \mid \mathcal{F}_t] = \frac{1}{B}\mathcal{S}_t(v_t) + (1 - \frac{1}{B})G\delta_t\delta_t^\top G$. Expanding $\delta_{t+1}\delta_{t+1}^\top = (\delta_t - \eta_t\hat{g}_t)(\delta_t - \eta_t\hat{g}_t)^\top$ and using the conditional mean in the cross terms, $\mathbb{E}[\delta_{t+1}\delta_{t+1}^\top \mid \mathcal{F}_t] = H_t\delta_t\delta_t^\top H_t + \frac{\eta_t^2}{B}\big(\mathcal{S}_t(v_t) - G\delta_t\delta_t^\top G\big)$, and averaging over $\mathcal{F}_t$,

$$C_{t+1} = H_tC_tH_t + \frac{\eta_t^2}{B}\big(\mathbb{E}[\gamma_{t1}\gamma_{t1}^\top] - GC_tG\big).$$

The subtracted matrix is GC_tG with the *full* second moment C_t (not the covariance); the bracket equals $\mathbb{E}\operatorname{Cov}(\gamma_{t1} \mid \mathcal{F}_t) \succeq 0$. Substituting (30) leaves exactly one GC_tG and proves (9). The closure of Theorem 1 is $v_t = \delta_t + v_*^S$; W_t enters (30) only through $(W_t)_{kk}$; the risk identity is $\mathbb{E}[\delta_t^\top G\delta_t] = \operatorname{tr}(GC_t)$; and with $p_{\text{fw},t} = p_{\text{bw},t} = \infty$ all cell coefficients vanish. □

General initialization. The proof does not use $v_0 = 0$: for any deterministic v_0, the mean recursion, (9), its closure and the risk identity hold as stated with $\tilde{\mu}_0 = v_0 - v_*^S$ and $C_0 = (v_0 - v_*^S)(v_0 - v_*^S)^\top$.

Theorem 1 first says that rounding does not change the mean: $\tilde{\mu}_t$ follows exact-arithmetic gradient descent whatever the precisions, so SR changes the expected excess risk only through the covariance of the error. The forward term is driven by the size of the parameters, since the forward copy multiplies v_t in the prediction, and it is *aligned* with G, largest on the leading modes; as the norm of the best predictor grows with the width in the regime $1 < b < a$ (§2), this term grows through the run and with the width. The backward term is driven by the size of the residual, since the backward copy multiplies $a_\ell^\top v_t - y_\ell$ in the gradient, and it is *isotropic*, of the same order in every feature coordinate. In other words, backward rounding is a replica of the SGD noise scaled by about $0.09 \cdot 4^{-p_{\text{bw}}}$, whereas forward rounding is a different noise, with a different driver and a different geometry. The interaction bracket is present only when both copies are rounded: it is the part of the noise that an experiment rounding a single copy cannot see. Moreover, SR adds no bias, and raising either precision at any step can only lower the expected risk of every fixed-weight output, with no condition on the step sizes (Corollary B.5); thus at a fixed number of steps more bits are always better, and the trade-off between bits and steps is the subject of §4.

B.4 AVERAGING OVER THE TARGET: COROLLARY B.4

Let $\Sigma_{\perp S} = \Sigma_d - \Sigma_d S^\top G^{-1} S \Sigma_d$ be the matrix in the approximation error of (2), so that $A_S = \frac{1}{2}\theta_*^\top \Sigma_{\perp S}\theta_*$, and define the target averages

$$\bar{C}_t = \mathbb{E}_{\theta_*} C_t, \quad \bar{W}_t = \mathbb{E}_{\theta_*} W_t, \quad V_* = \mathbb{E}_{\theta_*}[v_*^S v_*^{S\top}], \quad \bar{\sigma}^2_{\text{eff}} = \mathbb{E}_{\theta_*}\sigma^2_{\text{eff}} = \sigma^2 + \operatorname{tr}(C_\theta \Sigma_{\perp S}). \quad (31)$$

For $s \le t$ write $P_{s\to t} = H_{t-1}\cdots H_s$ for the product of the exact-arithmetic steps from step s to step t, and $P_t = P_{0\to t}$.

Corollary B.4 (averaging over the target)**.** *The right-hand side of* (9) *is linear in* $(C_t, W_t, \sigma^2_{\text{eff}})$ *with coefficients free of* θ_**, so the target averages* $\bar{C}_t, \bar{W}_t$ *obey* (9) *with* $\sigma^2_{\text{eff}} \to \bar{\sigma}^2_{\text{eff}} = \sigma^2 + \operatorname{tr}(C_\theta \Sigma_{\perp S})$*,* $\Sigma_{\perp S} = \Sigma_d - \Sigma_d S^\top G^{-1} S \Sigma_d$*, and the closure* $\bar{C}_0 = V_*$*,* $\bar{W}_t = \bar{C}_t + V_* - P_t V_* - V_* P_t$*,* $V_* = G^{-1} S \Sigma_d C_\theta \Sigma_d S^\top G^{-1}$*.*

Proof. Since $v_*^S = G^{-1} S \Sigma_d \theta_*$ is linear in the centred teacher, $\mathbb{E}_{\theta_*} v_*^S = 0$, $\mathbb{E}_{\theta_*}[v_*^S v_*^{S\top}] = V_*$, and $\mathbb{E}_{\theta_*}[\theta_*^\top \Sigma_{\perp S}\theta_*] = \operatorname{tr}(C_\theta \Sigma_{\perp S}) = \operatorname{tr}(C_\theta \Sigma_d) - \operatorname{tr}(G^{-1} S \Sigma_d C_\theta \Sigma_d S^\top) = \sum_j \lambda_j \vartheta_j - \operatorname{tr}(G V_*)$, which gives $\bar{A}_S$ and $\bar{\sigma}^2_{\text{eff}} = \sigma^2 + 2\bar{A}_S$. Write $\Phi_t(C, W, s)$ for the right-hand side of (9) with $\mathsf{R} = s + \operatorname{tr}(GC)$ and $\mathsf{F} = \operatorname{tr}(M_{a,t} W)$. Each of $C \mapsto H_t C H_t$, $C \mapsto GCG$, $C \mapsto \Lambda_{b,t}(GCG)$, $W \mapsto A_{a,t}(W)$, $W \mapsto \Delta_{ab,t}(W)$ and $(C, W, s) \mapsto (s + \operatorname{tr}(GC) + \operatorname{tr}(M_{a,t}W))(G + M_{b,t})$ is linear with coefficients that are functions of $(G, \eta_t, B, p_{\text{fw},t}, p_{\text{bw},t})$ only, hence free of θ_* (the schedules are deterministic); there is no constant term and no product of two teacher-dependent quantities. By linearity of $\mathbb{E}_{\theta_*}$ and integrability, $\mathbb{E}_{\theta_*}\Phi_t(C_t, W_t, \sigma^2_{\text{eff}}) = \Phi_t(\bar{C}_t, \bar{W}_t, \bar{\sigma}^2_{\text{eff}})$, so the averaged state obeys (9). Averaging the closure of Theorem 1 with $\tilde{\mu}_t = P_t(v_0 - v_*^S)$, the terms with a single factor v_*^S vanish and $\mathbb{E}_{\theta_*}[v_*^S \tilde{\mu}_t^\top] = -V_* P_t$, giving the closure of Corollary B.4; the diagonal identity uses the symmetry of V_* and P_t. Finally $\bar{K}_t = \bar{C}_t - \mathbb{E}_{\theta_*}[\tilde{\mu}_t \tilde{\mu}_t^\top] = \bar{C}_t - P_t(v_0 v_0^\top + V_*)P_t$ is an average of conditional covariances, hence $\succeq 0$, and $\mathbb{E}[\mathcal{E}(v_t) \mid S] = \bar{A}_S + \frac{1}{2}\operatorname{tr}(G\bar{C}_t)$ is the averaged risk identity. □

B.5 COROLLARY B.5 AND ITS PROOF

Corollary B.5 (SR is unbiased noise; more bits never hurt)**.** *For every precision schedule,* $\mathbb{E}[v_t \mid S, \theta_*] = v_*^S + P_t(v_0 - v_*^S)$*. If* $p'_{\text{fw},t} \ge p_{\text{fw},t}$ *and* $p'_{\text{bw},t} \ge p_{\text{bw},t}$ *for all* t*, with everything else equal, then* $C'_t \preceq C_t$ *and* $\mathbb{E}[\mathcal{E}(v'_{\text{out}})] \le \mathbb{E}[\mathcal{E}(v_{\text{out}})]$ *for every output* $v_{\text{out}} = \sum_{t\le T}\omega_t v_t$ *with fixed real weights, with no restriction on the step sizes.*

The proof compares the two runs step by step through a positive map; in particular, comparing a precision schedule with exact arithmetic ($p'_{\text{fw},t} = p'_{\text{bw},t} = \infty$) shows that the rounding penalty of §2 is nonnegative for every such output, so the number of bits that keeps the penalty below a given fraction of the exact-arithmetic risk is well defined (§4).

Positive maps. $X \succeq Y$ means that $X - Y$ is symmetric positive semidefinite. Two facts are used: $\mathsf{A} X \mathsf{A}^\top \succeq 0$ for every A when $X \succeq 0$ (congruence), and $\operatorname{tr}(\mathsf{K}X) \ge 0$ when $\mathsf{K}, X \succeq 0$. A linear map $\mathcal{J}(X) = \sum_i c_i \mathsf{A}_i X \mathsf{A}_i^\top$ with $c_i \ge 0$, or an expectation of such maps, is therefore *positive*: $X \succeq 0 \Rightarrow \mathcal{J}(X) \succeq 0$, hence $X \succeq Y \Rightarrow \mathcal{J}(X) \succeq \mathcal{J}(Y)$. The mean statement of the corollary is the mean recursion of Theorem 1, which is gradient descent on $\mathcal{E}$ and does not involve the precisions.

Step 1: the batch augmented map.

Lemma B.6 (batch augmented map and its precision order)**.** *Fix* (S, θ_*)*, a step* t*, and the batch* $\mathcal{D} = (z_\ell, y_\ell)_{\ell \le B}$*. Let* $u = (v^\top, 1)^\top \in \mathbb{R}^{M+1}$*,* $\mathsf{c}_\ell = (z_\ell^\top, 0)^\top$*,* $\mathsf{r}_\ell = (z_\ell^\top, -y_\ell)^\top$*,* $\mathsf{f}_j = (e_j^\top, 0)^\top$*, and*

$$\mathsf{A}_0 = I_{M+1} - \tfrac{\eta_t}{B}\textstyle\sum_{\ell\le B}\mathsf{c}_\ell \mathsf{r}_\ell^\top, \quad \mathsf{A}^{\text{fw}}_{\ell j} = -\tfrac{\eta_t}{B}\mathsf{c}_\ell \mathsf{f}_j^\top, \quad \mathsf{A}^{\text{bw}}_{\ell k} = -\tfrac{\eta_t}{B}\mathsf{f}_k \mathsf{r}_\ell^\top, \quad \mathsf{A}^{\text{fwbw}}_{\ell k j} = -\tfrac{\eta_t}{B}\mathsf{f}_k \mathsf{f}_j^\top. \quad (32)$$

Let $\xi_{\ell j} = (a_\ell)_j - z_{\ell j}$ *and* $\zeta_{\ell k} = (b_\ell)_k - z_{\ell k}$ *be the* $2MB$ *rounding errors of the batch, and* $d^{\text{fw}}_{\ell j} = \nu_{p_{\text{fw},t}}(z_{\ell j})$*,* $d^{\text{bw}}_{\ell k} = \nu_{p_{\text{bw},t}}(z_{\ell k})$*. Then, given* $\mathcal{D}$ *and* $v_t = v$*,*

$$u_{t+1} = \Big[\mathsf{A}_0 + \sum_{\ell,j}\xi_{\ell j}\mathsf{A}^{\text{fw}}_{\ell j} + \sum_{\ell,k}\zeta_{\ell k}\mathsf{A}^{\text{bw}}_{\ell k} + \sum_{\ell,k,j}\xi_{\ell j}\zeta_{\ell k}\mathsf{A}^{\text{fwbw}}_{\ell k j}\Big]u, \quad (33)$$

and the coin average of $u_{t+1}u_{t+1}^\top$ is $\mathcal{J}_{\mathcal{D}}^p(uu^\top)$, where for symmetric X

$$\mathcal{J}_{\mathcal{D}}^p(X) = \mathsf{A}_0 X \mathsf{A}_0^\top + \sum_{\ell,j} d_{\ell j}^{\mathrm{fw}} \mathsf{A}_{\ell j}^{\mathrm{fw}} X \mathsf{A}_{\ell j}^{\mathrm{fw}\top} + \sum_{\ell,k} d_{\ell k}^{\mathrm{bw}} \mathsf{A}_{\ell k}^{\mathrm{bw}} X \mathsf{A}_{\ell k}^{\mathrm{bw}\top} + \sum_{\ell,k,j} d_{\ell j}^{\mathrm{fw}} d_{\ell k}^{\mathrm{bw}} \mathsf{A}_{\ell k j}^{\mathrm{fwbw}} X \mathsf{A}_{\ell k j}^{\mathrm{fwbw}\top}. \tag{34}$$

$\mathcal{J}_{\mathcal{D}}^p$ is a positive map. If $p'_{\mathrm{fw},t} \geq p_{\mathrm{fw},t}$ and $p'_{\mathrm{bw},t} \geq p_{\mathrm{bw},t}$, then $\mathcal{J}_{\mathcal{D}}^p - \mathcal{J}_{\mathcal{D}}^{p'}$ is a positive map as well.

Proof. Expanding (6) with $a_\ell = z_\ell + \xi_\ell$, $b_\ell = z_\ell + \zeta_\ell$ gives $v_{t+1} = v - \frac{\eta_t}{B}\sum_\ell \big[z_\ell(z_\ell^\top v - y_\ell) + z_\ell \xi_\ell^\top v + \zeta_\ell(z_\ell^\top v - y_\ell) + \zeta_\ell \xi_\ell^\top v\big]$; the four brackets are the top blocks of $\mathsf{c}_\ell \mathsf{r}_\ell^\top u$, $\sum_j \xi_{\ell j} \mathsf{c}_\ell \mathsf{f}_j^\top u$, $\sum_k \zeta_{\ell k} \mathsf{f}_k \mathsf{r}_\ell^\top u$ and $\sum_{k,j} \xi_{\ell j}\zeta_{\ell k} \mathsf{f}_k \mathsf{f}_j^\top u$, and the last row of A_0 is $(0, \ldots, 0, 1)$ while the last rows of the noise matrices vanish, so the last coordinate of u_{t+1} is 1. This is (33): one clean matrix containing the batch *sum* (all B gradients use the same v_t; a batch is not a composition of B single-sample steps), and one pair of families of rounding decisions per occurrence ℓ.

Given $\mathcal{D}$ and v, the $2MB$ errors are independent and centred (the rounding decisions are independent across entries, samples and copies), with $\mathbb{E}\xi_{\ell j}^2 = d_{\ell j}^{\mathrm{fw}}$, $\mathbb{E}\zeta_{\ell k}^2 = d_{\ell k}^{\mathrm{bw}}$. Form $u_{t+1}u_{t+1}^\top$ from (33): it is a sum of products of pairs of the monomials 1, $\xi_{\ell j}$, $\zeta_{\ell k}$, $\xi_{\ell j}\zeta_{\ell k}$, each of which is square-free in the independent centred variables. The product of two *distinct* monomials contains some variable to the first power, and that variable is independent of all the other factors of the product, so the expectation of the product is zero. This covers every case: two different coordinates or occurrences of the same family, the mixed pairs $\xi_{\ell j}\zeta_{\ell k}$ against $\xi_{\ell j}$ or $\zeta_{\ell k}$ (the other factor survives to the first power), the case $j = k$ (then $\xi_{\ell k}$ and $\zeta_{\ell k}$ are still two distinct variables), and monomials of distinct occurrences $\ell \neq \ell'$. The squares have expectations 1, $d_{\ell j}^{\mathrm{fw}}$, $d_{\ell k}^{\mathrm{bw}}$ and $\mathbb{E}[\xi_{\ell j}^2 \zeta_{\ell k}^2] = d_{\ell j}^{\mathrm{fw}} d_{\ell k}^{\mathrm{bw}}$ (independence of the two families). This proves (34) for $X = uu^\top$, hence for all symmetric X by linearity. Every coefficient in (34) is a cell variance or a product of two, hence nonnegative: $\mathcal{J}_{\mathcal{D}}^p$ is positive.

For the refinement, $\nu_{p'} \leq \nu_p$ pointwise when $p' \geq p$: every cell $[\ell, u]$ of $\mathbb{F}_p$ (including the cells ending at a binade boundary) is bisected at its midpoint c by $\mathbb{F}_{p+1}$, and $\nu_p(x) - \nu_{p+1}(x)$ equals $\frac{1}{2}(u-\ell)(x-\ell)$ for $x \in [\ell, c]$ and $\frac{1}{2}(u-\ell)(u-x)$ for $x \in [c, u]$, both nonnegative; iterate, and use $\nu_\infty = 0$. Thus $d^{\mathrm{fw}} - d'^{\mathrm{fw}} \geq 0$, $d^{\mathrm{bw}} - d'^{\mathrm{bw}} \geq 0$ and $d^{\mathrm{fw}} d^{\mathrm{bw}} - d'^{\mathrm{fw}} d'^{\mathrm{bw}} = (d^{\mathrm{fw}} - d'^{\mathrm{fw}}) d^{\mathrm{bw}} + d'^{\mathrm{fw}}(d^{\mathrm{bw}} - d'^{\mathrm{bw}}) \geq 0$ at every batch entry, while the matrices (32) do not depend on the precisions. Hence $\mathcal{J}_{\mathcal{D}}^p - \mathcal{J}_{\mathcal{D}}^{p'}$ is (34) with the coefficients replaced by their nonnegative differences: a positive map. The comparison is between the two rounding laws; no coupling of the coarse and fine rounding decisions is used. □

Step 2: propagation at a fixed data stream. Fix the whole stream $\mathbf{D} = (\mathcal{D}_0, \ldots, \mathcal{D}_{T-1})$ of batches (a realization of the fresh data; only its independence from the rounding decisions is used) and let $\mathcal{M}_{t|\mathbf{D}}^p = \mathbb{E}_{\mathrm{SR}}[u_t u_t^\top \mid S, \theta_*, \mathbf{D}]$. The rounding decisions of step t are independent of u_t given $\mathbf{D}$, so Lemma B.6 gives $\mathcal{M}_{t+1|\mathbf{D}}^p = \mathcal{J}_{\mathcal{D}_t}^p(\mathcal{M}_{t|\mathbf{D}}^p)$. Both runs start from $u_0 = (v_0^\top, 1)^\top$, so $E_0 = 0$ for $E_t = \mathcal{M}_{t|\mathbf{D}}^p - \mathcal{M}_{t|\mathbf{D}}^{p'}$, and

$$E_{t+1} = \mathcal{J}_{\mathcal{D}_t}^p(E_t) + \big(\mathcal{J}_{\mathcal{D}_t}^p - \mathcal{J}_{\mathcal{D}_t}^{p'}\big)\big(\mathcal{M}_{t|\mathbf{D}}^{p'}\big). \tag{35}$$

If $E_t \succeq 0$, the first term is $\succeq 0$ because $\mathcal{J}_{\mathcal{D}_t}^p$ is positive, and the second is $\succeq 0$ because the difference map is positive and is applied to the refined run's second moment $\mathcal{M}_{t|\mathbf{D}}^{p'} \succeq 0$. By induction $E_t \succeq 0$ for all $t \leq T$; averaging the nonnegative quadratic forms $w^\top E_t w$ over the stream gives $\mathcal{M}_t - \mathcal{M}'_t \succeq 0$. (The rounding-only inequality holds for every fixed numerical stream; the Gaussian recursion (9) is not asserted at a fixed stream.)

Step 3: readouts. With $J = [I, -v_*^S] \in \mathbb{R}^{M \times (M+1)}$, $C_t = J\mathcal{M}_t J^\top$ and $W_t = [I, 0]\mathcal{M}_t[I, 0]^\top$, so $C'_t \preceq C_t$ and $W'_t \preceq W_t$ by congruence, which is the order relation of Corollary B.5. The excess risk is a positive-semidefinite readout with the factor $\frac{1}{2}$,

$$\mathbb{E}[\mathcal{E}(v_t) \mid S, \theta_*] = A_S + \mathrm{tr}(\mathsf{K}_{\mathrm{last}} \mathcal{M}_t), \qquad \mathsf{K}_{\mathrm{last}} = \tfrac{1}{2} J^\top G J \succeq 0,$$

so the difference of the two expected risks is $\mathrm{tr}(\mathsf{K}_{\mathrm{last}}(\mathcal{M}_t - \mathcal{M}'_t)) \geq 0$; A_S is common to both runs.

Step 4: weighted outputs. For fixed real weights $\omega_1, \dots, \omega_T$ append the exact register $s_0 = 0$, $s_{t+1} = s_t + \omega_{t+1} v_{t+1}$, so that $v_{\mathrm{out}} = s_T$, and let $\tilde{u} = (v^\top, s^\top, 1)^\top \in \mathbb{R}^{2M+1}$. Writing the matrices of (32) in blocks, $\mathsf{A}_0 = \left(\begin{smallmatrix} R_0 & r_0 \\ 0 & 1 \end{smallmatrix}\right)$ and $\mathsf{A}_i = \left(\begin{smallmatrix} R_i & r_i \\ 0 & 0 \end{smallmatrix}\right)$ for every noise matrix, the update of $\tilde{u}$ is (33) with the *same* monomials and coefficients and the lifted matrices

$$\tilde{\mathsf{A}}_0 = \begin{pmatrix} R_0 & 0 & r_0 \\ \omega_{t+1} R_0 & I & \omega_{t+1} r_0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \tilde{\mathsf{A}}_i = \begin{pmatrix} R_i & 0 & r_i \\ \omega_{t+1} R_i & 0 & \omega_{t+1} r_i \\ 0 & 0 & 0 \end{pmatrix}. \tag{36}$$

The second block row of (36) produces $s + \omega_{t+1} v^+$. Lemma B.6 and the induction (35) apply verbatim to $\tilde{\mathcal{M}}_t = \mathbb{E}[\tilde{u}_t \tilde{u}_t^\top]$ (congruence positivity needs no sign of the entries, hence no sign or normalization of the weights), giving $\tilde{\mathcal{M}}'_T \preceq \tilde{\mathcal{M}}_T$, and $\mathbb{E}[\mathcal{E}(v_{\mathrm{out}}) \mid S, \theta_*] = A_S + \mathrm{tr}(\mathsf{K}_{\mathrm{av}} \tilde{\mathcal{M}}_T)$ with $\mathsf{K}_{\mathrm{av}} = \frac{1}{2} J_{\mathrm{av}}^\top G J_{\mathrm{av}} \succeq 0$, $J_{\mathrm{av}} = [0, I, -v_*^S]$. The uniform average is $\omega_t \equiv 1/T$; the last iterate is $\omega_T = 1, \omega_{t<T} = 0$.

Step 5: teacher average. Each inequality above says that a scalar $(w^\top(\mathcal{M}_t - \mathcal{M}'_t)w$ for fixed w, or the difference of two expected risks) is nonnegative; it has finite teacher expectation by the moment bound (and Cauchy–Schwarz for the register), and integrating it over θ_* at fixed S preserves the sign. (One must not average $\mathsf{K}_{\mathrm{last}}$ and $\mathcal{M}_t$ separately.) Nowhere was a sign of α_p or c_{ab} used: the comparison is made in (34), before the Gaussian average is regrouped into the coefficients of (9). □

Remark B.7 (scope of the comparison argument). The proof uses only that $\nu_{p'} \le \nu_p$ pointwise when $p' \ge p$ (Step 1) and the independence and centring of the rounding decisions at every rounding location; no sign of α_p or c_{ab} and no step-size condition enter, and Step 3 also gives $W'_t \preceq W_t$ under the hypotheses of Corollary B.5. Behind Step 1, Lemma B.6 rests on two properties of the two-read graph: after conditioning on the batch, the matrices (32) and the variance weights depend on the data, not on the state, and the mean of the rounding decision is precision-free. Rounding the weight read, the residual $a_\ell^\top v - y_\ell$, the completed gradient $\hat{g}$ or the stored update v^+ injects covariances whose cell arguments are state-dependent (for storage, $D_{p_u}(v^+)$): the second moment of $(Q_p(\cdot), 1)$ is then not a function of the incoming second moment, since two laws with equal first two moments can round to different second moments, and there is no state-independent positive map through which to push the order. For shared copies (one rounded copy $a = b = Q_p(z)$, or the shared input copy P2 of Appendix C) the mean of the rounding decision acquires the precision-dependent term $D_p(z)v$ or $S D_{p_x}(x) S^\top v$, the monomials 1 and e^2 are no longer orthogonal, and the coarse-minus-fine augmented map is indefinite. At these locations the argument does not establish monotonicity and does not show that it fails, and no comparison of schedules with equal total bit budget follows from it.

B.6 Two-time moments: Corollary B.8

Corollary B.8 (two-time moments; risk of the average)**.** $\mathbb{E}[\delta_t \delta_s^\top] = P_{s\to t} C_s$ *for* $s \le t$, $P_{s\to t} = H_{t-1} \cdots H_s$*; hence* $\mathbb{E}[\mathcal{E}(\bar{v}_T) \mid S, \theta_*] = A_S + \frac{1}{2T^2} \mathrm{tr}\big(G \sum_{t \le T} (X_t + X_t^\top - C_t)\big)$, $X_1 = C_1$, $X_{t+1} = H_t X_t + C_{t+1}$*: one extra product per step.*

Proof. From (e), $\mathbb{E}[\delta_{u+1} \mid \mathcal{F}_u] = H_u \delta_u$; the tower property and induction on t give $\mathbb{E}[\delta_t \mid \mathcal{F}_s] = P_{s\to t} \delta_s$ for $t \ge s$, and multiplying by the $\mathcal{F}_s$-measurable $\delta_s^\top$, $\mathbb{E}[\delta_t \delta_s^\top] = P_{s\to t} C_s$ (the propagator stands on the left; the transpose $C_s P_{s\to t}$ supplies the other half of each cross-time pair). For $\sum_t \omega_t = 1$, $v_{\mathrm{out}} - v_*^S = \sum_t \omega_t \delta_t$, and expanding its outer product gives $\sum_t \omega_t^2 C_t + \sum_{s<t} \omega_s \omega_t (P_{s\to t} C_s + C_s P_{s\to t})$. For uniform weights put $X_t = \sum_{s \le t} P_{s\to t} C_s$; then $X_1 = C_1$, $X_{t+1} = H_t X_t + C_{t+1}$ because $P_{s\to t+1} = H_t P_{s\to t}$, and $X_t + X_t^\top - C_t = C_t + \sum_{s<t} (P_{s\to t} C_s + C_s P_{s\to t})$; summing over $t \le T$, dividing by T^2 and using $\mathcal{E}(v) = A_S + \frac{1}{2}\|v - v_*^S\|_G^2$ gives the formula of Corollary B.8. No commutation of C_s with a propagator is used. The coefficients of these finite sums are teacher-free and every cross-time product is integrable (Cauchy–Schwarz), so teacher averaging replaces C_s by $\bar{C}_s$.

□

C ROUNDING THE INPUT: PROPOSITION C.1 AND ITS PROOFS

Our main results take the input x to be exact. If the input is stored at p_x bits before the feature map, the two protocols of §2 behave differently; the following summarizes what Proposition C.1 below proves. With independent copies of the rounded input, the mean of the iterate still follows exact-arithmetic gradient descent and C_t still obeys an exact affine recursion: the rounded input adds variance and no bias. For a shared rounded copy, let $\bar{\Gamma} = S\,\mathrm{diag}_j(m_{p_x}(\lambda_j))S^\top$ be the matrix of the mean input-rounding variances seen through the feature map; it satisfies $0.0814 \cdot 4^{-p_x} G \preceq \bar{\Gamma} \preceq \frac{1}{4}4^{-p_x} G$. The mean update is then gradient descent on $\mathcal{E}(v) + \frac{1}{2}v^\top \bar{\Gamma} v$, so SGD turns into ridge regression: for step sizes below the stability threshold of this ridge problem, the mean iterate converges to a ridge minimizer, whose excess risk is $\asymp 16^{-p_x}\|v_*^S\|_G^2$. This is a bias floor at every batch size, and it stays below the approximation error of the feature map, of order M^{1-b}, if and only if $p_x \gtrsim \frac{b-1}{4}\log_2 M$. The rounded-input dynamics can be computed through a Gaussian approximation of the input-rounding error. Section 5 and Appendix G measure the accuracy of this approximation, which is not proved.

Proposition C.1 (rounding the input: a shared copy is a ridge, independent copies are variance)**.** *Let the input be rounded to p_x bits before the feature map and $\bar{\Gamma} = S\,\mathrm{diag}_j(m_{p_x}(\lambda_j))S^\top$, so $0.0814 \cdot 4^{-p_x} G \preceq \bar{\Gamma} \preceq \frac{1}{4}4^{-p_x} G$. (i) With independent copies the mean is exact-arithmetic gradient descent and C_t obeys an exact affine recursion (Theorem C.4): no bias floor. (ii) With a shared copy the mean gradient is that of $\mathcal{E}(v) + \frac{1}{2}v^\top \bar{\Gamma} v$, and for $\eta_t \lambda_{\max}(G + \bar{\Gamma}) \le c < 1$, $\sum_t \eta_t = \infty$, $\mathbb{E}v_t$ converges to a ridge minimiser with excess risk $\asymp 16^{-p_x}\|v_*^S\|_G^2$ (Theorem C.8): a bias floor at every batch size, below the exact-arithmetic order M^{1-b} iff $p_x \gtrsim \frac{b-1}{4}\log_2 M$. (iii) A Gaussian approximation of the rounded-input error gives a computable solver whose accuracy is measured (Appendix G), not proved.*

Proof. The Loewner bounds on $\bar{\Gamma}$ are (37). (i) is Theorem C.4: its mean recursion $\tilde{\mu}_{t+1} = (I - \eta_t G)\tilde{\mu}_t$ is that of exact-arithmetic gradient descent, so there is no bias floor. (ii) The mean gradient and the convergence of $\mathbb{E}v_t$ to the ridge minimiser are Theorem C.8, and the two-sided order $\asymp 16^{-p_x}\|v_*^S\|_G^2$ of its excess risk is (52) of Corollary C.9, which also shows that $\mathbb{E}\,\mathcal{E}(v_t) - A_S$ is eventually at least this floor. In the regime $1 < b < a$ of §2, $\|v_*^S\|_G^2 \asymp 1$, so the floor is $\asymp 16^{-p_x} = 4^{-2p_x}$, and it is at most a constant times M^{1-b} if and only if $p_x \ge \frac{b-1}{4}\log_2 M - O(1)$. (iii) states what Appendix G measures and is not a mathematical claim. □

Standing assumptions are those of (6): fresh i.i.d. (x, ϵ) at every step, S fixed of full row rank, θ_* fixed, deterministic v_0 and schedules $(\eta_t, p_{x,t}, p_{\mathrm{fw},t}, p_{\mathrm{bw},t})$ with precisions in $\mathbb{N}_0 \cup \{\infty\}$, no clipping, independent rounding decisions at every rounding location, each centred given its argument; $\mathbb{E}$ averages data, labels and rounding decisions conditionally on (S, θ_*), and the time index is suppressed within one step.

Notation. For a current iterate v put $w = S^\top v$, $u = w - \theta_* \in \mathbb{R}^d$ and $r = z^\top v - y = x^\top u - \epsilon$; then $S\Sigma_d u = G\delta$ and $u^\top \Sigma_d u = \delta^\top G \delta + 2A_S$ (projection orthogonality). The rounded input is $\hat{x} = x + e$ with, given x, independent centred coordinates e_j of variance $d_j(x) = \nu_{p_x}(x_j)$; $D_x(x) = \mathrm{diag}_j\, d_j(x)$, $\Gamma(x) = SD_x(x)S^\top = \mathrm{Cov}(Se \mid x)$, $\eta = Se$ (the sketched input error; the step size always carries its index η_t), $\hat{z} = z + \eta$. Under P3, $a = z + \eta^{(1)} + \xi$, $b = z + \eta^{(2)} + \zeta$ with $e^{(1)}, e^{(2)}$ independent given x; under P2, $a = \hat{z} + \xi$, $b = \hat{z} + \zeta$; in both, $\mathbb{E}[\xi\xi^\top \mid \hat{z}^{(1)}] = D_{p_{\mathrm{fw}}}(\hat{z}^{(1)})$, $\mathbb{E}[\zeta\zeta^\top \mid \hat{z}^{(2)}] = D_{p_{\mathrm{bw}}}(\hat{z}^{(2)})$, and $g = b(a^\top v - y)$. Input-cell coefficients: $m_j = m_{p_x}(\lambda_j) = \mathbb{E}d_j(x)$, $\alpha_j = \alpha_{p_x}(\lambda_j)$, $V_j = \mathbb{E}d_j(x)^2 - m_j^2$; $\bar{\Gamma} = S\,\mathrm{diag}_j(m_j)S^\top = \mathbb{E}\Gamma(x)$, $\hat{G} = G + \bar{\Gamma}$. State functionals: $\mu_t = \mathbb{E}v_t$, $\tilde{\mu}_t = \mu_t - v_*^S$, $\omega_j = (S^\top W_t S)_{jj} = \mathbb{E}w_j^2$, $\mathcal{U}_t = \mathbb{E}[uu^\top] = S^\top W_t S - S^\top \mu_t \theta_*^\top - \theta_* \mu_t^\top S + \theta_* \theta_*^\top$, $\bar{U}_j = (\mathcal{U}_t)_{jj}$, $\mathsf{L}_j = \mathbb{E}[w_j u_j] = \omega_j - \theta_{*,j} s_j^\top \mu_t$, $\mathsf{R}_t = \sigma_{\mathrm{eff}}^2 + \mathrm{tr}(GC_t) = \sigma^2 + \mathrm{tr}(\Sigma_d \mathcal{U}_t)$, $\mathsf{F}_{x,t} = \mathrm{tr}(\bar{\Gamma} W_t) = \sum_j m_j \omega_j$, and $W_t = C_t + v_*^S v_*^{S\top} + v_*^S \tilde{\mu}_t^\top + \tilde{\mu}_t v_*^{S\top}$. Post-sketch coefficients at the dressed variances $\hat{G}_{kk} = G_{kk} + \bar{\Gamma}_{kk}$ carry a hat: $\hat{M}_{a,t} = \mathrm{diag}_k\, m_{p_{\mathrm{fw},t}}(\hat{G}_{kk})$, $\hat{M}_{b,t} = \mathrm{diag}_k\, m_{p_{\mathrm{bw},t}}(\hat{G}_{kk})$, $\hat{\mathsf{F}}_t = \mathrm{tr}(\hat{M}_{a,t} W_t)$, $\hat{A}_{a,t} = \mathrm{diag}_k\big(\alpha_{p_{\mathrm{fw},t}}(\hat{G}_{kk})(W_t)_{kk}/\hat{G}_{kk}\big)$, $\hat{\Lambda}_{b,t}(X) = \mathrm{diag}_k\big(\alpha_{p_{\mathrm{bw},t}}(\hat{G}_{kk})X_{kk}/\hat{G}_{kk}\big)$; c_{xa}, c_{xb} denote the pair moment c_{ab} of (25) with precisions (p_x, p_{fw}), (p_x, p_{bw}).

Moments. $|Q_p(s)| \le 2|s|$ and $\nu_p(s) \le 4^{-p}s^2/4$ give $\|a\|, \|b\| \le 4\|S\|\|x\|$, hence $\|g\| \le c_S(\|x\|^2\|v\| + \|x\|\|y\|)$ with $c_S = \max(16\|S\|^2, 4\|S\|)$ and, as v_t is independent of the fresh sample, $\mathbb{E}\|v_{t+1}\|^2 \le (2 + 4\eta_t^2 c_S^2 \mathbb{E}\|x\|^4)\mathbb{E}\|v_t\|^2 + 4\eta_t^2 c_S^2 \mathbb{E}[\|x\|^2 y^2]$: all second moments exist at every finite horizon, and every coefficient integrand below is dominated by a degree-four polynomial of the Gaussian input, which justifies all interchanges. *Aligned bound.* $a^\top \bar{\Gamma} a = \sum_j m_j (s_j^\top a)^2$, $a^\top G a = \sum_j \lambda_j (s_j^\top a)^2$ and $\kappa_- 4^{-p}\lambda \le m_p(\lambda) \le \frac{1}{4}4^{-p}\lambda$ (Lemma B.1; $\kappa_- = 0.0814$ is computer-assisted) give, with $q_x = 4^{-p_x}$,

$$\kappa_- q_x\, G \preceq \bar{\Gamma} \preceq \tfrac{1}{4} q_x\, G, \qquad \text{so } \hat{G} \succ 0. \tag{37}$$

C.1 The nine terms of the P3 gradient second moment

Lemma C.2 (P3 factorization)**.** *Let $\tilde{E}_{\mathrm{fw}}(x) = \mathbb{E}[D_{p_{\mathrm{fw}}}(z+Se) \mid x]$, $\tilde{E}_{\mathrm{bw}}(x) = \mathbb{E}[D_{p_{\mathrm{bw}}}(z+Se) \mid x]$ (true input-rounding law). Under P3, for fixed (x, ϵ, v),*

$$\mathbb{E}[g \mid x, \epsilon, v] = zr, \qquad \mathbb{E}[gg^\top \mid x, \epsilon, v] = \big(zz^\top + \Gamma(x) + \tilde{E}_{\mathrm{bw}}(x)\big)\big(r^2 + v^\top \Gamma(x) v + v^\top \tilde{E}_{\mathrm{fw}}(x) v\big), \tag{38}$$

hence $\mathbb{E}[g \mid v] = G\delta$, $\tilde{\mu}_{t+1} = (I - \eta_t G)\tilde{\mu}_t$ and $\mu_t = v_*^S + P_t(v_0 - v_*^S)$, $P_t = \prod_{s<t}(I - \eta_s G)$. Conditioning on (x, v) only replaces r by $x^\top u$ and r^2 by $(x^\top u)^2 + \sigma^2$ (an expression with a realized r on the right is not $\sigma(x, v)$-measurable).

Proof. Independence used: (I1) each rounding error is centred given its argument, with diagonal conditional covariance; (I2) given (x, ϵ, v), the rounding decisions $(e^{(1)}, \xi)$ of the forward branch are independent of the rounding decisions $(e^{(2)}, \zeta)$ of the backward branch; (I3) v is independent of the fresh sample and $\epsilon \perp x$. Since b is a function of $(x, e^{(2)}, \zeta)$ and $a^\top v - y = r + \eta^{(1)\top} v + \xi^\top v$ of $(x, \epsilon, v, e^{(1)}, \xi)$, (I2) gives $\mathbb{E}[g \mid x, \epsilon, v] = \mathbb{E}[b \mid x]\,\mathbb{E}[a^\top v - y \mid x, \epsilon, v]$ and likewise for $gg^\top$. By (I1) and the tower property $\mathbb{E}[\eta \mid x] = 0$, $\mathbb{E}[\zeta \mid x] = \mathbb{E}[\mathbb{E}[\zeta \mid \hat{z}] \mid x] = 0$, $\mathbb{E}[\eta\zeta^\top \mid x] = \mathbb{E}[\eta\,\mathbb{E}[\zeta \mid x, \eta]^\top \mid x] = 0$, so $\mathbb{E}[b \mid x] = z$, $\mathbb{E}[bb^\top \mid x] = zz^\top + \Gamma(x) + \tilde{E}_{\mathrm{bw}}(x)$ and $\mathbb{E}[(r + \eta^\top v + \xi^\top v)^2 \mid x, \epsilon, v] = r^2 + v^\top \Gamma(x) v + v^\top \tilde{E}_{\mathrm{fw}}(x) v$. Averaging ϵ, then x, in zr gives $S\Sigma_d u = G\delta$. □

Lemma C.3 (weighted Gaussian identity)**.** *For centred jointly Gaussian (X, Y, Z) with X scalar, $\mathrm{Var}\, X = g > 0$, and every p,*

$$\mathbb{E}[\nu_p(X) Y Z^\top] = m_p(g)\,\mathrm{Cov}(Y, Z) + \tfrac{2\alpha_p(g)}{g}\,\mathrm{Cov}(Y, X)\,\mathrm{Cov}(X, Z). \tag{39}$$

Proof. With $c = \mathrm{Cov}(Y, X)$, $d = \mathrm{Cov}(Z, X)$ write $Y = cX/g + Y_0$, $Z = dX/g + Z_0$; (Y_0, Z_0) is centred Gaussian, independent of X, with $\mathrm{Cov}(Y_0, Z_0) = \mathrm{Cov}(Y, Z) - cd^\top/g$. The terms linear in (Y_0, Z_0) vanish, so $\mathbb{E}[\nu_p(X) Y Z^\top] = \frac{cd^\top}{g^2}\mathbb{E}[\nu_p(X) X^2] + m_p(g)(\mathrm{Cov}(Y, Z) - \frac{cd^\top}{g})$, and $\mathbb{E}[\nu_p(X) X^2] = g(m_p(g) + 2\alpha_p(g))$ by (25). □

The four pre-sketch terms. Expanding (38) gives nine terms; the four involving only $z, r, \Gamma(x)$ are exact Gaussian moments, by (39) with $X = x_j$ and the covariances $\mathrm{Cov}(z, x_j) = \lambda_j s_j$, $\mathrm{Cov}(x_j, x^\top u) = \lambda_j u_j$, $\mathrm{Cov}(z, x^\top u) = G\delta$, $\mathrm{Var}(x^\top u) = u^\top \Sigma_d u$ (distinct x_j independent). Integrating the sample at fixed v and then v ($\delta\delta^\top \to C_t$, $vv^\top \to W_t$, $w_j^2 \to \omega_j$, $u_j^2 \to \bar{U}_j$):

$$\begin{aligned}
\mathcal{A}_1 &:= \mathbb{E}[zz^\top r^2] = \mathsf{R}_t G + 2GC_tG, \\
\mathcal{A}_2 &:= \mathbb{E}[zz^\top\, v^\top \Gamma(x) v] = \mathsf{F}_{x,t} G + 2S\,\mathrm{diag}_j(\alpha_j\lambda_j\omega_j)S^\top, \\
\mathcal{A}_3 &:= \mathbb{E}[\Gamma(x) r^2] = \mathsf{R}_t \bar{\Gamma} + 2S\,\mathrm{diag}_j(\alpha_j\lambda_j\bar{U}_j)S^\top, \\
\mathcal{A}_4 &:= \mathbb{E}[\Gamma(x)\, v^\top \Gamma(x) v] = \mathsf{F}_{x,t}\bar{\Gamma} + S\,\mathrm{diag}_j(V_j\omega_j)S^\top.
\end{aligned} \tag{40}$$

Indeed, $\mathcal{A}_1$ is Isserlis, $\mathbb{E}_x[zz^\top(x^\top u)^2] = (u^\top\Sigma_d u)G + 2(G\delta)(G\delta)^\top$, plus $\sigma^2 G$ from ϵ; $\mathcal{A}_2$: $v^\top\Gamma(x)v = \sum_j d_j(x) w_j^2$ and $\mathbb{E}[d_j(x) zz^\top] = m_j G + 2\alpha_j\lambda_j s_j s_j^\top$; $\mathcal{A}_3$: $\mathbb{E}[d_j(x)(x^\top u)^2] = m_j u^\top\Sigma_d u + 2\alpha_j\lambda_j u_j^2$, plus $m_j\sigma^2$, times $s_j s_j^\top$; $\mathcal{A}_4$: $\mathbb{E}[d_j(x) d_{j'}(x)] = m_j m_{j'} + \mathbf{1}_{j=j'} V_j$ by independence of the coordinates—under P3 the two factors $\Gamma(x)$ are conditional variances of the two *independent* copies given the same x, so the $j = j'$ entry is $\mathbb{E}d_j(x)^2$, not $\mathbb{E}e_j^4$. Note that $\mathcal{A}_3$ depends on the mean through $\bar{U}_j$; it cannot be rewritten through ω_j or $G\delta$ alone.

The five true-law terms. The other five terms contain $\tilde{E}_{\mathrm{fw}}$ or $\tilde{E}_{\mathrm{bw}}$, the post-sketch cell variance at the rounded feature $\hat{z}_k = z_k + \eta_k$, whose law given x is a convolution of two-point laws. With the coefficient tensors (true law; the two copies in $\mathcal{P}$ are independent given x, so $\mathcal{P}_{mk} = \mathbb{E}_x[(\tilde{E}_{\mathrm{bw}})_{mm}(\tilde{E}_{\mathrm{fw}})_{kk}]$ is not a product of means)

$$\begin{aligned} \mathcal{Z}_{kl;m} &= \mathbb{E}[z_k z_l \nu_{p_{\mathrm{fw}}}(\hat{z}_m)]\ (M^3), \qquad \mathcal{T}_{m;jj'} = \mathbb{E}[x_j x_{j'} \nu_{p_{\mathrm{bw}}}(\hat{z}_m)]\ (Md^2), \\ \mathcal{X}^{\mathrm{fw}}_{j;m} &= \mathbb{E}[d_j(x)\nu_{p_{\mathrm{fw}}}(\hat{z}_m)], \qquad \mathcal{X}^{\mathrm{bw}}_{j;m} = \mathbb{E}[d_j(x)\nu_{p_{\mathrm{bw}}}(\hat{z}_m)]\ (dM \text{ each}), \\ \beta_m &= \mathbb{E}\nu_{p_{\mathrm{bw}}}(\hat{z}_m)\ (M), \qquad \mathcal{P}_{mk} = \mathbb{E}\big[\nu_{p_{\mathrm{bw}}}(\hat{z}^{(2)}_m)\nu_{p_{\mathrm{fw}}}(\hat{z}^{(1)}_k)\big]\ (M^2), \end{aligned} \tag{41}$$

the terms $\mathcal{E}_1 = \mathbb{E}[zz^\top v^\top \tilde{E}_{\mathrm{fw}} v]$, $\mathcal{E}_2 = \mathbb{E}[\Gamma(x)\, v^\top \tilde{E}_{\mathrm{fw}} v]$, $\mathcal{E}_3 = \mathbb{E}[\tilde{E}_{\mathrm{bw}} r^2]$, $\mathcal{E}_4 = \mathbb{E}[\tilde{E}_{\mathrm{bw}}\, v^\top \Gamma(x) v]$, $\mathcal{E}_5 = \mathbb{E}[\tilde{E}_{\mathrm{bw}}\, v^\top \tilde{E}_{\mathrm{fw}} v]$ are exactly

$$\begin{aligned} \mathcal{E}_1 &= \sum_k (W_t)_{kk} \mathcal{Z}_{\cdot;k}, \qquad \mathcal{E}_2 = S\,\mathrm{diag}_j\Big(\sum_k (W_t)_{kk}\mathcal{X}^{\mathrm{fw}}_{j;k}\Big) S^\top, \qquad \mathcal{E}_4 = \mathrm{diag}_m\Big(\sum_j \omega_j \mathcal{X}^{\mathrm{bw}}_{j;m}\Big), \\ \mathcal{E}_3 &= \mathrm{diag}_m\Big(\sum_{j,j'} \mathcal{T}_{m;jj'}(\mathcal{U}_t)_{jj'} + \sigma^2\beta_m\Big), \qquad \mathcal{E}_5 = \mathrm{diag}_m\Big(\sum_k (W_t)_{kk}\mathcal{P}_{mk}\Big) : \end{aligned} \tag{42}$$

integrate the input copies given x (diagonality of $\tilde{E}_{\mathrm{fw}}$ gives $v^\top \tilde{E}_{\mathrm{fw}} v = \sum_k v_k^2 (\tilde{E}_{\mathrm{fw}})_{kk}$), use $v \perp$ sample, and for $\mathcal{E}_3$ integrate ϵ and expand $(x^\top u)^2 = \sum_{j,j'} x_j x_{j'} u_j u_{j'}$. $\mathcal{E}_3$ is affine in (W_t, μ_t) through $\mathcal{U}_t$, not linear in W_t alone.

C.2 The exact P3 recursion (independent copies)

Theorem C.4 (P3, exact closure)**.** *Under P3, conditionally on (S, θ_*), $\tilde{\mu}_{t+1} = (I - \eta_t G)\tilde{\mu}_t$ and*

$$C_{t+1} = (I - \eta_t G) C_t (I - \eta_t G) + \frac{\eta_t^2}{B}\big(\mathcal{B}_t - G C_t G\big), \qquad \mathcal{B}_t = \mathbb{E}[g_t g_t^\top] = \textstyle\sum_{i=1}^4 \mathcal{A}_i + \sum_{i=1}^5 \mathcal{E}_i, \tag{43}$$

all coefficients (including the tensors (41)) at $(p_{x,t}, p_{\mathrm{fw},t}, p_{\mathrm{bw},t})$, $C_0 = (v_0 - v_*^S)(v_0 - v_*^S)^\top$, W_t from the closure relation. Substituting W_t and the known mean, $\mathcal{B}_t = \mathcal{L}_t(C_t) + \Phi_t$ with a linear map $\mathcal{L}_t$ depending only on (S, Σ_d, p_t) and a forcing Φ_t that is a polynomial of degree at most two in θ_* whose linear part carries v_0. At $p_x = \infty$ ($\bar{\Gamma} = 0$, $m_j = \alpha_j = V_j = 0$, $\hat{z} = z$) (43) is (9).

Proof. Write the batch gradient as $\bar{g}_t = G\delta_t + \varsigma_t$ with $\mathbb{E}[\varsigma_t \mid v_t] = 0$ (Lemma C.2); the B samples are i.i.d. given v_t, so $\mathbb{E}[\varsigma_t \varsigma_t^\top \mid v_t] = \frac{1}{B}(\mathbb{E}[gg^\top \mid v_t] - G\delta_t\delta_t^\top G)$. Expanding $\delta_{t+1} = (I - \eta_t G)\delta_t - \eta_t \varsigma_t$, the cross terms vanish, and averaging v_t gives (43), where the subtracted matrix is GC_tG with the *full* second moment C_t. The decisive bookkeeping: $\mathcal{A}_1$ contributes $2GC_tG$ to the raw moment $\mathcal{B}_t$, so the innovation $\mathcal{B}_t - GC_tG$ contains exactly $+GC_tG$, as in (9); using the innovation in place of $\mathcal{B}_t$ would remove GC_tG twice. The nine terms are (40), (42). Closure: R_t is affine in C_t; ω_j, $S^\top W_t S$ are linear in W_t; $\mathcal{U}_t = S^\top W_t S - S^\top \mu_t \theta_*^\top - \theta_* \mu_t^\top S + \theta_* \theta_*^\top$ with $\mu_t = P_t v_0 + (I - P_t) L\theta_*$, $L = G^{-1} S \Sigma_d$; and $W_t = C_t + v_*^S v_*^{S\top} + v_*^S \tilde{\mu}_t^\top + \tilde{\mu}_t v_*^{S\top}$ with $v_*^S = L\theta_*$, $\tilde{\mu}_t = P_t(v_0 - v_*^S)$. The coefficient tensors depend on (S, Σ_d, p_t) only, and σ^2_{eff} and the displayed products have degree at most two in θ_*. At $p_x = \infty$ the tensors reduce to the Gaussian moments of Theorem 1 and $\mathcal{A}_2 = \mathcal{A}_3 = \mathcal{A}_4 = 0$. □

Remark C.5 (the mean is part of the state). The coefficients of (43) depend on μ_t (through $\bar{U}_j$ in $\mathcal{A}_3$ and $\mathcal{U}_t$ in $\mathcal{E}_3$), which Theorem C.4 supplies in closed form. No autonomous map $(C_t, W_t) \mapsto C_{t+1}$ valid for all moment laws exists: for $M = 1$, $d = 2$, $S = (1, 2)$, $\Sigma_d = \lambda I$, $\theta_* = (2, -1)^\top$ (so $v_*^S = 0$), finite p_x and $p_{\mathrm{fw}} = p_{\mathrm{bw}} = \infty$, the laws $v = a$ and $v = -a$ have the same $C = W = a^2$ but their $\mathcal{A}_3$ terms differ by $48\alpha_{p_x}(\lambda)\lambda a \neq 0$ ($\alpha_p > 0$ by Lemma B.1), hence so do their next second moments.

Corollary C.6 (teacher average)**.** *Let θ_* be centred with $\mathbb{E}\theta_*\theta_*^\top = C_\theta$, independent of S and of the training randomness; put $L = G^{-1}S\Sigma_d$, $V_* = LC_\theta L^\top$, $\bar{C}_t = \mathbb{E}_{\theta_*} C_t$, $\bar{W}_t = \mathbb{E}_{\theta_*} W_t$. Then $K_t := \mathbb{E}[v_t \theta_*^\top \mid S] = (I - P_t) L C_\theta$, and $\bar{C}_t$ obeys (43) with every coefficient tensor unchanged and $\bar{W}_t = \bar{C}_t + V_* - P_t V_* - V_* P_t^\top$, $\bar{\mathcal{U}}_t = S^\top \bar{W}_t S - S^\top K_t - K_t^\top S + C_\theta$, $\bar{\sigma}^2_{\mathrm{eff}} = \sigma^2 + \mathrm{tr}\big(C_\theta(\Sigma_d - \Sigma_d S^\top G^{-1} S \Sigma_d)\big)$, $\bar{C}_0 = v_0 v_0^\top + V_*$; at $p_x = \infty$ this is Corollary B.4.*

Proof. $\mathcal{L}_t$ is teacher independent and Φ_t has degree ≤ 2 in θ_*, so averaging replaces $\theta_*\theta_*^\top$ by C_θ and kills the linear terms (the moment bound above is teacher integrable). K_t follows from $\mu_t = v_*^S + P_t(v_0 - v_*^S)$, $v_*^S = L\theta_*$, $\mathbb{E}\theta_* = 0$; then $\mathbb{E}_{\theta_*}[v_*^S \mu_t^\top] = V_* - V_* P_t^\top$ gives $\bar{W}_t$, $\mathbb{E}_{\theta_*}[\mu_t \theta_*^\top] = K_t$ gives $\bar{\mathcal{U}}_t$, and (2) gives $\bar{\sigma}^2_{\text{eff}}$. Iterate and teacher are dependent; no product of means replaces K_t. □

C.3 The Gaussian-smoothing oracle A′

Proposition C.7 (hybrid evaluation)**.** *In $\mathcal{E}_1, \dots, \mathcal{E}_5$ only, replace the conditional law of $\eta^{(1)}, \eta^{(2)}$ given x by independent $N(0, \bar{\Gamma})$ vectors independent of (x, ϵ) (equivalently: replace the independent triples $(x_j, e_j^{(1)}, e_j^{(2)})$ there by centred Gaussian triples with covariance $\mathrm{diag}(\lambda_j, m_j, m_j)$); $\mathcal{A}_1$–$\mathcal{A}_4$ keep the true law. Under this hybrid law the variables in $\mathcal{E}_i$ are, given v, jointly Gaussian with $\mathrm{Var}\,\hat{z}_k^{(i)} = \hat{G}_{kk}$, $\mathrm{Cov}(\hat{z}^{(i)}, z) = G$, $\mathrm{Cov}(\hat{z}_k^{(1)}, \hat{z}_m^{(2)}) = G_{km}$ also for $k = m$, $\mathrm{Cov}(\hat{z}_k^{(i)}, x_j) = \lambda_j S_{kj}$, $\mathrm{Cov}(\hat{z}^{(i)}, x^\top u) = G\delta$; the correlations are $\varkappa_{jk} = S_{kj}\sqrt{\lambda_j/\hat{G}_{kk}}$ and $\hat{\rho}_{km} = G_{km}/\sqrt{\hat{G}_{kk}\hat{G}_{mm}}$, with $\hat{\rho}_{kk} = G_{kk}/\hat{G}_{kk} \in [(1 + q_x/4)^{-1}, 1)$ strictly below one for $p_x < \infty$. The evaluated innovation is*

$$\begin{aligned}\mathcal{V}_t^{A'} &= (\mathsf{R}_t + \mathsf{F}_{x,t} + \hat{\mathsf{F}}_t)\big(G + \bar{\Gamma} + \hat{M}_{b,t}\big) + GC_tG + 2G\hat{A}_{a,t}G + 2\hat{\Lambda}_{b,t}(GC_tG) \\ &\quad + 2S\,\mathrm{diag}_j\big(\alpha_j\lambda_j(\omega_j + \bar{U}_j)\big)S^\top + S\,\mathrm{diag}_j\Big(V_j\omega_j + \sum_k (W_t)_{kk}\, c_{xa}(\lambda_j, \hat{G}_{kk}, \varkappa_{jk})\Big)S^\top \\ &\quad + \mathrm{diag}_m\Big(\sum_j \omega_j\, c_{xb}(\lambda_j, \hat{G}_{mm}, \varkappa_{jm}) + \sum_k (W_t)_{kk}\, c_{ab}(\hat{G}_{kk}, \hat{G}_{mm}, \hat{\rho}_{km})\Big),\end{aligned} \tag{44}$$

and the oracle recursion is $C_{t+1} = (I - \eta_t G)C_t(I - \eta_t G) + \frac{\eta_t^2}{B}\mathcal{V}_t^{A'}$ with the mean and W_t of Theorem C.4 (teacher average as in Corollary C.6). It is an exact identity for the hybrid law, it preserves positive semidefiniteness of $K_t = C_t - \tilde{\mu}_t\tilde{\mu}_t^\top$, and at $p_x = \infty$ it is the bracket of (9)*. No bound on its distance from* (43) *is claimed; that distance is measured in Section 5 and Appendix G, with the diagonal pair coefficients computed by the nested integral* (45)*.*

Proof. The covariances are direct ($\mathrm{Cov}(\hat{z}_k^{(1)}, \hat{z}_m^{(2)})$ has no $\bar{\Gamma}$ part because the two proxies are independent, also on the diagonal; $\hat{\rho}_{kk} \ge (1 + q_x/4)^{-1}$ by (37)). Lemma C.3 with $X = \hat{z}_k$ (variance $\hat{G}_{kk}$, cross covariances $G_{\cdot k}$ and $(G\delta)_k$, *not* $\hat{G}_{\cdot k}$) gives $\mathcal{E}_1^{A'} = \hat{\mathsf{F}}_t G + 2G\hat{A}_{a,t}G$ and $\mathcal{E}_3^{A'} = \mathsf{R}_t\hat{M}_{b,t} + 2\hat{\Lambda}_{b,t}(GC_tG)$; the definition of c_{ab} gives $\mathcal{E}_2^{A'} = \hat{\mathsf{F}}_t\bar{\Gamma} + S\,\mathrm{diag}_j(\sum_k (W_t)_{kk}c_{xa})S^\top$, $\mathcal{E}_4^{A'} = \mathsf{F}_{x,t}\hat{M}_{b,t} + \mathrm{diag}_m(\sum_j \omega_j c_{xb})$, $\mathcal{E}_5^{A'} = \hat{\mathsf{F}}_t\hat{M}_{b,t} + \mathrm{diag}_m(\sum_k (W_t)_{kk}c_{ab})$ with the arguments of (44). The nine products of energies and covariances in $\mathcal{A}_1$–$\mathcal{A}_4$, $\mathcal{E}_1^{A'}$–$\mathcal{E}_5^{A'}$ collect into $(\mathsf{R}_t + \mathsf{F}_{x,t} + \hat{\mathsf{F}}_t)(G + \bar{\Gamma} + \hat{M}_{b,t})$; the Wick term of $\mathcal{A}_1$ is $2GC_tG$, and subtracting GC_tG leaves (44). *Positivity:* if (μ_t, C_t, W_t) are the moments of some $v = \mu_t + K_t^{1/2}\mathsf{n}$, $\mathsf{n} \sim N(0, I)$, then $\mathcal{V}_t^{A'} = \mathbb{E}_v\,\mathrm{Cov}(zr \mid v) + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4 + \sum_i \mathcal{E}_i^{A'}$, and each summand is, under a genuine (true or hybrid) law, the expectation of a positive semidefinite matrix ($zz^\top$, $\Gamma(x)$, $D_{p_{\text{bw}}}$) times a nonnegative scalar (r^2, $v^\top\Gamma v$, $v^\top D_{p_{\text{fw}}} v$), hence $\succeq 0$; since $\tilde{\mu}_{t+1} = (I - \eta_t G)\tilde{\mu}_t$, $K_{t+1} = (I - \eta_t G)K_t(I - \eta_t G) + \frac{\eta_t^2}{B}\mathcal{V}_t^{A'}$, and induction from $K_0 = 0$ closes. *Reduction:* at $p_x = \infty$ the hybrid law is the true law, $\bar{\Gamma} = 0$, $\hat{G} = G$, $c_{xa} = c_{xb} = 0$, $\hat{\rho}_{km} = \rho_{km}$ ($= 1$ for $k = m$), and (44) is the bracket of (9). □

The diagonal pair coefficient in (44) is a functional of the whole pair law near $\hat{\rho}_{kk} \to 1$ and must be computed exactly, not by a low-degree Hermite (Mehler) truncation: conditionally on z_k the two proxies are independent $N(z_k, \bar{\Gamma}_{kk})$, so with the Gaussian smoothings $\mathcal{G}_{\text{fw}}(\mathfrak{z}; \gamma) = \mathbb{E}_{Z\sim N(0,1)}\nu_{p_{\text{fw}}}(\mathfrak{z} + \sqrt{\gamma}Z)$, likewise $\mathcal{G}_{\text{bw}}$, and φ_g the $N(0, g)$ density,

$$c_{ab}(\hat{G}_{kk}, \hat{G}_{kk}, \hat{\rho}_{kk}) + m_{p_{\text{fw}}}(\hat{G}_{kk})\, m_{p_{\text{bw}}}(\hat{G}_{kk}) = \int_{\mathbb{R}} \mathcal{G}_{\text{fw}}(\mathfrak{z}; \bar{\Gamma}_{kk})\, \mathcal{G}_{\text{bw}}(\mathfrak{z}; \bar{\Gamma}_{kk})\, \varphi_{G_{kk}}(\mathfrak{z})\, d\mathfrak{z}, \tag{45}$$

which the oracle implementation uses for all diagonal and near-diagonal pairs (Appendix G).

C.4 THE EXACT P2 RECURSION (SHARED COPY)

Under P2 the shared $\hat{z} = z + \eta$ enters both branches. Put $\hat{r} = \hat{z}^\top v - y$, $h_* = S\Sigma_d\theta_* = Gv_*^S$, $\mathsf{c}_t = \bar{\Gamma}_t v_*^S$, $\hat{H}_t = I - \eta_t\hat{G}_t$, $\mathfrak{g}(v) = \hat{G}v - h_* = G\delta + \bar{\Gamma}v = \hat{G}\delta + \mathsf{c}_t = \nabla_v[\mathcal{E}(v) + \frac{1}{2}v^\top\bar{\Gamma}v]$ and $\mathcal{J}_t = \mathbb{E}[\mathfrak{g}(v_t)\mathfrak{g}(v_t)^\top] = \hat{G}C_t\hat{G} + \hat{G}\tilde{\mu}_t\mathsf{c}_t^\top + \mathsf{c}_t\tilde{\mu}_t^\top\hat{G} + \mathsf{c}_t\mathsf{c}_t^\top$. Two-point cell moments: for x_j in the cell $[\ell_{p_x}(x_j), u_{p_x}(x_j)]$ of width Δ_j (endpoints as in Section 2), $e_j = u_{p_x}(x_j) - x_j$ with probability $(x_j - \ell_{p_x}(x_j))/\Delta_j$ and $e_j = \ell_{p_x}(x_j) - x_j$ otherwise, so $\mathbb{E}[e_j^2 \mid x] = \nu_{p_x}(x_j)$ and

$$\begin{aligned}\tau_3(x_j) &:= \mathbb{E}[e_j^3 \mid x] = \nu_{p_x}(x_j)\big(u_{p_x}(x_j) + \ell_{p_x}(x_j) - 2x_j\big),\\ \tau_4(x_j) &:= \mathbb{E}[e_j^4 \mid x] = \nu_{p_x}(x_j)\big(\Delta_j^2 - 3\nu_{p_x}(x_j)\big);\end{aligned} \tag{46}$$

let $b_j^{(3)} = \mathbb{E}[x_j\tau_3(x_j)]$ (an even integrand; $\mathbb{E}\tau_3(x_j) = 0$ by grid symmetry) and $b_j^{(4)} = \mathbb{E}\tau_4(x_j)$.

Theorem C.8 (P2, exact recursion)**.** *Under P2, conditionally on (S, θ_*): (a) $\mathbb{E}[g \mid x, e, \epsilon, v] = \hat{z}\hat{r}$, $\mathbb{E}[gg^\top \mid x, e, \epsilon, v] = (\hat{z}\hat{z}^\top + D_{p_{\mathrm{bw}}}(\hat{z}))(\hat{r}^2 + v^\top D_{p_{\mathrm{fw}}}(\hat{z})v)$, and $\mathbb{E}[g \mid v] = \mathfrak{g}(v)$; hence $\mu_{t+1} = \hat{H}_t\mu_t + \eta_t h_*$ and $\tilde{\mu}_{t+1} = \hat{H}_t\tilde{\mu}_t - \eta_t\mathsf{c}_t$. (b) With $\mathcal{B}_t^{\mathrm{P2}} = \mathbb{E}[gg^\top] = \mathcal{Q}_0 + \mathcal{Q}_{\mathrm{fw}} + \mathcal{Q}_{\mathrm{bw}} + \mathcal{Q}_{\mathrm{fwbw}}$ the expansion of the conditional second moment in (a),*

$$C_{t+1} = \hat{H}_tC_t\hat{H}_t - \eta_t\big(\hat{H}_t\tilde{\mu}_t\mathsf{c}_t^\top + \mathsf{c}_t\tilde{\mu}_t^\top\hat{H}_t\big) + \eta_t^2\mathsf{c}_t\mathsf{c}_t^\top + \frac{\eta_t^2}{B}\big(\mathcal{B}_t^{\mathrm{P2}} - \mathcal{J}_t\big) \tag{47}$$

(the mean shifts are not divided by B; the subtraction is $\mathcal{J}_t$, neither $\hat{G}C_t\hat{G}$ nor GC_tG), where the pre-sketch fourth moment is exactly

$$\mathcal{Q}_0 := \mathbb{E}[\hat{z}\hat{z}^\top\hat{r}^2] = (\mathsf{R}_t + \mathsf{F}_{x,t})\hat{G} + 2\mathcal{J}_t + S\,\mathrm{diag}_j(\Psi_j)S^\top, \tag{48}$$

$$\Psi_j = 2\alpha_j\lambda_j\big(\omega_j + \bar{U}_j + 4\mathsf{L}_j\big) + 2b_j^{(3)}\big(\omega_j + \mathsf{L}_j\big) + \big(b_j^{(4)} - 3m_j^2\big)\omega_j, \tag{49}$$

and $\mathcal{Q}_{\mathrm{fw}} = \mathbb{E}[\hat{z}\hat{z}^\top v^\top D_{p_{\mathrm{fw}}}(\hat{z})v] = \sum_k (W_t)_{kk}\mathbb{E}[\hat{z}\hat{z}^\top\nu_{p_{\mathrm{fw}}}(\hat{z}_k)]$, $\mathcal{Q}_{\mathrm{bw}} = \mathbb{E}[D_{p_{\mathrm{bw}}}(\hat{z})\hat{r}^2]$ (diagonal; a contraction of $\mathbb{E}[\hat{z}_k\hat{z}_l\nu_{p_{\mathrm{bw}}}(\hat{z}_m)]$, $\mathbb{E}[\hat{z}_k x_j\nu_{p_{\mathrm{bw}}}(\hat{z}_m)]$, $\mathbb{E}[x_jx_{j'}\nu_{p_{\mathrm{bw}}}(\hat{z}_m)]$, $\mathbb{E}\nu_{p_{\mathrm{bw}}}(\hat{z}_m)$ against W_t, $\mu_t\theta_*^\top$, $\theta_*\theta_*^\top$, σ^2) and $\mathcal{Q}_{\mathrm{fwbw}} = \mathrm{diag}_m(\sum_k (W_t)_{kk}\mathbb{E}[\nu_{p_{\mathrm{bw}}}(\hat{z}_m)\nu_{p_{\mathrm{fw}}}(\hat{z}_k)])$ are true-law contractions affine in (W_t, μ_t), so (47) *is an exact affine recursion in C_t of the same kind as Theorem C.4. (c) The hybrid evaluation (one shared $\eta \sim N(0, \bar{\Gamma})$ independent of (x, ϵ) in $\mathcal{Q}_{\mathrm{fw}}, \mathcal{Q}_{\mathrm{bw}}, \mathcal{Q}_{\mathrm{fwbw}}$ only, the same proxy in the cell argument and in the polynomial factors; $\mathcal{Q}_0$ keeps its exact form) has $\mathrm{Cov}(\hat{z}) = \hat{G}$, $\mathrm{Cov}(\hat{z}, \hat{r} \mid v) = \mathfrak{g}(v)$, $\mathrm{Var}(\hat{r} \mid v) = \sigma^2 + u^\top\Sigma_d u + v^\top\bar{\Gamma}v$, and gives the innovation*

$$\begin{aligned}\mathcal{V}_t^{A',\mathrm{P2}} = (\mathsf{R}_t + \mathsf{F}_{x,t} + \hat{\mathsf{F}}_t)\big(\hat{G} + \hat{M}_{b,t}\big) + \mathcal{J}_t + S\,\mathrm{diag}_j(\Psi_j)S^\top + 2\hat{G}\hat{A}_{a,t}\hat{G} + 2\hat{\Lambda}_{b,t}(\mathcal{J}_t)\\ + \mathrm{diag}_m\Big(\sum_k (W_t)_{kk}\,c_{ab}\big(\hat{G}_{kk}, \hat{G}_{mm}, \hat{G}_{km}/\sqrt{\hat{G}_{kk}\hat{G}_{mm}}\big)\Big),\end{aligned} \tag{50}$$

with diagonal pair correlation 1 (shared feature), to be used in place of $\mathcal{B}_t^{\mathrm{P2}} - \mathcal{J}_t$ in (47)*; it is exact for the hybrid law, preserves $K_t \succeq 0$, and reduces to* (9) *at $p_x = \infty$ ($\hat{G} = G$, $\mathsf{c} = 0$, $\mathcal{J}_t = GC_tG$, $\Psi_j = 0$).*

Proof. (a) Given (x, e, ϵ, v) only the post-sketch rounding decisions remain, independent across branches and centred with covariances $D_{p_{\mathrm{fw}}}(\hat{z})$, $D_{p_{\mathrm{bw}}}(\hat{z})$; integrating e at fixed (x, ϵ, v) gives $\mathbb{E}[g \mid x, \epsilon, v] = (zz^\top + \Gamma(x))v - zy$, then x, ϵ give $G\delta + \bar{\Gamma}v$. (b) As in Theorem C.4 with $\bar{g}_t = \mathfrak{g}(v_t) + \varsigma_t$, $\mathbb{E}[\varsigma_t\varsigma_t^\top] = \frac{1}{B}(\mathcal{B}_t^{\mathrm{P2}} - \mathcal{J}_t)$ and $\delta_{t+1} = \hat{H}_t\delta_t - \eta_t\mathsf{c}_t - \eta_t\varsigma_t$; expanding the affine shift gives every term of (47). Two-point moments (46): with $a_- = x_j - \ell_{p_x}(x_j)$, $a_+ = u_{p_x}(x_j) - x_j$, $\Delta_j = a_- + a_+$ and $\nu = a_-a_+$, $\mathbb{E}[e_j^3 \mid x] = a_+a_-(a_+^2 - a_-^2)/\Delta_j = \nu(a_+ - a_-)$ and $\mathbb{E}[e_j^4 \mid x] = a_+a_-(a_+^3 + a_-^3)/\Delta_j = \nu(a_+^2 - a_+a_- + a_-^2) = \nu(\Delta_j^2 - 3\nu)$. *Coordinate route to* (48)–(49). Fix v and put $\mathsf{a}_j = x_j + e_j$, $\mathsf{b}_j = u_jx_j + w_je_j$, so that $\hat{z} = \sum_j s_j\mathsf{a}_j$, $\hat{r} = \sum_j \mathsf{b}_j - \epsilon$; the pairs $(\mathsf{a}_j, \mathsf{b}_j)$ are independent across j and centred, with $\mathbb{E}\mathsf{a}_j^2 = \lambda_j + m_j$, $\mathbb{E}\mathsf{b}_j^2 = \lambda_ju_j^2 + m_jw_j^2$, $\mathbb{E}[\mathsf{a}_j\mathsf{b}_j] = \lambda_ju_j + m_jw_j$. In $\mathbb{E}[\hat{z}\hat{z}^\top\hat{r}^2 \mid v]$ an index occurring once gives zero, so the sum splits into the three pairings, $\hat{G}\,\mathrm{Var}(\hat{r} \mid v) + 2\,\mathfrak{g}(v)\mathfrak{g}(v)^\top$ (as $\sum_j s_js_j^\top\mathbb{E}\mathsf{a}_j^2 = \hat{G}$, $\mathrm{Var}(\hat{r} \mid v) = \sigma^2 + u^\top\Sigma_d u + v^\top\bar{\Gamma}v$, $\sum_j s_j\mathbb{E}[\mathsf{a}_j\mathsf{b}_j] = G\delta + \bar{\Gamma}v$), plus the equal-index correction $\sum_j s_js_j^\top(\mathbb{E}[\mathsf{a}_j^2\mathsf{b}_j^2] - \mathbb{E}\mathsf{a}_j^2\mathbb{E}\mathsf{b}_j^2 - 2\mathbb{E}[\mathsf{a}_j\mathsf{b}_j]^2)$. Expanding $(x_j + e_j)^2(u_jx_j + w_je_j)^2$ and conditioning on x_j (terms linear in e_j vanish; $\mathbb{E}[x_j^2e_j^2] = \lambda_j(m_j + 2\alpha_j)$, $\mathbb{E}[x_je_j^3] = b_j^{(3)}$, $\mathbb{E}e_j^4 = b_j^{(4)}$)

gives $\mathbb{E}[\mathsf{a}_j^2\mathsf{b}_j^2] = 3\lambda_j^2u_j^2 + (u_j^2 + 4u_jw_j + w_j^2)\lambda_j(m_j + 2\alpha_j) + 2(u_jw_j + w_j^2)b_j^{(3)} + w_j^2b_j^{(4)}$, so the correction is $2\alpha_j\lambda_j(u_j^2 + 4u_jw_j + w_j^2) + 2b_j^{(3)}(u_jw_j + w_j^2) + (b_j^{(4)} - 3m_j^2)w_j^2$. Averaging v ($u_j^2 \to \bar{U}_j$, $u_jw_j \to \mathsf{L}_j$, $w_j^2 \to \omega_j$, $\mathrm{Var}(\hat{r} \mid v) \to \mathsf{R}_t + \mathsf{F}_{x,t}$, $\mathfrak{g}\mathfrak{g}^\top \to \mathcal{J}_t$) proves (48)–(49); neither Gaussianity of e_j nor independence of e_j from x_j is used (independent Gaussian input noise would have $\alpha_j = b_j^{(3)} = 0$, $b_j^{(4)} = 3m_j^2$, $\Psi_j = 0$). The tensor contractions follow by expanding $\hat{r}$ and integrating ϵ. (c) Lemma C.3 with $X = \hat{z}_k$, $Y = Z = \hat{z}$ (cross covariance $\hat{G}_{\cdot k}$) gives $\mathcal{Q}_{\mathrm{fw}}^{A'} = \hat{\mathsf{F}}_t\hat{G} + 2\hat{G}\hat{A}_{a,t}\hat{G}$; with $Y = Z = \hat{r}$ it gives $\mathcal{Q}_{\mathrm{bw}}^{A'} = (\mathsf{R}_t + \mathsf{F}_{x,t})\hat{M}_{b,t} + 2\hat{\Lambda}_{b,t}(\mathcal{J}_t)$; the pair definition gives $\mathcal{Q}_{\mathrm{fwbw}}^{A'} = \hat{\mathsf{F}}_t\hat{M}_{b,t} + \mathrm{diag}_m(\sum_k (W_t)_{kk}c_{ab}(\hat{G}_{kk}, \hat{G}_{mm}, \hat{G}_{km}/\sqrt{\hat{G}_{kk}\hat{G}_{mm}}))$. Adding (48) and subtracting $\mathcal{J}_t$ once gives (50). Positivity as in Proposition C.7: $\mathcal{V}_t^{A',\mathrm{P2}} = \mathbb{E}_v\,\mathrm{Cov}(\hat{z}\hat{r} \mid v) + \mathcal{Q}_{\mathrm{fw}}^{A'} + \mathcal{Q}_{\mathrm{bw}}^{A'} + \mathcal{Q}_{\mathrm{fwbw}}^{A'} \succeq 0$, and $K_{t+1} = \hat{H}_tK_t\hat{H}_t + \frac{\eta_t^2}{B}\mathcal{V}_t^{A',\mathrm{P2}}$ because the mean-shift terms of (47) cancel exactly against $\tilde{\mu}_{t+1}\tilde{\mu}_{t+1}^\top$. □

Teacher average (P2). For a centred teacher as in Corollary C.6, $K_0 = 0$, $K_{t+1} = \hat{H}_tK_t + \eta_tS\Sigma_dC_\theta$, and (47) averages term by term with $\bar{W}_t = \bar{C}_t + K_tL^\top + LK_t^\top - V_*$, $\mathbb{E}_{\theta_*}\mathsf{L}_j = (S^\top\bar{W}_tS)_{jj} - (S^\top K_t)_{jj}$, $\bar{\mathcal{U}}_t$ as before, $\mathbb{E}_{\theta_*}[\tilde{\mu}_t\mathsf{c}_t^\top] = (K_t - LC_\theta)L^\top\bar{\Gamma}_t$, $\mathbb{E}_{\theta_*}[\mathsf{c}_t\mathsf{c}_t^\top] = \bar{\Gamma}_tV_*\bar{\Gamma}_t$, and in $\mathcal{Q}_{\mathrm{bw}}$ the replacements $(\mu_t)_k\theta_{*,j} \to (K_t)_{kj}$, $\theta_{*,j}\theta_{*,j'} \to (C_\theta)_{jj'}$.

Corollary C.9 (fixed input precision: ridge coordinates and the bias floor)**.** *Let $p_{x,t} = p_x$ for the run (post-sketch precisions and step sizes may vary) and put $v_\dagger = \hat{G}^{-1}h_* = \hat{G}^{-1}Gv_*^S$, $b_\dagger = v_\dagger - v_*^S = -\hat{G}^{-1}\bar{\Gamma}v_*^S$, $\delta_t^\dagger = v_t - v_\dagger$, $C_t^\dagger = \mathbb{E}[\delta_t^\dagger\delta_t^{\dagger\top}]$, $\tilde{\mu}_t^\dagger = \mathbb{E}\delta_t^\dagger$, $\hat{\sigma}_{\mathrm{eff}}^2 = \sigma^2 + \theta_*^\top\Sigma_d\theta_* - h_*^\top\hat{G}^{-1}h_* = \sigma_{\mathrm{eff}}^2 + v_*^{S\top}G\hat{G}^{-1}\bar{\Gamma}v_*^S \geq \sigma_{\mathrm{eff}}^2$. Then $\mathfrak{g}(v) = \hat{G}(v - v_\dagger)$, $\mathcal{J}_t = \hat{G}C_t^\dagger\hat{G}$, $\mathsf{R}_t + \mathsf{F}_{x,t} = \hat{\sigma}_{\mathrm{eff}}^2 + \mathrm{tr}(\hat{G}C_t^\dagger)$, the mean shifts disappear,*

$$\tilde{\mu}_{t+1}^\dagger = \hat{H}_t\tilde{\mu}_t^\dagger, \qquad C_{t+1}^\dagger = \hat{H}_tC_t^\dagger\hat{H}_t + \frac{\eta_t^2}{B}\big(\mathcal{B}_t^{\mathrm{P2}} - \hat{G}C_t^\dagger\hat{G}\big),$$
$$\mathbb{E}\,\mathcal{E}(v_t) = A_S + \tfrac{1}{2}\big[\mathrm{tr}(GC_t^\dagger) + 2b_\dagger^\top G\tilde{\mu}_t^\dagger + b_\dagger^\top Gb_\dagger\big], \tag{51}$$

and (50) *takes the form of the bracket of* (9) *with $(G, v_*^S, \sigma_{\mathrm{eff}}^2) \to (\hat{G}, v_\dagger, \hat{\sigma}_{\mathrm{eff}}^2)$ plus the exact term $S\,\mathrm{diag}_j(\Psi_j)S^\top$. Moreover, with $\kappa_- = 0.0814$ and $\kappa_+ = \frac{1}{4}$ from* (37)*,*

$$\Big(\frac{\kappa_-q_x}{1+\kappa_-q_x}\Big)^2\|v_*^S\|_G^2 \;\leq\; b_\dagger^\top Gb_\dagger = 2\big(\mathcal{E}(v_\dagger) - A_S\big) \;\leq\; \Big(\frac{\kappa_+q_x}{1+\kappa_+q_x}\Big)^2\|v_*^S\|_G^2 = \Big(\frac{q_x}{4+q_x}\Big)^2\|v_*^S\|_G^2, \tag{52}$$

so $\mathcal{E}(v_\dagger) - A_S \asymp 16^{-p_x}\|v_*^S\|_G^2$ whenever $v_*^S \neq 0$ (and $= 0$ when $v_*^S = 0$, e.g. $\theta_* \in \ker S\Sigma_d$). If $\sup_t\eta_t < 2/\|\hat{G}\|$ and $\sum_t\eta_t = \infty$, then $\tilde{\mu}_t^\dagger \to 0$ and, since $C_t^\dagger \succeq \tilde{\mu}_t^\dagger\tilde{\mu}_t^{\dagger\top}$, (51) *gives $\mathbb{E}\,\mathcal{E}(v_t) - A_S \geq \frac{1}{2}\|\tilde{\mu}_t^\dagger + b_\dagger\|_G^2 \to \frac{1}{2}b_\dagger^\top Gb_\dagger$: the floor the two-sided floor of Proposition C.1(ii).*

Proof. $\hat{G}v_\dagger = h_*$ gives $\mathfrak{g}(v) = \hat{G}(v - v_\dagger)$ and $\mathcal{J}_t = \hat{G}C_t^\dagger\hat{G}$. Completing the square, $\sigma^2 + u^\top\Sigma_du + v^\top\bar{\Gamma}v = \hat{\sigma}_{\mathrm{eff}}^2 + (v - v_\dagger)^\top\hat{G}(v - v_\dagger)$, which after averaging gives the second identity, and subtracting $\sigma_{\mathrm{eff}}^2 = \sigma^2 + \theta_*^\top\Sigma_d\theta_* - v_*^{S\top}Gv_*^S$ gives $\hat{\sigma}_{\mathrm{eff}}^2 - \sigma_{\mathrm{eff}}^2 = v_*^{S\top}(G - G\hat{G}^{-1}G)v_*^S = v_*^{S\top}G\hat{G}^{-1}\bar{\Gamma}v_*^S \geq 0$. The batch expansion of Theorem C.8(b) with the fixed centre $v_\dagger$ has no shift, which is (51); the readout is $\delta_t = \delta_t^\dagger + b_\dagger$. For (52) let $B_x = G^{-1/2}\bar{\Gamma}G^{-1/2}$ and $a_* = G^{1/2}v_*^S$; by congruence of (37), $\kappa_-q_xI \preceq B_x \preceq \kappa_+q_xI$, and $G^{1/2}b_\dagger = -G^{1/2}(G + \bar{\Gamma})^{-1}\bar{\Gamma}v_*^S = -(I + B_x)^{-1}B_xa_*$. In an eigenbasis of B_x each multiplier is $\beta/(1+\beta)$ with $\beta \in [\kappa_-q_x, \kappa_+q_x]$, increasing in β; squaring and summing gives (52). Finally $\tilde{\mu}_t^\dagger = \prod_{s<t}\hat{H}_s(v_0 - v_\dagger)$ and $\|\hat{H}_s\| \leq 1 - c\,\eta_s$ with $c = \min\{\lambda_{\min}(\hat{G}),\, 2/\sup_t\eta_t - \|\hat{G}\|\} > 0$. □

D THE DETERMINISTIC-EQUIVALENT DYNAMICS AND CONJECTURE 2

This appendix states the deterministic-equivalent dynamics in the form used by the proofs (§D.2) and records what is proved about the passage from the exact recursion (9) to the reduced system (10): the exact projection identities (§D.1), two obstructions that fix the form of the hypotheses (§D.3), a deterministic transfer theorem that reduces the conjecture to three finite-dimensional statements

(§D.4), a static defect of one coefficient replacement (§D.5), an explicit admissible family (§D.6) and what a proof of the conjecture requires (§D.7); §D.8 derives the law for the algorithm from Conjecture 2. Nothing in §§D.1–D.7 is used by Theorem 4, which concerns (10) alone, or by Theorem E.12.

Conventions. $r_j = 1-\eta g_j$, $\varrho = \eta/B$ ($\eta^2/B = \varrho\eta$), $\varpi = 1/a$, $\upsilon = (a-b+1)/a$, $\chi = (b-1)/a$ (so $0 < \varpi < \upsilon < 1$, $0 < \chi < 1-\varpi$), $\theta_G = \operatorname{tr} G \asymp 1$, $\eta_{\max} = c_0/c$, $\tau_* = 32(1+\eta_{\max}+c^{-1})$. The *window* is $\tau_* \le \tau_t \le M^a$ ($k_t = \tau_t^{1/a}$); it contains at least eight steps and a top-mode relaxation time in its last quarter. Unless a subscript σ or a constant $C(\sigma)$ says otherwise, constants depend only on $(a, b, c, C, c_0, \theta, \Sigma_{\max})$, the constants of Assumption 3 and the margins $\kappa_0, \gamma_*, \kappa_c$, and stay bounded as $\sigma^2 \downarrow 0$. Routine sum–integral comparisons are compressed.

Risks and penalties of the two systems. For each output rule $\mathrm{out} \in \{\mathrm{last}, \mathrm{avg}, \mathrm{cooled}\}$ the *risk of the deterministic-equivalent dynamics*, $\mathcal{E}_{\mathrm{DE}}^{\mathrm{out}}$, and the *risk of the algorithm*, $\mathcal{E}^{\mathrm{out}}$ (both target-averaged, both in excess of the approximation error A_S), are

$$\begin{gathered}\mathcal{E}_{\mathrm{DE}}^{\mathrm{last}}(t) = \tfrac{1}{2}\sum_j g_j \bar{e}_j(t), \quad \mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}}(t),\ \mathcal{E}_{\mathrm{DE}}^{\mathrm{cooled}}(t) \text{ from the two-time moments (§D.2)},\\ \mathcal{E}^{\mathrm{out}}(t) = \mathbb{E}\big[\mathcal{E}(v_{\mathrm{out}}) \mid S\big] - A_S,\end{gathered} \tag{53}$$

their exact-arithmetic versions are $\mathcal{E}_0^{\mathrm{out}}(t)$ (all rounding coefficients zero; one symbol serves both systems because without rounding the reduction is exact, Corollary B.4), and the two *rounding penalties* are

$$\begin{aligned}\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) &= \mathcal{E}_{\mathrm{DE}}^{\mathrm{out}}(t) - \mathcal{E}_0^{\mathrm{out}}(t) \ \text{(deterministic-equivalent dynamics)},\\ \Delta^{\mathrm{out}}(t) &= \mathcal{E}^{\mathrm{out}}(t) - \mathcal{E}_0^{\mathrm{out}}(t) \ \text{(algorithm)}.\end{aligned} \tag{54}$$

Note that $\mathcal{E}_{\mathrm{DE}}^{\mathrm{out}}(t)$ is not a limit of the algorithm's risk: it is the risk that the dynamics predicts at step t, and $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$ is computed by running (10) twice, with and without the rounding coefficients, and subtracting.

D.1 The projection identities (exact)

Let $G = \sum_j g_j \phi_j \phi_j^\top$ and $e_j(t) = \phi_j^\top \bar{C}_t \phi_j$. Projecting the target-averaged (9) on ϕ_j gives, exactly,

$$\begin{aligned}e_j(t+1) = (1-\eta g_j)^2 e_j(t) + \frac{\eta^2}{B}\Big[&(\bar{\mathsf{R}}_t + \mathsf{F}_t)\big(g_j + \phi_j^\top M_b \phi_j\big) + g_j^2 e_j(t)\\ &+ 2g_j^2\, \phi_j^\top A_a \phi_j + 2\,\phi_j^\top \Lambda_b(G\bar{C}_t G)\phi_j + \phi_j^\top \Delta_{ab}\phi_j\Big],\end{aligned} \tag{55}$$

with $\phi_j^\top M_b \phi_j = \sum_k \phi_j(k)^2 m_{p_{\mathrm{bw}}}(G_{kk})$, $\phi_j^\top A_a \phi_j = \sum_k \phi_j(k)^2 \frac{\alpha_{p_{\mathrm{fw}}}(G_{kk})}{G_{kk}} (\bar{W}_t)_{kk}$, $\phi_j^\top \Lambda_b(X)\phi_j = \sum_k \phi_j(k)^2 \frac{\alpha_{p_{\mathrm{bw}}}(G_{kk})}{G_{kk}} X_{kk}$ and $\phi_j^\top \Delta_{ab}\phi_j = \sum_{k,m} \phi_j(m)^2 (\bar{W}_t)_{kk} c_{ab}(G_{kk}, G_{mm}, \rho_{km})$; the off-diagonal part of $\bar{C}_t$ in the eigenbasis does not enter. The reduced system is (55) with every $\phi_j(k)^2$ replaced by $1/M$ and every product of a coordinate weight with a rounding coefficient replaced by the product of their averages. These five replacements are the content of (DE2′) and (DE3′).

D.2 The limit system, positivity and the Neumann series

Let $\pi_j(t,u) = \prod_{v=u}^{t-1}(1-\eta_v g_j)$, $\pi_j(t) = \pi_j(t,0)$, with $\eta_s \equiv \eta$ except in Section E.7. For $j \le M$, $t \ge 0$,

$$\begin{aligned}\bar{e}_j(t+1) &= r_j^2 \bar{e}_j(t) + \varrho\eta\Big[(\bar{\mathsf{R}}_t + \iota_{\mathrm{fw}}\mathsf{U}_t)(g_j + \iota_{\mathrm{bw}}) + g_j^2 \bar{e}_j(t) + \tfrac{2g_j^2 \bar{\alpha}_{\mathrm{fw}}}{M}\mathsf{U}_t + 2\bar{\alpha}_{\mathrm{bw}}\bar{\mathsf{S}}_t + \tfrac{\gamma_{\mathrm{fwbw}}}{M}\mathsf{U}_t\Big],\\ \bar{\mathsf{R}}_t &= \sigma^2 + \textstyle\sum_l g_l \bar{e}_l(t), \qquad \bar{\mathsf{S}}_t = \frac{1}{M}\sum_l g_l^2 \bar{e}_l(t),\\ \mathsf{U}_t &= \sum_l \big[w_l(1-2\pi_l(t)) + \bar{e}_l(t)\big], \qquad \bar{e}_j(0) = w_j;\end{aligned} \tag{56}$$

the exact-arithmetic system has all five coefficients zero, solution $e_j^0(t)$, moments $\mathsf{R}_t^0, \mathsf{S}_t^0, \mathsf{U}_t^0$. The outputs are $\mathcal{E}_{\mathrm{DE}}^{\mathrm{last}}(t) = \frac{1}{2}\sum_j g_j \bar{e}_j(t)$ and

$$\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}}(t) = \tfrac{1}{2t^2}\sum_j g_j y_j(t),\ \ x_j(t{+}1) = r_j x_j(t) + \bar{e}_j(t{+}1),\ \ y_j(t{+}1) = y_j(t) + 2r_j x_j(t) + \bar{e}_j(t{+}1), \tag{57}$$

$x_j(0) = y_j(0) = 0$; $\Delta_{\mathrm{DE}}^{\mathrm{out}} = \mathcal{E}_{\mathrm{DE}}^{\mathrm{out}} - \mathcal{E}_0^{\mathrm{out}}$. Put $m = \iota_{\mathrm{fw}}\iota_{\mathrm{bw}} + \gamma_{\mathrm{fwbw}}/M \asymp q_{\mathrm{fw}} q_{\mathrm{bw}} \theta^2/M^2$, $h_j = \iota_{\mathrm{fw}} g_j + 2\bar{\alpha}_{\mathrm{fw}} g_j^2/M + m$, $\mathsf{H}_{jl} = g_j g_l + g_j^2 \mathbf{1}_{j=l} + \iota_{\mathrm{bw}} g_l + 2\bar{\alpha}_{\mathrm{bw}} g_l^2/M + h_j$, $\mathsf{D} = \mathrm{diag}(r_j^2)$.

Proposition D.1 (exact difference recursion; positivity)**.** *$\bar{e}_j(t) \geq r_j^{2t} w_j$, $\mathsf{U}_t \geq 0$, $\delta_j(t) := \bar{e}_j(t) - e_j^0(t) \geq 0$, and exactly*

$$\delta(t+1) = \mathsf{D}\delta(t) + \varrho\eta\big[\Phi(t) + \mathsf{H}\delta(t)\big], \quad \delta(0) = 0, \qquad \Phi_j(t) = \iota_{\mathrm{bw}}\mathsf{R}_t^0 + 2\bar{\alpha}_{\mathrm{bw}}\mathsf{S}_t^0 + h_j\,\mathsf{U}_t^0. \tag{58}$$

Proof. If $\bar{e}_l(t) \geq r_l^{2t} w_l$ for all l then $\mathsf{U}_t = \sum_l w_l(1-r_l^t)^2 + \sum_l [\bar{e}_l(t) - r_l^{2t} w_l] \geq 0$, the bracket in (56) is ≥ 0 and $\bar{e}_j(t+1) \geq r_j^{2(t+1)} w_j$; induct. Subtracting the exact-arithmetic recursion ($\bar{\mathsf{R}}_t - \mathsf{R}_t^0 = \sum_l g_l \delta_l$, etc.) gives (58), whose coefficients are nonnegative, so $\delta \geq 0$. □

Lemma D.2 (exact response of one mode)**.** *If $e_j(t+1) = r_j^2 e_j(t) + f_j(t)$, $e_j(0) = w_j$, then $e_j(t) = r_j^{2t} w_j + \sum_{s<t} r_j^{2(t-1-s)} f_j(s)$ and*

$$y_j(t) = \textstyle\sum_{u,u'=1}^{t} r_j^{|u-u'|} e_j(\min(u,u')) = \big(\sum_{u=1}^{t} r_j^u\big)^2 w_j + \sum_{s<t}\big(\frac{1-r_j^{t-s}}{\eta g_j}\big)^2 f_j(s), \tag{59}$$

by iterating (57)*: an injection at s enters only $u, u' \geq s+1$, with weight $(\sum_{h<t-s} r_j^h)^2$ since $|u-u'| + 2(\min(u,u') - 1 - s) = (u-1-s) + (u'-1-s)$.*

Both outputs are nonnegative linear functionals of the injections. The *first insertion* F solves $F(t+1) = \mathsf{D}F(t) + \varrho\eta\Phi(t)$, $F(0) = 0$, and

$$\delta = F + \mathcal{K}\delta, \quad (\mathcal{K}v)(t) = \varrho\eta \textstyle\sum_{s<t} \mathsf{D}^{t-1-s}\mathsf{H}v(s), \qquad \text{so} \qquad \delta = \textstyle\sum_{n\geq 0} \mathcal{K}^n F \geq F, \tag{60}$$

a finite series with nonnegative terms: $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) \geq I^{\mathrm{out}}(t)$, the output of F, unconditionally.

D.3 Two obstructions

Proposition D.3 (no width-independent contraction gap)**.** *For any spectrum $g_1 \geq \cdots \geq g_M > 0$, any $\eta \geq 0$ and $B \geq 1$, the spectral radius of the homogeneous part of the exact-arithmetic recursion, $C \mapsto (I-\eta G)C(I-\eta G) + \frac{\eta^2}{B}[\mathrm{tr}(GC)\,G + GCG]$, is at least $(1-\eta g_M)^2 + \frac{2}{B}(\eta g_M)^2$. If $\eta g_1 \leq c_0$ this is $\geq 1 - 2c_0 g_M/g_1$; under the power law $g_M/g_1 \asymp M^{-a}$, no gap $\rho \leq 1 - c'$ with c' independent of M can hold, for constant steps and, per step, for any capped schedule.*

Proof. In the eigenbasis the map sends $\mathrm{diag}(e)$ to $\mathrm{diag}\big((1-\eta g_j)^2 e_j + \frac{\eta^2}{B}[g_j \sum_l g_l e_l + g_j^2 e_j]\big)$, a nonnegative matrix whose spectral radius is at least its last diagonal entry $(1-\eta g_M)^2 + \frac{2\eta^2}{B} g_M^2 \geq 1 - 2c_0 g_M/g_1$ (Perron–Frobenius); for a schedule the ordered product of the per-step maps has radius $\geq \prod_t (1-\eta_t g_M)^2$ by the same diagonal bound. □

So a Grönwall argument with a uniform contraction cannot transfer the coefficient errors to the risk: the weak modes are neither learned nor contracting on the horizon $\tau_T \asymp M^a$, and the transfer has to be organised on the scale of the penalty itself, which is what §D.4 does.

Proposition D.4 (per-mode delocalisation fails for the leading modes)**.** *If $\Phi = (\phi_1, \ldots, \phi_M)$ is Haar-distributed on the orthogonal group and independent of the eigenvalues, then for every j, $\mathbb{E}\sum_k \phi_j(k)^2 G_{kk} = (\mathrm{tr}\,G + 2g_j)/(M+2)$. The relative deviation from $\mathrm{tr}\,G/M$ is $2g_j/\mathrm{tr}\,G = O(1)$ for $j = O(1)$ and $O(j^{-a})$ for $j \to \infty$.*

Proof. $G_{kk} = \sum_l g_l \phi_l(k)^2$ and $\mathbb{E}[\phi_j(k)^2 \phi_l(k)^2] = (1 + 2\cdot\mathbf{1}_{j=l})/(M(M+2))$ (fourth moments of a uniform unit vector); sum over k. □

The same computation applies to $\sum_k \phi_j(k)^2 f(G_{kk})$ for $f = m_p, \alpha_p/\mathrm{id}$ to leading order in the fluctuation of G_{kk}, so a per-mode statement $\phi_j^\top M_b \phi_j = \iota_{\mathrm{bw}}(1 + o(1))$ is false for $j = O(1)$; this is why (DE2′) weights the errors by the response, which the leading modes carry an $O(1/k_t)$ share of for the isotropic terms.

What is known, what is measured. Three obstructions shape the hypotheses: no width-independent contraction gap exists for (9), so (G) is stated on the scale of the penalty; per-mode delocalisation fails for the leading modes, so (DE2′) is response-weighted; and the unweighted $\bar{\alpha}_{\mathrm{fw}}$ has a static defect that does not vanish with M, so (10) may use the mode-dependent coefficient $\phi_j^\top A \phi_j$ instead. Under (G) the transfer is a theorem (Theorem D.6); the probabilistic premises are open. The error is *measured* (Appendix G): for $(2, 2)$ its maximum over the regime is $13.1, 7.8, 2.5, 2.4, 0.6\%$ (last iterate) and $9.7, 5.7, 2.6, 1.5, 0.5\%$ (average) at $M = 16, \ldots, 256$.

Sufficient hypotheses for Conjecture 2. Sufficient hypotheses: power laws (DE0), Lemma B.2 (DE1), response-weighted agreement of the eigenvector-weighted sums with their coordinate averages (DE2′–DE3′), and a small feedback gain (G): $\mathcal{K}F \preceq \gamma F$, $\gamma < 1$, for the first-order response F and feedback operator $\mathcal{K}$ of Proposition D.5.

D.4 The transfer contract (proved)

Work with the target-averaged recursion at constant controls, write $D_t = \bar{C}_t - \bar{C}_t^{\mathrm{clean}}$ (also $= \bar{W}_t - \bar{W}_t^{\mathrm{clean}}$, since the means coincide), $h = \eta^2/B$, and define the maps

$$\begin{aligned} \mathcal{L}^0(X) &= HXH + h\big[GXG + \mathrm{tr}(GX)G\big], \\ \mathcal{J}(X) &= h\,\mathbb{E}_z\big[zz^\top \mathrm{tr}(D_{\mathrm{fw}}(z)X) + D_{\mathrm{bw}}(z)\, z^\top X z + D_{\mathrm{bw}}(z)\,\mathrm{tr}(D_{\mathrm{fw}}(z)X)\big], \end{aligned} \tag{61}$$

and the source $f_t = h\,\mathbb{E}_z[zz^\top \mathrm{tr}(D_{\mathrm{fw}} W_t^{\mathrm{clean}}) + D_{\mathrm{bw}}(z)(z^\top C_t^{\mathrm{clean}} z + \sigma_{\mathrm{eff}}^2) + D_{\mathrm{bw}}\,\mathrm{tr}(D_{\mathrm{fw}} W_t^{\mathrm{clean}})]$, $z \sim N(0, G)$, $D_r(z) = \mathrm{diag}_k\, \nu_{p_r}(z_k)$.

Proposition D.5 (exact perturbation equation; first-order response; sandwich)**.** *(i) $\mathcal{L}^0$ and $\mathcal{J}$ are positive maps, $f_t \succeq 0$, and $D_0 = 0$, $D_{t+1} = \mathcal{L}^0(D_t) + \mathcal{J}(D_t) + f_t$; hence $D_t \succeq 0$. (ii) Let F be the first-order response, $F_0 = 0$, $F_{t+1} = \mathcal{L}^0(F_t) + f_t$, and $\mathcal{K}$ the feedback operator on paths, $(\mathcal{K}X)_t = \sum_{s<t} (\mathcal{L}^0)^{t-1-s} \mathcal{J}(X_s)$. Then $D = \sum_{n<T} \mathcal{K}^n F \succeq F$ (componentwise Loewner order on paths). (iii) If $\Gamma_{\mathrm{ex}}(T) := \inf\{\gamma : (\mathcal{K}F)_t \preceq \gamma F_t\ \forall t \le T\} \le \gamma < 1$, then $F_t \preceq D_t \preceq F_t/(1-\gamma)$ for all $t \le T$ and, for both outputs, $I_{\mathrm{ex}}^{\mathrm{out}}(t) \le \Delta^{\mathrm{out}}(t) \le I_{\mathrm{ex}}^{\mathrm{out}}(t)/(1-\gamma)$, where $I_{\mathrm{ex}}^{\mathrm{out}}$ is the output functional of F (for the average, the readout of Corollary B.8 applied to the path F).*

Proof. (i) is the difference of (9) for the rounded and the exact-arithmetic run, with $\mathbb{E}[zz^\top(z^\top X z)] = \mathrm{tr}(GX)G + 2GXG$ and $W_t - W_t^{\mathrm{clean}} = D_t$; each term of (61) is a positive-semidefinite matrix times a nonnegative scalar. (ii) $\mathcal{K}$ advances the first nonzero time by at least one, so $\mathcal{K}^T = 0$ on paths with $X_0 = 0$, and iterating $D = F + \mathcal{K}D$ gives the finite series with nonnegative terms. (iii) $\mathcal{K}F \preceq \gamma F$ and positivity give $\mathcal{K}^n F \preceq \gamma^n F$; sum the geometric series. The readouts of both outputs are positive linear functionals of the path (for the average, because $\eta g_1 \le c_0 < 1$ makes every eigenvalue of H positive). □

Theorem D.6 (conditional transfer)**.** *Let $I_{\mathrm{r}}^{\mathrm{out}}$ and $\bar{\Delta}^{\mathrm{out}}$ be the first-order response and the penalty of the reduced system, with its own gain $\gamma_{\mathrm{r}} < 1$ (Proposition E.5). If $\Gamma_{\mathrm{ex}}(T) \le \gamma_{\mathrm{ex}} < 1$ and $\epsilon := \sup_{t \in \mathcal{T}, \mathrm{out}} |I_{\mathrm{ex}}^{\mathrm{out}}(t) - I_{\mathrm{r}}^{\mathrm{out}}(t)|/I_{\mathrm{r}}^{\mathrm{out}}(t) < 1$ on a set of terminal times $\mathcal{T}$, then uniformly on $\mathcal{T}$ and both outputs*

$$\frac{|\Delta^{\mathrm{out}}(t) - \bar{\Delta}^{\mathrm{out}}(t)|}{\Delta^{\mathrm{out}}(t)} \le \frac{1}{1-\epsilon}\Big[\epsilon + \frac{\gamma_{\mathrm{ex}}}{1-\gamma_{\mathrm{ex}}}(1+\epsilon) + \frac{\gamma_{\mathrm{r}}}{1-\gamma_{\mathrm{r}}}\Big].$$

Consequently $\Gamma_{\mathrm{ex}}(T_M) \to 0$, $\gamma_{\mathrm{r}}(T_M) \to 0$ and $\epsilon_M \to 0$ in probability over S imply Conjecture 2.

Proof. Write the difference as $(\Delta - I_{\mathrm{ex}}) + (I_{\mathrm{ex}} - I_{\mathrm{r}}) + (I_{\mathrm{r}} - \bar{\Delta})$; the two sandwiches bound the outer terms, the middle one is ϵI_{r}, and $\Delta \ge I_{\mathrm{ex}} \ge (1-\epsilon) I_{\mathrm{r}}$. □

The conjecture is thus reduced to three finite-dimensional statements about the realised feature map: the exact gain vanishes, the reduced gain vanishes (which Lemma E.6 gives under Assumption 3 on the family of §D.6), and the two first-order responses agree relatively on the regime. The last is the response-weighted form of (DE2′)–(DE3′); hypothesis (G) of the conjecture is the first.

D.5 A STATIC DEFECT OF THE FORWARD COEFFICIENT

Let $A_p = \operatorname{diag}_k \alpha_p(G_{kk})/G_{kk}$, $H_\eta = 2I - \eta(1 + B^{-1})G$, and consider the stationary forward response to a flat source, $\operatorname{tr}(G^2 H_\eta^{-1} A_p)$, against its reduced value $\bar{\alpha}_p \operatorname{tr}(G^2 H_\eta^{-1})$. For a Haar eigenbasis independent of a trace-class spectrum, along width sequences with $M/4^{\lfloor \log_4 M \rfloor} \to r$, the normalised defect $[\operatorname{tr}(G^2 H_\eta^{-1} A_p) - \bar{\alpha}_p \operatorname{tr}(G^2 H_\eta^{-1})]/q_p$ converges in probability to $\operatorname{Cov}\big(\sum_j \beta_j Z_j^2,\ a_p(\sum_j \lambda_j Z_j^2/r)\big)$, $\beta_j = \lambda_j^2/(2-\eta(1+B^{-1})\lambda_j)$, $a_p(g) = \alpha_p(g)/(q_p g)$, Z_j i.i.d. standard Gaussian; this covariance is provably nonzero for the admissible power law $\lambda_j = j^{-24}$, $d = 4M$, $B = 1$, $\eta = \frac{1}{2}$, every $p \geq 6$, and numerically nonzero ($\approx 2 \cdot 10^{-5}$ in units of q_p) for the paper's $a = 2$. The Gaussian feature map has the required ℓ^1 spectral convergence. So the unweighted replacement $\bar{\alpha}_{\mathrm{fw}}$ is not asymptotically exact even at high precision; the response-weighted coefficient $\phi_j^\top A_p \phi_j$ removes this particular defect exactly, and (10) may be solved with it at the same cost. Whether the defect is visible in the penalty of the dynamic problem is open; Appendix G measures the total error.

D.6 AN EXPLICIT ADMISSIBLE FAMILY

With $0 < \epsilon < b-1$, $\eta_M = \eta_0$ small, $B_M = \lceil M^\epsilon \rceil$, $p_{\mathrm{fw},M} = \lceil \frac{a-b}{2} \log_2 M \rceil$, $p_{\mathrm{bw},M} = \lceil \frac{b-1}{2} \log_2 M \rceil$ and $\tau_T \asymp M^a$, one has $\varrho \asymp M^{-\epsilon}$, $q_{\mathrm{fw}} \asymp M^{b-a}$, $q_{\mathrm{bw}} \asymp M^{1-b}$, and under Assumption 3 the reduced gain of Lemma E.6 is $O(M^{-\epsilon})$: the reduced sandwich holds with vanishing gain, and the forward penalty is comparable to the exact-arithmetic estimation risk $M^{1-b} + \varrho\sigma_{\mathrm{eff}}^2$. This is the family on which the conjecture should first be attacked; it also shows that at fixed positive precisions and ϱ the interaction mass $\varrho q_{\mathrm{fw}} q_{\mathrm{bw}} M^{a-1}$ of Lemma E.6 grows without bound, so the small-gain regime of Theorem 4 is not an unrestricted fixed-precision statement.

D.7 WHAT A PROOF OF CONJECTURE 2 REQUIRES

Theorem D.6 reduces the conjecture to three statements about the realised feature map; the reduced-gain statement holds on the family of §D.6 (Lemma E.6), and the other two split into four items, listed in the order in which the measurements say the errors are largest. First, the response-weighted agreement of the backward first-order responses, by leave-one-row-out or Gram-matrix delocalisation with the Haar correction of Proposition D.4 kept explicit. Second, the same agreement for the forward coordinate weights $(\bar{W}_t)_{kk}$ against $m_{p_{\mathrm{fw}}}(G_{kk})$ and $\alpha_{p_{\mathrm{fw}}}(G_{kk})/G_{kk}$, retaining the mode-dependent coefficient of §D.5 and the full, non-commuting V_*. Third, the interaction term, through the Hermite expansion of c_{ab} in ρ_{km}. Fourth, the vanishing of the exact gain Γ_{ex} on the family of §D.6. For backward rounding alone, Theorem E.12 bypasses all four.

D.8 THE LAW FOR THE ALGORITHM UNDER CONJECTURE 2

Corollary D.7 (the neural scaling law with mixed precision; full statement)**.** *Let the hypotheses of Theorem 4 hold and let Conjecture 2 hold for the feature map. Then, for both output rules, uniformly over $M^a/10 \leq \tau_t \leq M^a$ and with probability tending to one over S,*

$$\mathbb{E}[\mathcal{E}(v_{\mathrm{out}}) \mid S] - A_S = \mathcal{E}_0^{\mathrm{out}}(t) + \big(1 + o(1)\big)\Delta_{\mathrm{DE}}^{\mathrm{out}}(t); \tag{62}$$

consequently the expected excess risk of the rounded algorithm is, up to constants, $A_S + \mathcal{E}_0^{\mathrm{out}}(t) + \Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$, with $\mathcal{E}_0^{\mathrm{out}}(t)$ of Lemma E.1 and $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$ of Theorem 4.

Proof. Write $\mathcal{E}^{\mathrm{out}}(t) = \mathbb{E}[\mathcal{E}(v_{\mathrm{out}}) \mid S] - A_S$ and $\mathcal{E}^{\mathrm{out,clean}}(t) = \mathbb{E}[\mathcal{E}(v_{\mathrm{out}}^{\mathrm{clean}}) \mid S] - A_S$ for the target-averaged excess risks of the algorithm with and without rounding, so that $\Delta^{\mathrm{out}} = \mathcal{E}^{\mathrm{out}} - \mathcal{E}^{\mathrm{out,clean}}$ is the penalty of Conjecture 2. Without rounding, $M_b = A_a = \Lambda_b = \Delta_{ab} = 0$ and $\mathsf{F}_t = 0$, so the projection identity (55) reads $e_j(t+1) = (1 - \eta g_j)^2 e_j(t) + \frac{\eta^2}{B}\big[\bar{\mathsf{R}}_t g_j + g_j^2 e_j(t)\big]$, which is (10) with

all rounding coefficients zero, and the outputs are the same functionals of the e_j in both systems (Corollary B.8 and (57)). Hence $\mathcal{E}^{\mathrm{out,clean}}(t) = \mathcal{E}_0^{\mathrm{out}}(t)$ for every t and every S, and

$$\mathcal{E}^{\mathrm{out}}(t) = \mathcal{E}_0^{\mathrm{out}}(t) + \Delta^{\mathrm{out}}(t).$$

Let $\varepsilon_M = \sup_{M^a/10 \le \tau_t \le M^a} |\Delta^{\mathrm{out}}(t) - \Delta_{\mathrm{DE}}^{\mathrm{out}}(t)|/\Delta^{\mathrm{out}}(t)$, which tends to zero in probability by Conjecture 2. On the event $\{\varepsilon_M < 1\}$ every t in the window satisfies $(1-\varepsilon_M)\Delta^{\mathrm{out}}(t) \le \Delta_{\mathrm{DE}}^{\mathrm{out}}(t) \le (1+\varepsilon_M)\Delta^{\mathrm{out}}(t)$, that is, $\Delta^{\mathrm{out}}(t) = (1+\theta_t)\Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$ with $|\theta_t| \le \varepsilon_M/(1-\varepsilon_M)$ uniformly in t; this is (62). The last claim follows from the orders of $\mathcal{E}_0^{\mathrm{out}}(t)$ in Lemma E.1, the two-sided bounds of Theorem 4 for $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t)$ and $A_S \asymp M^{1-b}$. □

E Proofs of the scaling laws: Theorem 4, Proposition E.10 and Theorem E.12

E.1 Clean solution

Let $a_j^0 = r_j^2 + \varrho\eta g_j^2$ and $\beta_0 = \varrho \sum_j g_j/(2 - \eta(1+B^{-1})g_j)$.

Lemma E.1 (clean orders; σ^2-uniform residual; exact threshold)**.** *Under Assumption 3 and $\beta_0 \le 1-\kappa_0$, on the window $\mathsf{R}_t^0 \asymp r_t$, $\mathsf{U}_t^0 \asymp k_t^{a-b+1} + \varrho\sigma^2 k_t \asymp k_t^{a-b+1}$, $\mathsf{S}_t^0 \le CM^{-1}\mathsf{R}_t^0$, $\mathcal{E}_0^{\mathrm{last}}(t) \asymp k_t^{1-b} + \varrho\sigma^2$, $\mathcal{E}_0^{\mathrm{avg}}(t) \asymp k_t^{1-b} + \varrho\sigma^2 k_t/\tau_t$, the residual bound and the first energy comparison uniformly in $\sigma^2 \ge 0$; for $\eta t \le \tau_*$, $\mathsf{R}_t^0 \le C(1+\sigma^2)$ and $\mathsf{U}_t^0 \le C$. The exact-arithmetic homogeneous recursion is stable iff $\beta_0 < 1$; $\varrho\theta_G \le 1-c_0$ gives $\beta_0 \le \frac{1}{2}$. The form $\eta\theta_G/(2B) < 1$ is insufficient: for $g_j = j^{-2}$, $B = 1$, $\eta = \frac{3}{4}$ it holds (< 0.62), yet $\mathsf{R}_t^0 \ge g_1 e_1^0(t)$ gives $e_1^0(t+1) \ge [(1-\eta)^2 + 2\eta^2]e_1^0(t) = \frac{19}{16}e_1^0(t)$.*

Proof. Absorbing the diagonal term, $e_j^0(t+1) = a_j^0 e_j^0(t) + \varrho\eta g_j \mathsf{R}_t^0$, so $R_t := \mathsf{R}_t^0$ obeys the exact renewal equation $R_t = \sigma^2 + \sum_j g_j (a_j^0)^t w_j + \sum_{\ell=1}^t k_\ell R_{t-\ell}$, $k_\ell = \varrho\eta \sum_j g_j^2 (a_j^0)^{\ell-1}$, $\sum_{\ell \ge 1} k_\ell = \beta_0$. The decay exponent of a_j^0 is between fixed multiples of ηg_j, so Lemma E.3 with weight $g_j w_j \asymp j^{-b}$ gives $\sum_j g_j (a_j^0)^t w_j \le C_b(1+\eta t)^{-\chi}$ and $\mu_\chi := \sum_\ell k_\ell (\eta\ell)^\chi \le C\varrho[1 + \int_0^\infty (1+u)^{\varpi - 2 + \chi} du] < \infty$ as $\chi < 1 - \varpi$. With $\varphi_t = \sigma^2 + C_b(1+\eta t)^{-\chi}$, $\varphi_{t-h} \le (1+\eta h)^\chi \varphi_t$ and $(1 + \eta\sum_i \ell_i)^\chi \le 1 + \sum_i (\eta\ell_i)^\chi$, expanding the n-fold convolution gives the *weighted renewal bound*

$$(k^{*n} * \varphi)_t \le \varphi_t\{\beta_0^n + n\beta_0^{n-1}\mu_\chi\}, \qquad R_t \le \varphi_t\big[(1-\beta_0)^{-1} + \mu_\chi(1-\beta_0)^{-2}\big], \tag{63}$$

free of σ^2; and $R_t \ge \sigma^2 + \sum_j g_j r_j^{2t} w_j \ge c\, r_t$ by positivity. The other orders follow by inserting the exact-arithmetic injection $\varrho\eta[R_s g_j + g_j^2 e_j^0(s)] \in [\varrho\eta\sigma^2 g_j, 2\varrho\eta g_j R_s]$ into Lemma E.3(c),(d) (the decaying part with exponent $-\chi$, the noise part with exponent 0) and adding the pure bias terms $\sum_j g_j r_j^{2t} w_j \asymp k_t^{1-b}$ and $\sum_j w_j (1-r_j^t)^2 \asymp k_t^{a-b+1}$; $\mathsf{S}_t^0 \le g_1 R_t/M$. Threshold: $A_0 = \mathrm{diag}(a_j^0) + \varrho\eta g g^\top$ and $v = (I - \mathrm{diag}(a_j^0))^{-1} g > 0$ satisfy $A_0 v = v - (1-\beta_0)g$, so the v-weighted maximum norm contracts iff $\beta_0 < 1$, and $\beta_0 \ge 1$ gives $A_0^n v \ge v$. □

E.2 Spectral sums and envelopes

Lemma E.2 (cell–integral comparison)**.** *(i) $e^{-x/(1-c_0)} \le 1 - x \le e^{-x}$ on $[0, c_0]$, hence $\pi_j(t,u)^2 \in [e^{-2g_j(\tau_t - \tau_u)/(1-c_0)}, e^{-2g_j(\tau_t - \tau_u)}]$ when $\eta_v g_j \le c_0$; and $(1-e^{-1})\min(1,x) \le 1 - e^{-x} \le \min(1,x)$. (ii) On a cell of length $\ell \le \eta_{\max}$ with $g\ell \le c_0$ the factors $(1+v)^\xi$ and e^{-2gv} vary by at most $(1+\eta_{\max})^{|\xi|}$ and e^{2c_0}, so for a nonnegative product Ξ of such factors and any such partition (v_n, ℓ_n) of $[0,\tau]$, $\sum_n \ell_n \Xi(v_n) \asymp \int_0^\tau \Xi$; e.g. $\eta\sum_{s<t}(1+\eta s)^\xi \le C_\xi \int_0^{\tau_t}(1+u)^\xi du \le C'_\xi \tau_t(1+\tau_t)^\xi$ for $\xi > -1$.*

Proof. (i) For $x \in [0, c_0]$, $-x/(1-x) \le \log(1-x) \le -x$ and $x/(1-x) \le x/(1-c_0)$, which gives the first pair; the bounds on $\pi_j(t,u)^2 = \prod_v (1-\eta_v g_j)^2$ follow by summing over v. The map $x \mapsto 1 - e^{-x}$ is at most x and at most 1; it is concave, so on $[0,1]$ it lies above its chord $(1-e^{-1})x$, and for $x \ge 1$ it is at least $1 - e^{-1}$. (ii) On a cell $[v, v+\ell]$, $(1+v+\ell)/(1+v) \le 1 + \ell \le 1 + \eta_{\max}$ and $e^{-2g(v+\ell)}/e^{-2gv} = e^{-2g\ell} \ge e^{-2c_0}$, so Ξ varies by at most the product of these factors on each cell; hence $\ell_n \Xi(v_n)$ is within that factor of $\int_{\mathrm{cell}} \Xi$, and summing over the cells gives the claim. The example is the case $\Xi(u) = (1+u)^\xi$ with the cells $[\eta s, \eta(s+1)]$. □

Lemma E.3 (spectral calculus). *Let* $S_\ell(h) = \sum_j g_j^\ell r_j^{2h}$, $\mathsf{Q}_\ell(h) = \sum_j g_j^\ell (1-r_j^h)^2$, $S_\ell^{\mathrm{c}}(u) = \sum_j g_j^\ell e^{-2ug_j}$, $u = \eta h$, $J_\ell(t) = \varrho\eta \sum_{h<t} S_\ell(h) = \varrho \sum_j g_j^{\ell-1}(1-r_j^{2t})/(2-\eta g_j)$ *and* $\omega_\xi(s) = (1+\eta s)^\xi$. *(a)* Sums. *For* $u \le 1$*:* $S_\ell(h) \asymp \sum_j g_j^\ell$, $\mathsf{Q}_\ell(h) \asymp u^2 \sum_j g_j^{\ell+2}$. *For* $1 \le u \le M^a$*:* $S_\ell(h) \asymp u^{\varpi-\ell}$ *(*$\ell > \varpi$*),* $S_0(h) \asymp M$, $\mathsf{Q}_{-1}(h) \asymp u^{1+\varpi}$, $\mathsf{Q}_0(h) \asymp u^\varpi$, $\mathsf{Q}_1(h) \asymp 1$, $\sum_j (1-r_j^{2h}) \asymp u^\varpi$*; also* $\sum_j g_j^{-1} \asymp M^{a+1}$. *For* $u \ge M^a$*:* $S_\ell(h) \le e^{-2ug_M} \sum_j g_j^\ell$. *The same holds for* S_ℓ^{c}*;* $S_1^{\mathrm{c}}(u) \asymp (1+u)^{\varpi-1}$ *(*$u \le M^a$*),* $S_1^{\mathrm{c}}(u) \le C_1 M^{1-a} e^{-c_1 u/M^a}$ *(*$u \ge M^a$*). (b)* Masses. *On the window,* $J_2 \asymp \varrho\theta_G$, $J_1 \asymp \varrho k_t$, $J_0 \asymp \varrho M\tau_t$, $J_2(\infty) \le \varrho\theta_G/(2-c_0)$. *(c)* Polynomial sources. *Let* $v_s \ge 0$ *with* $v_s \le C_v(1+\eta s)^\xi$ *for all* s *and* $v_s \ge c_v(\eta s)^\xi$ *for* $\eta s \ge 1$, $\xi \ge 0$. *For injections* $f_j(s) = \varrho\eta A v_s g_j^r$ *(*$A \ge 0$*) and window* t *(*$\tau = \tau_t$, $k = k_t$*), the risk* $\sum_j g_j F_j(t)$*, coordinate sum* $\sum_j F_j(t)$ *and averaged output* $I^{\mathrm{avg}}(t)$ *of the response are* $\asymp \varrho A \tau^\xi$ *times* $(k,\, M\tau,\, k)$ *for* $r = 0$*,* $(1,\, k,\, k/\tau)$ *for* $r = 1$ *and* $(1,\, 1,\, 1/\tau)$ *for* $r = 2$. *Upper bounds use only the upper envelope; lower bounds need the lower envelope on the middle half* $t/4 \le s \le t/2$. *(d)* Power envelopes. *On the window,* $\varrho\eta \sum_{s<t} S_\ell(t-1-s)\omega_\xi(s) \le C J_\ell(t)\omega_\xi(t)$ *for* $\ell \in \{0,1\}$ *and all* $\xi > -1$*, and for* $\ell = 2$ *when* $\xi \ge 0$ *or* $-(1-\varpi) < \xi < 0$*; the averaged responses have the orders of (c) with* τ^ξ *replaced by* $\omega_\xi(t)$, $\xi > -1$. *Matching lower bounds hold when the source has one on the middle half (last quarter for* S_2*).*

Proof. (a) Reduce each filter to e^{-cug_j} or $\min(1, ug_j)$ by Lemma E.2(i); split the index sum at $k = u^\varpi$; the band $[k/2, k]$, for S_0 the last half of the indices, gives the lower bounds. (b) $1 - r_j^{2t} \asymp \min(1, \tau_t g_j)$; split at k_t. (c) Upper bounds: $v_s \le C\tau^\xi$ and (b); for the average, $I^{\mathrm{avg}}(t) = \frac{A}{2Bt^2} \sum_{s<t} v_s \mathsf{Q}_{r-1}(t-s)$ by (59), an increasing filter of maximal order $\tau^{1+\varpi}, \tau^\varpi, 1$. Lower bounds: the band $k/2 \le j \le k$ (where $g_j \asymp \tau^{-1}$) on the middle half, the band $[M/2, M]$ for the isotropic coordinate sum, and $j = 1$ over the last quarter for the top-mode entries. (d) Lemma E.2, splitting at $t/2$ for S_1; for S_2 with $\xi = -\alpha < 0$, $\omega_{-\alpha}(t-h)/\omega_{-\alpha}(t) \le (1+\eta h)^\alpha$ and $\varrho\eta \sum_h S_2(h)(1+\eta h)^\alpha \le C\varrho[1 + \int_0^\infty (1+u)^{\varpi-2+\alpha} du] < \infty$ iff $\alpha < 1 - \varpi$. □

E.3 First insertion

Split $\Phi = \Phi^{\mathrm{fw}} + \Phi^{\mathrm{bw}} + \Phi^{\mathrm{fwbw}}$, $\Phi_j^{\mathrm{fw}} = (\iota_{\mathrm{fw}} g_j + 2\bar\alpha_{\mathrm{fw}} g_j^2/M)\mathsf{U}^0$, $\Phi_j^{\mathrm{bw}} = \iota_{\mathrm{bw}}\mathsf{R}^0 + 2\bar\alpha_{\mathrm{bw}}\mathsf{S}^0$, $\Phi_j^{\mathrm{fwbw}} = m\mathsf{U}^0$, with separate responses I_r^{out}.

Proposition E.4 (the six first-insertion entries). *Under Assumption 3 and* $\beta_0 \le 1 - \kappa_0$*, on the window and with* σ^2*-free constants,* $I_{\mathrm{fw}}^{\mathrm{last}} \asymp \varrho q_{\mathrm{fw}} k_t^{a-b+1}/M$, $I_{\mathrm{fw}}^{\mathrm{avg}} \asymp \varrho q_{\mathrm{fw}} k_t^{2-b}/M$, $I_{\mathrm{bw}}^{\mathrm{last}} \asymp I_{\mathrm{bw}}^{\mathrm{avg}} \asymp \varrho q_{\mathrm{bw}} r_t k_t/M$, $I_{\mathrm{fwbw}}^{\mathrm{last}} \asymp I_{\mathrm{fwbw}}^{\mathrm{avg}} \asymp \varrho q_{\mathrm{fw}} q_{\mathrm{bw}} k_t^{a-b+2}/M^2$.

Proof. $\Phi_j^{\mathrm{fw}} \asymp (q_{\mathrm{fw}}/M)\mathsf{U}_s^0 g_j$ ($g_j^2 \le g_1 g_j$), $\iota_{\mathrm{bw}}\mathsf{R}_s^0 \le \Phi_j^{\mathrm{bw}}(s) \le (\iota_{\mathrm{bw}} + 2\bar\alpha_{\mathrm{bw}} g_1/M)\mathsf{R}_s^0 \le C\iota_{\mathrm{bw}}\mathsf{R}_s^0$ ($M\mathsf{S}_s^0 \le g_1 \mathsf{R}_s^0$), $\mathsf{U}_s^0 \asymp (\eta s)^\upsilon$ on the window with bounded onset, and $\mathsf{R}_s^0 \le \sigma^2 + C(1+\eta s)^{-\chi}$ with a matching lower bound on the middle half (Lemma E.1). The forward and mixed entries are Lemma E.3(c) with $r = 1, 0$ and $\xi = \upsilon$ ($\tau^\upsilon = k^{a-b+1}$), the backward entries Lemma E.3(d) with exponents $0, -\chi$. Two entries are written out.

Backward last iterate. With $\mathfrak{R}_{\mathrm{bw}}(s) = \mathsf{R}_s^0 + (2\bar\alpha_{\mathrm{bw}}/\iota_{\mathrm{bw}})\mathsf{S}_s^0$ and the geometric-history mean $\overline{R}_j(t) = \sum_{s<t} r_j^{2(t-1-s)} \mathfrak{R}_{\mathrm{bw}}(s) / \sum_{s<t} r_j^{2(t-1-s)}$, summing the geometric series gives exactly

$$I_{\mathrm{bw}}^{\mathrm{last}}(t) = \frac{\varrho\, \iota_{\mathrm{bw}}}{2} \sum_j \frac{1 - r_j^{2t}}{2 - \eta g_j}\, \overline{R}_j(t), \qquad \overline{R}_j(t) \asymp r_t \text{ uniformly in } j : \tag{64}$$

g_j cancels the $1/g_j$ of the stationary variance, so a saturated mode ($j \le k_t$) contributes $\asymp \varrho\iota_{\mathrm{bw}}\overline{R}_j$, an unsaturated one $\asymp \varrho\eta t g_j \iota_{\mathrm{bw}}\overline{R}_j$, and $\sum_j (1 - r_j^{2t}) \asymp k_t + \tau_t k_t^{1-a} \asymp k_t$. For the residual, $\mathfrak{R}_{\mathrm{bw}}(s) \ge \sigma^2 + \sum_l g_l w_l r_l^{2t} \ge c\, r_t$ for $s \le t$ (positivity), while $\mathfrak{R}_{\mathrm{bw}}(s) \le C[\sigma^2 + (1+\eta s)^{-\chi}]$ with geometric weights increasing in s against a decreasing envelope, so the weighted mean is at most the unweighted one (Chebyshev), $\le C[\sigma^2 + (1+\tau_t)^{-\chi}] \le C r_t$. Hence $I_{\mathrm{bw}}^{\mathrm{last}} \asymp \varrho\iota_{\mathrm{bw}} r_t k_t \asymp \varrho q_{\mathrm{bw}} \theta r_t k_t/M$. (The average weights the residual history by its own filter, so the two outputs carry different constants in front of k_t^{1-b}.)

Averaged mixed entry at $\tau_T \asymp M^a$. By (59), $I^{\mathrm{avg}}_{\mathrm{fwbw}}(t) = \frac{m}{2Bt^2}\sum_{s<t} \mathsf{U}^0_s \mathsf{Q}_{-1}(t-s)$ with $\mathsf{U}^0_s \asymp \tau_s^{\upsilon}$, $\mathsf{Q}_{-1}(h) \asymp \tau_h^{1+\varpi}$ both increasing: the upper bound is $\frac{m}{2Bt^2}\cdot Ct\tau^{\upsilon}\cdot\tau^{1+\varpi} = C\varrho m\tau^{\upsilon+\varpi}$, the lower bound retains $t/4 \le s \le t/2$, and $\tau^{\upsilon+\varpi} = k^{a-b+2}$ (no threshold at $b=2$: the limiting integral is $\tau^{\upsilon+\varpi+2}\mathrm{B}(\upsilon+1,\varpi+2)$). At $k=M$ this is $\varrho q_{\mathrm{fw}} q_{\mathrm{bw}}\theta^2 M^{a-b}$ against $\mathcal{E}_0^{\mathrm{avg}}(T) \asymp M^{1-b}$: the joint budget of Corollary F.2. The $\bar\alpha$ companions do not change these orders: the forward one is top-mode weighted like the main forward last term (an order-one change of its constant; Figure 4) and smaller by $1/k$ in the average; the backward one is pointwise $\le C$ times the main backward forcing. □

E.4 The forcing-relative gain

Put $f = \iota_{\mathrm{fw}} + 2\bar\alpha_{\mathrm{fw}} g_1/M \asymp q_{\mathrm{fw}}/M$, $\beta = \iota_{\mathrm{bw}} + 2\bar\alpha_{\mathrm{bw}} g_1/M \asymp q_{\mathrm{bw}}/M$ ($f\beta \asymp m$ when both copies are active). Since $g_j^2 v_j \le g_j R(v)$ and $g_l^2 \le g_1 g_l$,

$$\mathsf{H}v \le (2g+\beta\mathbf{1})R(v) + (fg+m\mathbf{1})U(v), \qquad R(v) = \textstyle\sum_j g_j v_j, \quad U(v) = \sum_j v_j, \qquad v \ge 0 \tag{65}$$

(the 2 is a conservative comparison, not the exact-arithmetic threshold β_0). Let $\psi_R = R(F)$, $\psi_U = U(F)$ (companions included) and, for pairs $v = (v_R, v_U) \ge 0$ and a terminal time u,

$$(\mathsf{K}v)(t) = \varrho\eta \sum_{s<t} \begin{pmatrix} 2S_2 + \beta S_1 & fS_2 + mS_1 \\ 2S_1 + \beta S_0 & fS_1 + mS_0 \end{pmatrix} (t-1-s)\, v(s), \tag{66}$$

$$\mathcal{A}_u(v) = \frac{\varrho\eta}{2u^2} \sum_{s<u}\sum_j g_j \Big(\frac{1-r_j^{u-s}}{\eta g_j}\Big)^2 \big\{(2g_j+\beta)v_R(s) + (fg_j+m)v_U(s)\big\}. \tag{67}$$

By (60), (65) and Lemma D.2, $(R(\delta), U(\delta)) \le \psi + \mathsf{K}(R(\delta), U(\delta))$ and $\Delta^{\mathrm{avg}}_{\mathrm{DE}}(u) \le I^{\mathrm{avg}}(u) + \mathcal{A}_u(R(\delta), U(\delta))$.

Proposition E.5 (forcing-relative gain; pointwise sandwich)**.** *With $0/0 := 0$, on a horizon t let*

$$\Gamma_m(t) = \max_{u\le t}\max_{i\in\{R,U\}} \frac{(\mathsf{K}\psi)_i(u)}{\psi_i(u)}, \qquad \Gamma_A(t) = \max_{u\le t}\frac{\mathcal{A}_u(\psi)}{I^{\mathrm{avg}}(u)}, \qquad \Gamma(t) = \max\{\Gamma_m(t), \Gamma_A(t)\}. \tag{68}$$

If $\Gamma(t) < 1$ then $I^{\mathrm{out}}(u) \le \Delta^{\mathrm{out}}_{\mathrm{DE}}(u) \le I^{\mathrm{out}}(u)/(1-\Gamma(t))$ for every $u \le t$ and both outputs.

Proof. Let $v = (R(\delta), U(\delta))$; by induction on u, if $v(s) \le \psi(s)/(1-\Gamma_m)$ for $s < u$ then $(\mathsf{K}v)(u) \le (\mathsf{K}\psi)(u)/(1-\Gamma_m) \le \Gamma_m\psi(u)/(1-\Gamma_m)$, so $v(u) \le \psi(u)/(1-\Gamma_m)$. Thus $\Delta^{\mathrm{last}}_{\mathrm{DE}} = R(\delta)/2 \le I^{\mathrm{last}}/(1-\Gamma_m)$ and $\Delta^{\mathrm{avg}}_{\mathrm{DE}}(u) \le I^{\mathrm{avg}}(u)[1+\Gamma_A/(1-\Gamma_m)] \le I^{\mathrm{avg}}(u)/(1-\Gamma)$. □

The sufficient condition quoted in Section 3 is

$$C_{\mathrm{sg}}\frac{\eta}{B}\Big[\theta_G + \frac{(q_{\mathrm{fw}}+q_{\mathrm{bw}})k_T}{M} + \frac{q_{\mathrm{fw}}q_{\mathrm{bw}}\tau_T}{M} + \min\{q_{\mathrm{bw}}(1+\sigma^2 k_T^{b-1}),\, q_{\mathrm{fw}}k_T^{a-1}\}\Big] \le \gamma_* \tag{69}$$

together with $\frac{\eta}{B}\theta_G \le 1 - c_0$, with an untracked constant C_{sg}; its first three terms are kernel sums, the last is the conversion of backward *energy*, accumulated isotropically on unlearned modes, into forward *risk* by one pass of the feedback loop.

Assumption 3(e) in full. Assumption 3(e) states the small-gain hypothesis in one line; its full form is the following. *Small feedback gain.* $\Gamma(T) \le \gamma_* < 1$. Here F is the first-order response of the dynamics to the rounding noise (the solution with the feedback of the rounding terms on $\bar{\mathsf{R}}_t, \mathsf{U}_t$ switched off), $\psi = (\psi_R, \psi_U) = (R(F), U(F))$ its residual and parameter-energy functionals, K the feedback map (one pass of "rounding noise raises the residual and the parameter norm, which raise the next rounding noise") and $\mathcal{A}_u$ the averaging map, and $\Gamma(T) = \max\big\{\max_{u\le T}\max_{i\in\{R,U\}}(\mathsf{K}\psi)_i(u)/\psi_i(u),\ \max_{u\le T}\mathcal{A}_u(\psi)/I^{\mathrm{avg}}(u)\big\}$, (68), is the largest factor by which one pass of the loop scales the first-order response (Proposition E.5). A sufficient condition, in the parameters of the model, is (69).

Lemma E.6 (upper bound for the gain)**.** *Assume $\beta_0 \le 1 - \kappa_0$ and for $1 \le k \le M$ let (0 if the numerator vanishes)*

$$Z_\sigma(k) = \frac{q_{\mathrm{fw}}q_{\mathrm{bw}}(\sigma^2 + k^{1-b})k^a}{q_{\mathrm{fw}}k^{a-b+1} + q_{\mathrm{bw}}(\sigma^2+k^{1-b})k + (q_{\mathrm{fw}}q_{\mathrm{bw}}/M)k^{a-b+2}}, \tag{70}$$

so that $Z_\sigma(k) \le \min\{q_{\rm bw}(1+\sigma^2 k^{b-1}), q_{\rm fw}k^{a-1}\}$. *Then for window* t, $\Gamma(t) \le C_{\rm sg}\varrho[\theta_G + (q_{\rm fw}+q_{\rm bw})k_t/M + q_{\rm fw}q_{\rm bw}\tau_t/M + \sup_{k\le k_t} Z_\sigma(k)]$*; as both bounds on* Z_σ *increase in* k, (69) *implies* $\Gamma(T) \le \gamma_*$. *The mixed energy-loop mass is two-sided,* $mJ_0(t) \asymp \varrho m M\tau_t \asymp \frac{\eta}{BM} q_{\rm fw}q_{\rm bw}\theta^2\tau_t$. $C_{\rm sg}$ *depends only on* $(a,b,c,C,c_0,\theta,\Sigma_{\max},\kappa_0)$ *and the constants of Assumption 3, not on* $(M,\eta,B,q_{\rm fw},q_{\rm bw},t,\sigma^2)$*, and is not tracked.*

Proof. Profiles, and their transfer to the exact ψ. With $V = k_u^{a-b+1}$, $X = fV$, $Y = \beta r_u + mV$, Proposition E.4 (risk and coordinate columns) gives at window times

$$\psi_R(u) \asymp \varrho(X + Yk_u), \qquad \psi_U(u) \asymp \varrho(Xk_u + YM\tau_u). \tag{71}$$

Let $\Pi_R(s) = f\omega_\upsilon + \beta[\sigma^2\omega_\varpi + \omega_{\varpi-\chi}] + m\omega_{\upsilon+\varpi}$ and $\Pi_U(s) = f\omega_{\upsilon+\varpi} + \beta M[\sigma^2\omega_1 + \omega_{1-\chi}] + mM\omega_{\upsilon+1}$, $\omega_\xi = \omega_\xi(s)$, which reproduce (71) at window times ($\tau^\upsilon = V$, $\tau^\varpi = k$, $\tau^{-\chi} = k^{1-b}$). The transfer from power envelopes to the exact first insertion, on the onset interval as well, is

$$\psi_i(s) \le C_1\varrho\,\Pi_i(s)\ \ (1 \le s \le t), \qquad \psi_i(u) \ge c_1\varrho\,\Pi_i(u)\ \ (\tau_* \le \eta u \le M^a), \qquad i \in \{R,U\}: \tag{72}$$

for $\eta s \le \tau_*$, Lemma E.1 gives $\Phi_j(v) \le \beta\mathsf{R}_v^0 + (fg_j+m)\mathsf{U}_v^0 \le C[\beta(1+\sigma^2)+fg_j+m]$, so summing at most τ_*/η injections with contraction ≤ 1, $\psi_R(s) \le C\varrho\tau_*[\beta(1+\sigma^2)\theta_G + f\sum_j g_j^2 + m\theta_G]$ and $\psi_U(s) \le C\varrho\tau_*[\beta(1+\sigma^2)M + f\theta_G + mM]$, while $\omega_\xi(s) \ge \min\{1,(1+\tau_*)^\xi\}$. Every exponent in Π_i exceeds -1, and the only possibly negative one, $\varpi - \chi = (2-b)/a$, has $|\varpi - \chi| < 1 - \varpi$ because $b < a$; so Lemma E.3(d) applies termwise, by linearity, to each ω_ξ in Π_i against each kernel of (66), and with Lemma E.3(b) and (72)

$$(\mathsf{K}\psi)(u) \le C\,\mathsf{J}(u)\,\psi(u)\ \text{ entrywise at window } u, \qquad \mathsf{J}(u) = \begin{pmatrix} 2J_2+\beta J_1 & fJ_2+mJ_1 \\ 2J_1+\beta J_0 & fJ_1+mJ_0 \end{pmatrix}(u); \tag{73}$$

no historical supremum of a decreasing backward profile is taken. With $k = k_u$, $\tau = \tau_u$, $\theta_G \asymp 1$, the two rows of (73) divided by ψ_R, ψ_U are at most $C\varrho$ times

$$\theta_G + \beta k + (f+mk)\frac{Xk + YM\tau}{X+Yk} \qquad \text{and} \qquad fk + mM\tau + (k+\beta M\tau)\frac{X+Yk}{Xk+YM\tau}.$$

Risk row. $k^2 \le M\tau$, so $k(Xk+YM\tau) \le M\tau(X+Yk)$ and the mk term is $\le mM\tau$. In the f term, $fXk/(X+Yk) \le fk$ and $f\cdot mVM\tau/(X+Yk) \le fmVM\tau/X = mM\tau$ ($X = fV$); what remains is $f\beta r_u M\tau/(X+Yk) = f\beta r_u M\tau/(fV + \beta r_u k + mVk) \asymp Z_\sigma(k)$ (substitute $f \asymp q_{\rm fw}/M$, $\beta \asymp q_{\rm bw}/M$, $m \asymp q_{\rm fw}q_{\rm bw}/M^2$, $\tau = k^a$): the backward-*energy*→forward-*risk* loop, not a kernel sum. βk and fk are the single-location orders. *Energy row.* $kX \le Xk$ and $Yk^2 \le YM\tau$ bound the sampling term by 1; $\beta M\tau X/(Xk+YM\tau) \le \beta X/Y \le \beta fV/(mV) \asymp 1$ when both copies are active (zero otherwise), and $\beta M\tau Yk/(Xk+YM\tau) \le \beta k$. *Averaged ratio.* A profile p has aligned and isotropic averaged responses $\varrho p(t)k/\tau$ and $\varrho p(t)k$ (Lemma E.3(d)); applying the four blocks of (67) to the (forward, backward, mixed) profiles $\psi_R/\varrho = (fV, \beta r_t k, mVk)$, $\psi_U/\varrho = (fVk, \beta r_t M\tau, mVM\tau)$ and comparing with $I^{\rm avg}/\varrho = (fVk/\tau, \beta r_t k, mVk)$ gives the twelve relative factors $\varrho, \varrho k/\tau, \varrho k/\tau$ (sampling); $\varrho, \varrho\beta k, \varrho\beta k$ (backward, by $\beta f \asymp m$); $\varrho fk, \varrho fM, \varrho fM$ (forward); $\varrho fk, \varrho mM\tau, \varrho mM\tau$ (mixed): $\Gamma_A \le C\varrho[\theta_G + (q_{\rm fw}+q_{\rm bw})k/M + mM\tau]$. *Small terminal times* $\eta u \le \tau_*$: positivity gives $\psi_i(s) \le c_*^{-1}\psi_i(u)$ for $s \le u$, $c_* = e^{-2C\tau_*/(1-c_0)}$, the shapes $1, g_j, g_j^2$ of Φ give $\psi_U(u) \le C_*(M/\theta_G)\psi_R(u)$, $\psi_R(u) \le g_1\psi_U(u)$, and with $S_2 \le \sum_j g_j^2$, $S_1 \le \theta_G$, $S_0 \le M$, $fM \lesssim q_{\rm fw} \le 1$, $\beta M \lesssim q_{\rm bw} \le 1$ this yields $(\mathsf{K}\psi)_i(u) \le C\varrho\psi_i(u)$ and, interchanging the two finite sums in (67) with $c_*(u-s) \le \sum_{h<u-s} r_j^h \le u-s$, $\mathcal{A}_u(\psi) \le C\varrho\tau_* I^{\rm avg}(u)$. Finally, dropping denominator terms in (70) gives its two bounds; the mixed mass is Lemma E.3(b); $\beta_0 \le \frac{1}{2}$ is Lemma E.1. □

Beyond the small-gain regime. The certificate (69) contains the term $\mathsf{G} := \varrho q_{\rm fw}q_{\rm bw}\tau_T/M$; at fixed positive bits and noise its last term $\varrho\min\{q_{\rm bw}(1+\sigma^2k_T^{b-1}), q_{\rm fw}k_T^{a-1}\}$ grows with k_T whatever M is, so the certificate can hold only on a finite range of horizons. The next proposition describes the dynamics without it. Two mechanisms appear. Downward: the isotropic interaction noise raises the mean squared parameter norm U_t, which raises the interaction noise; the unlearned modes contract only by $1 - 2\eta g_j \ge 1 - C\eta M^{-a}$ per step, so once $\varrho q_{\rm fw}q_{\rm bw}/M$ exceeds a multiple of M^{-a} their excess energy grows geometrically, by a factor $e^{c\mathsf{G}}$ over the run. Upward: an exponentially weighted

resolvent bound shows that nothing in the dynamics grows faster than $e^{L\mathsf{G}}$ for a fixed constant L. Together they fix the compute frontier of the average with both operands rounded, to logarithmic accuracy, without any small-gain hypothesis. Write $\mathsf{D}_\times = \varrho q_{\rm fw} q_{\rm bw}$, $\mathsf{D}_{\rm bw} = \varrho q_{\rm bw}\sigma^2$ (sans-serif, to keep them apart from the rounding matrices $D_{\rm bw}(z)$ of Appendix B), $\theta = \operatorname{tr}\Sigma_d$, $x_j(t) = \bar e_j(t) - \bar e_j^0(t)$ for the excess energy of mode j over the exact-arithmetic run, $J = \{\lceil M/2\rceil, \dots, M\}$, $X_t = \sum_{j\in J} x_j(t)$, and U_t, $\bar{\mathsf{R}}_t$ for the parameter norm and residual of (56).

Proposition E.7 (the unlearned modes at fixed precision: amplification, growth bound, frontier)**.** *Let both copies be rounded at fixed precisions $q_{\rm fw}, q_{\rm bw} \in (0,1]$ with exact input, constant controls, Assumption 3(a)–(d), and horizons on the window $\tau_* \le \tau_T \le M^a$, $k_T = \tau_T^{1/a}$. Constants $c, C, c', L_0, \dots$ depend only on $(a, b, c, C, c_0, \theta, \Sigma_{\max}, \kappa_0)$ and, where indicated, on L; none depends on M, T, $q_{\rm fw}$, $q_{\rm bw}$, ϱ or σ^2 unless displayed.*

(i) (recurrence) for every $t \ge 0$,

$$X_{t+1} \;\ge\; \Big(1 - C\eta M^{-a} + c\,\eta\,\frac{\mathsf{D}_\times}{M}\Big) X_t + c\,\eta\,\mathsf{D}_{\rm bw}; \tag{74}$$

(ii) (amplification) let $M_{\rm amp} := (2C/c)^{1/(a-1)}\mathsf{D}_\times^{-1/(a-1)}$. For $M \ge M_{\rm amp}$ the multiplier in (74) is $\mu = 1 + z$ with $\frac{c}{2}\eta\mathsf{D}_\times/M \le z \le 1$, and

$$\begin{gathered} X_T \;\ge\; \frac{2\sigma^2 M}{q_{\rm fw}}\big(e^{c'\mathsf{G}} - 1\big), \\ \Delta_{\rm DE}^{\rm last}(T) \;\ge\; c\,\frac{\sigma^2 M^{1-a}}{q_{\rm fw}}\big(e^{c'\mathsf{G}} - 1\big), \qquad \Delta_{\rm DE}^{\rm avg}(T) \;\ge\; c\,\frac{\sigma^2 M^{1-a}}{q_{\rm fw}}\big(e^{c'\mathsf{G}/2} - 1\big), \end{gathered} \tag{75}$$

with $c' = c/4$;

(iii) (growth bound) there are L_0 and, for each $L \ge L_0$, $M_0(L)$ and C_L such that for $M \ge M_0(L)$, with $\lambda := L\mathsf{D}_\times/M$ and for every t (any horizon),

$$\bar{\mathsf{R}}_t \le C_L e^{\lambda\tau_t}, \qquad \frac{q_{\rm fw}\mathsf{U}_t}{M} \le C_L e^{\lambda\tau_t}; \tag{76}$$

(iv) (averaged risk) under the hypotheses of (iii), on the window,

$$\mathcal{E}_{\rm DE}^{\rm avg}(T) \;\le\; C\,k_T^{1-b} + C_L\,e^{L\mathsf{G}}\,\tau_T^{-(a-1)/a}\,\mathsf{G}^{-1-1/a}; \tag{77}$$

(v) (frontier) fix the bits, ϱ and σ^2, and let $f = \bar\kappa\,k_T^a M$ with $\bar\kappa$ fixed. With $x := f/(\bar\kappa\mathsf{D}_\times)$ and $r := (b-1)/(2a)$, for all f large enough,

$$c\,\big(x\log x\big)^{-r} \;\le\; \inf_{\bar\kappa k_T^a M = f,\; k_T \le M} \mathcal{E}_{\rm DE}^{\rm avg}(T) \;\le\; C\,\big(x\log x\big)^{-r}, \tag{78}$$

the upper bound attained by the policy $\mathsf{G} = \varepsilon\log x$ with any fixed $0 < \varepsilon < \frac{a-b}{2aL}$, i.e. $M = \mathsf{D}_\times\tau_T/(\varepsilon\log x) \asymp \mathsf{D}_\times k_T^a/\log f$.

Proof. Throughout, $\eta^2/B = \eta\varrho$, and (56) is written as $\bar e_j(t+1) = d_j\bar e_j(t) + \eta\varrho\,\Phi_j(t)$ with $d_j := (1-\eta g_j)^2 + \eta\varrho g_j^2$ and

$$\Phi_j(t) = (\bar{\mathsf{R}}_t + \iota_{\rm fw}\mathsf{U}_t)(g_j + \iota_{\rm bw}) + \frac{2g_j^2\bar\alpha_{\rm fw}}{M}\mathsf{U}_t + 2\bar\alpha_{\rm bw}\bar{\mathsf{S}}_t + \frac{\gamma_{\rm fwbw}}{M}\mathsf{U}_t \;\ge 0,$$

all coefficients being nonnegative. Note $1 - d_j = \eta g_j[2 - \eta(1+B^{-1})g_j] \ge c\,\eta g_j$ by Assumption 3(b) ($\eta g_1 \le c_0 < 1$, $B \ge 1$).

(i) Subtracting the exact-arithmetic recursion ($\iota_r = \bar\alpha_r = \gamma_{\rm fwbw} = 0$) and using Proposition D.1 ($x_j \ge 0$, $\bar{\mathsf{R}}_t - \mathsf{R}_t^0 = \sum_l g_l x_l \ge 0$, $\mathsf{U}_t - \mathsf{U}_t^0 = \sum_l x_l \ge 0$) we may drop the nonnegative terms $(\bar{\mathsf{R}}_t - \mathsf{R}_t^0)g_j$, $g_j^2 x_j$, $\iota_{\rm fw}\mathsf{U}_t g_j$, $2g_j^2\bar\alpha_{\rm fw}\mathsf{U}_t/M$, $2\bar\alpha_{\rm bw}\bar{\mathsf{S}}_t$ to get

$$x_j(t+1) \;\ge\; (1-\eta g_j)^2 x_j(t) + \eta\varrho\Big[\sigma^2\iota_{\rm bw} + \Big(\iota_{\rm fw}\iota_{\rm bw} + \frac{\gamma_{\rm fwbw}}{M}\Big)X_t\Big],$$

using $\bar{\mathsf{R}}_t \ge \sigma^2$ and $\mathsf{U}_t \ge \mathsf{U}_t^0 + X_t \ge X_t$ ($\mathsf{U}_t^0 \ge 0$ is a mean squared norm). Sum over $j \in J$ ($|J| \ge M/2$): $(1-\eta g_j)^2 \ge 1 - 2\eta g_{\lceil M/2 \rceil} \ge 1 - C\eta M^{-a}$ by Assumption 3(a), and by (c) $\frac{M}{2}\,\iota_{\mathrm{fw}}\iota_{\mathrm{bw}} \ge c\,q_{\mathrm{fw}}q_{\mathrm{bw}}\theta^2/M$ and $\frac{M}{2}\,\sigma^2\iota_{\mathrm{bw}} \ge c\,q_{\mathrm{bw}}\theta\sigma^2$, while $\gamma_{\mathrm{fwbw}} \ge 0$. This is (74); no lower bound on γ_{fwbw} is used.

(ii) For $M \ge M_{\mathrm{amp}}$, $C\eta M^{-a} \le \frac{c}{2}\eta\mathsf{D}_\times/M$, so $z := \mu - 1 \ge \frac{c}{2}\eta\mathsf{D}_\times/M$; and $z \le c\eta\mathsf{D}_\times/M \le c\,\eta\varrho \le 1$ by (b). Since $X_0 = 0$ and $\sum_{s<T}\mu^s$ is increasing in μ, (74) gives, with $z_- := \frac{c}{2}\eta\mathsf{D}_\times/M$,

$$X_T \;\ge\; c\,\eta\mathsf{D}_{\mathrm{bw}}\sum_{s<T}(1+z_-)^s = c\,\eta\mathsf{D}_{\mathrm{bw}}\,\frac{(1+z_-)^T-1}{z_-} \;\ge\; \frac{c\,\eta\mathsf{D}_{\mathrm{bw}}}{z_-}\big(e^{z_-T/2}-1\big)$$
$$= \frac{2\sigma^2 M}{q_{\mathrm{fw}}}\big(e^{c\mathsf{G}/4}-1\big),$$

because $\log(1+z) \ge z/2$ for $0 < z \le 1$, $z_-T/2 = \frac{c}{4}\mathsf{D}_\times\eta T/M = \frac{c}{4}\mathsf{G}$ and $\eta\mathsf{D}_{\mathrm{bw}}/z_- = 2\mathsf{D}_{\mathrm{bw}}M/(c\mathsf{D}_\times) = 2\sigma^2 M/(c\,q_{\mathrm{fw}})$. The last-iterate penalty is $\Delta_{\mathrm{DE}}^{\mathrm{last}}(T) = \frac{1}{2}\sum_j g_j x_j(T) \ge \frac{1}{2}g_M X_T$ with $g_M \ge cM^{-a}$. For the average, by (59) the averaged excess is $\frac{1}{2T^2}\sum_j g_j\sum_{s<T}\big(\frac{1-r_j^{T-s}}{\eta g_j}\big)^2 N_j(s)$ with the nonnegative one-step injections $N_j(s) = \bar{e}_j(s+1) - r_j^2\bar{e}_j(s)$, and the same identity for the exact-arithmetic run; the excess $\Delta_{\mathrm{DE}}^{\mathrm{avg}}(T)$ is therefore $\frac{1}{2T^2}\sum_j g_j\sum_s(\cdot)^2[N_j(s)-N_j^0(s)]$, and $N_j - N_j^0 \ge 0$ termwise by positivity. Keep only $j \in J$ and $s \le T/2$, and drop all but the part of $N_j(s) - N_j^0(s)$ that builds x_j: since $x_j(T/2) = \sum_{s<T/2} r_j^{2(T/2-1-s)}[N_j(s)-N_j^0(s)]$ and $r_j^{2(T/2-1-s)} \le 1$, we get $\sum_{s<T/2}[N_j(s)-N_j^0(s)] \ge x_j(\lfloor T/2\rfloor)$. On the window, for $j \in J$ and $s \le T/2$, $\eta g_j(T-s) \le g_j\tau_T \le C(M/2)^{-a}M^a = C'$, so $\frac{1-r_j^{T-s}}{\eta g_j} \ge (T-s)\frac{1-e^{-C'}}{C'} \ge c\,T$ (using $1-(1-u)^h \ge 1-e^{-uh} \ge \frac{1-e^{-C'}}{C'}uh$ for $uh \le C'$). Hence $\Delta_{\mathrm{DE}}^{\mathrm{avg}}(T) \ge \frac{c^2}{2}\sum_{j\in J} g_j x_j(\lfloor T/2\rfloor) \ge c\,M^{-a}X_{\lfloor T/2\rfloor}$, and (i)–(ii) run to $\lfloor T/2\rfloor$ give the last bound of (75).

(iii) Unrolling, $\bar{e}_j(t) = d_j^t w_j + \eta\varrho\sum_{s<t} d_j^{t-1-s}\Phi_j(s)$. Put $Z_t := q_{\mathrm{fw}}\mathsf{U}_t/M$ and, for a horizon T, $R_\lambda := \max_{t\le T} e^{-\lambda\tau_t}\bar{\mathsf{R}}_t$, $Z_\lambda := \max_{t\le T} e^{-\lambda\tau_t}Z_t$. By (c), $\iota_{\mathrm{fw}}\mathsf{U}_t \le C\theta Z_t$, $\bar{\alpha}_{\mathrm{fw}}\mathsf{U}_t/M \le CZ_t$, $\gamma_{\mathrm{fwbw}}\mathsf{U}_t/M \le Cq_{\mathrm{bw}}\theta^2 Z_t/M$, $\iota_{\mathrm{bw}} \le Cq_{\mathrm{bw}}\theta/M$, $\bar{\mathsf{S}}_t \le g_1\bar{\mathsf{R}}_t/M$, so

$$\Phi_j(s) \;\le\; g_j\bar{\mathsf{R}}_s + C\Big(g_j + \frac{q_{\mathrm{bw}}}{M}\Big)\big(\bar{\mathsf{R}}_s + Z_s\big) \quad\text{and}\quad \Phi_j(s) \le C\Big(g_j + \frac{q_{\mathrm{bw}}}{M}\Big)(\bar{\mathsf{R}}_s + Z_s),$$

where the first form keeps the exact-arithmetic residual feedback $g_j\bar{\mathsf{R}}_s$ separate (its constant is 1) and the second absorbs it. Three sums are used: for $\lambda > 0$, $\sum_{s<t} d_j^{t-1-s}e^{\lambda\tau_s} \le e^{\lambda\tau_t}\,\eta/(1-d_je^{-\eta\lambda})$, and $1-d_je^{-\eta\lambda} \ge 1-d_j$, $1-d_je^{-\eta\lambda} \ge e^{-\eta\lambda}\eta(\lambda+cg_j)$; and by Assumption 3(a), $\sum_j g_j/(\lambda+cg_j) \le C\lambda^{-1/a}$, $\sum_j 1/(\lambda+cg_j) \le M/\lambda$. *Residual.* $\bar{\mathsf{R}}_t = \sigma^2 + \sum_j g_j\bar{e}_j(t)$; the first term of Φ_j contributes $\eta\varrho\sum_j g_j^2\sum_s d_j^{t-1-s}\bar{\mathsf{R}}_s \le e^{\lambda\tau_t}R_\lambda\,\varrho\sum_j g_j^2\eta/(1-d_j) = e^{\lambda\tau_t}R_\lambda\,\varrho\sum_j\frac{g_j}{2-\eta(1+B^{-1})g_j} = \beta_0\,e^{\lambda\tau_t}R_\lambda$, with $\beta_0 \le 1-\kappa_0$ by (d); the remaining terms of $\Phi_j(s)$ are, by (c), at most $C\big[(g_j + q_{\mathrm{bw}}/M)\,\frac{q_{\mathrm{bw}}\theta}{M} + \frac{g_1 q_{\mathrm{bw}}}{M}\big]\bar{\mathsf{R}}_s$ (backward and $\bar{\alpha}_{\mathrm{bw}}$ couplings) plus $C\big[\theta g_j + g_j^2 + \frac{q_{\mathrm{bw}}\theta^2}{M}\big]Z_s$ (forward, $\bar{\alpha}_{\mathrm{fw}}$ and mixed couplings), and after the weighted sums they contribute $e^{\lambda\tau_t}\big[\epsilon_M R_\lambda + C_1 Z_\lambda\big]$ with $\epsilon_M := C\varrho q_{\mathrm{bw}}e^{\eta\lambda}\big[\theta M^{-1}\lambda^{-1/a} + g_1M^{-1}\lambda^{-1/a}\big] = C\varrho q_{\mathrm{bw}}M^{1/a-1}(L\mathsf{D}_\times)^{-1/a}(1+o(1)) \to 0$ as $M\to\infty$ (using $\sum_j g_j/(\lambda+cg_j) \le C\lambda^{-1/a}$ and $\sum_j g_j\cdot q_{\mathrm{bw}}/M\cdot 1/(\lambda+cg_j) \le q_{\mathrm{bw}}\lambda^{-1/a}C/M$) and $C_1 := C\varrho e^{\eta\lambda}\big[(\theta+g_1)\sum_j g_j/c + q_{\mathrm{bw}}\theta^2M^{-1}\lambda^{-1/a}\big] \le C\varrho\theta(\theta+g_1)/c + o(1)$, a constant. So $R_\lambda \le \sigma^2 + \sum_j g_j w_j + (\beta_0+\epsilon_M)R_\lambda + C_1Z_\lambda$. *Parameter norm.* $\mathsf{U}_t = \sum_j w_j(1-r_j^t)^2 + \sum_j[\bar{e}_j(t) - w_jr_j^{2t}]$; the first sum is $\le C(1+\tau_t)^u$, $u = (a-b+1)/a < 1$ (Lemma E.1), so its weighted contribution to Z_λ is $\le Cq_{\mathrm{fw}}M^{-1}\lambda^{-u} = Cq_{\mathrm{fw}}M^{u-1}(L\mathsf{D}_\times)^{-u} \to 0$; the exact-arithmetic sampling part $\sum_j(d_j^t - r_j^{2t})w_j$ is bounded by the exact-arithmetic parameter norm, $\le C(1+\tau_t)^u$ (Lemma E.1), and is absorbed in the same $o(1)$; the Φ part contributes $\frac{q_{\mathrm{fw}}}{M}\eta\varrho\sum_j\sum_s d_j^{t-1-s}\Phi_j(s) \le e^{\lambda\tau_t}(R_\lambda + Z_\lambda)\,\frac{q_{\mathrm{fw}}}{M}C\varrho e^{\eta\lambda}\big[\sum_j\frac{g_j}{\lambda+cg_j} + \frac{q_{\mathrm{bw}}}{M}\sum_j\frac{1}{\lambda+cg_j}\big] \le e^{\lambda\tau_t}(R_\lambda+Z_\lambda)\big[C\varrho q_{\mathrm{fw}}M^{-1}\lambda^{-1/a} + C\frac{\varrho q_{\mathrm{fw}}q_{\mathrm{bw}}}{M\lambda}\big]$, and $\varrho q_{\mathrm{fw}}q_{\mathrm{bw}}/(M\lambda) = \mathsf{D}_\times/(M\lambda) = 1/L$. Hence

$$\begin{pmatrix}R_\lambda\\ Z_\lambda\end{pmatrix} \le \begin{pmatrix}C\\ o(1)\end{pmatrix} + \begin{pmatrix}\beta_0+\epsilon_M & C_1\\ C_2/L+\epsilon'_M & C_2/L+\epsilon'_M\end{pmatrix}\begin{pmatrix}R_\lambda\\ Z_\lambda\end{pmatrix},$$

a nonnegative 2×2 system whose spectral radius is < 1 once $L \geq L_0 := 4C_1C_2/\kappa_0 + 4C_2$ and $M \geq M_0(L)$ makes $\epsilon_M, \epsilon'_M \leq \kappa_0/4$: then $(1-\beta_0-\epsilon_M)(1-C_2/L-\epsilon'_M) \geq \frac{\kappa_0}{2}\cdot\frac{1}{2} > C_1(C_2/L+\epsilon'_M)$. So $R_\lambda, Z_\lambda \leq C_L$ with C_L independent of the horizon, which is (76). Only Assumption 3(a)–(d) is used; not (e).

(iv) The one-step injections satisfy $N_j(t) = \eta\varrho[g_j^2\bar{e}_j(t) + \Phi_j(t)] \leq C\eta\varrho(g_j + q_{\mathrm{bw}}/M)e^{\lambda\tau_t}$ by (76) and $g_j^2\bar{e}_j \leq g_j\bar{\mathsf{R}}$. By (59), $\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}}(T) = \frac{1}{2T^2}\sum_j g_j\big(\sum_{u\leq T} r_j^u\big)^2 w_j + \frac{1}{2T^2}\sum_j g_j \sum_{s<T}\big(\frac{1-r_j^{T-s}}{\eta g_j}\big)^2 N_j(s)$; the first term is the exact-arithmetic averaged bias, $\leq Ck_T^{1-b}$ on the window (Lemma E.1); in the second, $\sum_j g_j(g_j + q_{\mathrm{bw}}/M)\big(\frac{1-r_j^h}{\eta g_j}\big)^2 = \eta^{-2}[\mathsf{Q}_0(h) + \frac{q_{\mathrm{bw}}}{M}\mathsf{Q}_{-1}(h)]$ with $\mathsf{Q}_0(h) \leq C(\eta h)^{1/a}$, $\mathsf{Q}_{-1}(h) \leq C(\eta h)^{1+1/a}$ on the window (Lemma E.3(a)), and $\eta\sum_{h\geq 1} e^{-\lambda\eta h}(\eta h)^p \leq C_p\lambda^{-p-1}$. Hence the second term is at most $\frac{C\varrho}{\tau_T^2}e^{\lambda\tau_T}\big[\lambda^{-1-1/a} + \frac{q_{\mathrm{bw}}}{M}\lambda^{-2-1/a}\big]$ with $\tau_T = \eta T$. Since $\lambda\tau_T = L\mathsf{G}$ and $\frac{q_{\mathrm{bw}}}{M\lambda} = \frac{1}{L\varrho q_{\mathrm{fw}}}$, this equals $C_L\, e^{L\mathsf{G}}\,\tau_T^{-2}(\tau_T/\mathsf{G})^{1+1/a} = C_L\, e^{L\mathsf{G}}\tau_T^{-(a-1)/a}\mathsf{G}^{-1-1/a}$, with C_L depending on L, ϱ, q_{fw} (fixed).

(v) Write $F := f/\bar{\kappa} = k_T^a M = \tau_T M$. *Lower bound*, three cases for a policy (k_T, M), $k_T \leq M$. If $M < M_{\mathrm{amp}}$: $\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}} \geq \mathcal{E}_0^{\mathrm{avg}} \geq ck_T^{1-b} \geq cM^{1-b} \geq cM_{\mathrm{amp}}^{1-b}$, a constant, which exceeds $(x\log x)^{-r}$ for large x. If $M \geq M_{\mathrm{amp}}$ and $\mathsf{G} \leq C_0$: $M \geq \mathsf{D}_\times k_T^a/C_0$, so $F \geq \mathsf{D}_\times k_T^{2a}/C_0$, $k_T \leq (C_0x)^{1/(2a)}$ and $\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}} \geq ck_T^{1-b} \geq c(C_0x)^{-r} \geq c(x\log x)^{-r}$. If $M \geq M_{\mathrm{amp}}$ and $\mathsf{G} > C_0$: $M = \mathsf{D}_\times k_T^a/\mathsf{G}$, $F = \mathsf{D}_\times k_T^{2a}/\mathsf{G}$, so $k_T^{2a} = x\mathsf{G}$ and $M^{1-a} = (\mathsf{D}_\times x/\mathsf{G})^{-(a-1)/2}$; by (ii),

$$\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}} \;\geq\; c\,(x\mathsf{G})^{-r} + c\,\frac{\sigma^2}{q_{\mathrm{fw}}}\mathsf{D}_\times^{-\frac{a-1}{2}}\,x^{-\frac{a-1}{2}}\,\mathsf{G}^{\frac{a-1}{2}}\big(e^{c'\mathsf{G}/2}-1\big).$$

As $\frac{a-1}{2} > r$ (equivalent to $a(a-1) > b-1$, true for $1 < b < a$), the second term exceeds $(x\log x)^{-r}$ unless $\mathsf{G} \leq C_1\log x$, in which case the first term is $\geq c(C_1x\log x)^{-r}$; either way $\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}} \geq c(x\log x)^{-r}$ for x large. *Upper bound.* Take $\mathsf{G} = \varepsilon\log x$ and $M = \mathsf{D}_\times\tau_T/\mathsf{G}$, so $F = \mathsf{D}_\times\tau_T^2/\mathsf{G}$, $\tau_T^2 = x\mathsf{G}$, and $M \geq k_T$ iff $\mathsf{D}_\times\tau_T^{1-1/a} \geq \varepsilon\log x$, true for x large; the run is on the window. By (iv), the noise term over the bias $k_T^{1-b} = \tau_T^{-(b-1)/a}$ is at most $C_Le^{L\mathsf{G}}\tau_T^{-(a-b)/a}\mathsf{G}^{-1-1/a} = C_L\,x^{L\varepsilon}(x\mathsf{G})^{-\frac{a-b}{2a}}\mathsf{G}^{-1-1/a} \to 0$ because $L\varepsilon < \frac{a-b}{2a}$. Hence $\mathcal{E}_{\mathrm{DE}}^{\mathrm{avg}}(T) \leq Ck_T^{1-b} = C(x\varepsilon\log x)^{-r}$. □

Remark E.8 (the widening policy, the accumulated amplification and the backward-only case). (a) *What the widening policy does.* Along $M \asymp \mathsf{D}_\times k_T^a$ the ratio of weak-mode feedback to weak-mode contraction is $\mathsf{D}_\times M^{a-1} \to \infty$: the unlearned modes *are* amplifying ($\mu > 1$ for $M > M_{\mathrm{amp}}$). The policy does not keep the feedback below the contraction; it keeps the amplification accumulated over the run, $e^{c'\mathsf{G}}$ to $e^{L\mathsf{G}}$, bounded. The log-optimal policy of (v) even lets G grow like $\varepsilon\log f$, trading a polynomial factor $f^{L\varepsilon}$ of amplification against the prefactor M^{1-a} of the weak modes' risk. (b) *Two regimes.* For $\mathsf{G} \lesssim 1$ the loss is within constants of the first-order response (Proposition E.5 where its certificate holds), whose own frontier exponent is $(b-1)/(a+\max\{b, \frac{a+1}{2}\})$ with the policy $M \asymp \max\{k_T, \mathsf{D}_{\mathrm{bw}}k_T^b, \mathsf{D}_\times^{1/2}k_T^{(a+1)/2}\}$; along that policy $\mathsf{G} \asymp \min\{(\mathsf{D}_\times/\mathsf{D}_{\mathrm{bw}})k_T^{a-b}, \mathsf{D}_\times^{1/2}k_T^{(a-1)/2}\}$ reaches order one at $k_\times \asymp \max\{(\mathsf{D}_{\mathrm{bw}}/\mathsf{D}_\times)^{1/(a-b)}, \mathsf{D}_\times^{-1/(a-1)}\}$, beyond which (v) takes over. (c) *Numbers.* For the realised Gaussian feature map at $(a,b) = (2, \frac{3}{2})$, $\gamma_{\mathrm{fwbw}}M \approx 0.015$; the weak-mode multiplier of (i) changes sign near $M \approx 6\cdot 10^4$ at $q_{\mathrm{fw}} = q_{\mathrm{bw}} = 1$, $B = 32$, $\eta = 0.5/g_1$ ($M_{\mathrm{amp}} \propto (B/\eta)^{1/(a-1)}$ at fixed η), so every width of the experiments has $\mu < 1$; with the mixed coefficient rescaled so that the feedback is $1.5\times$ the contraction at $M = 256$, the measured growth rate of X_t on the second half of a run to $\tau_T = 2M^a$ is $9.8\cdot 10^{-5}$ per step against the bound's $\log\mu \approx 3.0\cdot 10^{-5}$, and the penalty reaches $4.7\cdot 10^8$ against a exact-arithmetic excess risk of $2\cdot 10^{-3}$; at the unscaled coefficient the growth on the same window is algebraic — an observation consistent with (iii), not a consequence of (i). This was checked numerically. (d) *Backward rounding only* ($q_{\mathrm{fw}} = 0$): the loop of (i) is absent, the last term of (69) vanishes, and along $M \asymp \mathsf{D}_{\mathrm{bw}}k_T^b$ the certificate reduces to $\varrho\theta_G + \varrho q_{\mathrm{bw}}k_T/M \leq \gamma_*/C_{\mathrm{sg}}$: the backward-only bend $(b-1)/(a+b)$ of the compute scaling law (§4.1; derivation in Appendix F.4) is the first-order one and is certified by Proposition E.5 on the whole window. A forward copy in a fixed reference format ($q_{\mathrm{fw}} = 4^{-p_{\mathrm{ref}}} > 0$) is a different model: its loop is present with $\mathsf{D}_\times = \varrho 4^{-p_{\mathrm{ref}}}q_{\mathrm{bw}}$ and (v) applies on the scale $k_T \gtrsim (\sigma^2/4^{-p_{\mathrm{ref}}})^{1/(a-b)}$. (e)

Scope. Exact input; fixed bits, controls and noise (C_L and the onset in (v) depend on them); uniform average; the window; nothing about the algorithm (Conjecture 2 covers $\tau_T \geq M^a/10$ only).

E.5 Proof of Theorem 4

Proposition E.5 gives $I^{\rm out} \leq \Delta^{\rm out}_{\rm DE} \leq I^{\rm out}/(1-\gamma_*)$ for $t \leq T$, and Proposition E.4 the orders of $I^{\rm out} = \sum_r I^{\rm out}_r$ (a sum of nonnegative two-sided bounds is two-sided); the mixed last-iterate entry is $q_{\rm bw}k_t/M \leq C$ times the forward one, giving the last-iterate law of (12), while in the average law of (12) no term dominates uniformly. The lower bound uses (60) only, the orders of $I^{\rm out}$ only $\beta_0 \leq 1-\kappa_0$, and the constants are σ^2-uniform by Lemma E.1; if $\Gamma(T) \to 0$ then $1 \leq \Delta/I \leq (1-\Gamma)^{-1} = 1 + O(\Gamma)$. The sufficient condition (69) is Lemma E.6; the fourth term is Proposition E.9.

The forward term is driven by the squared parameter norm $\mathsf{U}_t \asymp k_t^{a-b+1}$ and is aligned with G, so it lands on the modes being fitted; averaging contracts the part that entered early, which is why the forward term of the average is the smaller $\varrho q_{\rm fw} k_t^{2-b}/M$. The backward term is driven by the residual scale r_t and is isotropic, so it lands on all modes and is the same for both outputs. The interaction term is seen only by the average and carries $q_{\rm fw}q_{\rm bw}$: it is the term that a run rounding a single copy cannot see. At the end of the window the forward term of the last iterate, $\varrho q_{\rm fw} M^{a-b}$, and the interaction term of the average, $\varrho q_{\rm fw}q_{\rm bw}M^{a-b}$, grow with the width, since $a > b$, whereas the exact-arithmetic law does not. The backward term of the last iterate is a factor $q_{\rm bw}$ below the SGD-noise floor $\varrho\sigma^2$ of (11): while the rate is constant, backward rounding is hidden by SGD noise.

E.6 The natural gain and the fourth term

Proposition E.9 (natural gain: a fourth last-iterate term)**.** *There is $C_{\rm nat}(\sigma)$ such that under $\beta_0 \leq 1-\kappa_0$ and $C_{\rm nat}(\sigma)\varrho[\theta_G + (q_{\rm fw}+q_{\rm bw})k_T/M + q_{\rm fw}q_{\rm bw}\tau_T/M] \leq \gamma_* < 1$ (the three masses, without Z_σ), on the window $\Delta^{\rm last}_{\rm DE}(t) \asymp_\sigma I^{\rm last}(t) + \varrho^2 q_{\rm fw}q_{\rm bw}r_t\tau_t/M$ and $\Delta^{\rm avg}_{\rm DE}(t) \asymp_\sigma I^{\rm avg}(t)$. The extra term can dominate: for $a=2$, $b=\frac{3}{2}$, $g_j = j^{-2}$, $w_j = j^{1/2}$, $B=1$, $q_{\rm fw} = M^{-3/4}$, $q_{\rm bw} = M^{-1/4}$, $\iota_r = q_r/M$, $\bar\alpha_r = q_r$, $T = M^2/\eta_0$ with η_0 small, the natural gain and β_0 are $O(\eta_0)$ while $I^{\rm last}(T) \asymp_\sigma \eta_0 M^{-1/4}$ and $\Delta^{\rm last}_{\rm DE}(T) \geq c(\sigma)\eta_0^2$.*

Proof. Use (73) with $\mathsf{J}(t)$ in the norm $\max\{|v_R|, |v_U|/\ell_t\}$, $\ell_t = k_t + \beta M\tau_t$: its row sums are $\leq C$ times the natural gain ($f\beta \asymp m$, $\beta k \leq C$, $k^2 \leq M\tau$); historical suprema of the profiles cost $\leq C(1+\sigma^{-2})$ (only k^{2-b} can decrease, and $\sigma^2 k \geq \sigma^2$ dominates it), hence $C_{\rm nat}(\sigma)$. Inverting the 2×2 resolvent (determinant $\geq (1-\gamma_*)^2$) with $\mathsf{J}_{21}\psi_R \leq C\varrho\psi_U$ gives $U(\delta) \lesssim_\sigma \psi_U$, $R(\delta) \lesssim_\sigma \psi_R + \mathsf{J}_{12}\psi_U$, and every term of $(fJ_2 + mJ_1)\psi_U$ is absorbed by ψ_R except $(\varrho f)(\varrho\beta r_t M\tau_t) \asymp \varrho^2 q_{\rm fw}q_{\rm bw}r_t\tau_t/M$. Lower bound: $U(F^{\rm bw}(s)) \asymp \varrho\iota_{\rm bw}r_s M\tau_s$ at middle and late s, and one forward feedback $\varrho\eta\iota_{\rm fw}g_1 U(F^{\rm bw}(s))$ on $j=1$ over the last quarter gives $\geq c\varrho^2 q_{\rm fw}q_{\rm bw}r_t\tau_t/M$, a nonnegative term of (60). For the average the extra profile's aligned and isotropic responses are smaller than the backward average by the bounded factors $\varrho^2 q_{\rm fw}$ and $\varrho\cdot\varrho q_{\rm fw}q_{\rm bw}\tau_t/M$. The family: substitute $k_T = M$. □

E.7 Cooldown: proof of Proposition E.10

Proposition E.10 (Proposition E.10, full statement)**.** *Let only backward rounding be active, at rate $\eta_{\rm peak}$ for optimisation time τ_1 and then along a cooldown $0 \leq \eta_s \leq \eta_{\rm peak}$ of optimisation time $\tau_c = \phi\tau_1$, with $\eta_{\rm peak}g_1 \leq c_0$ and $2\varrho_{\rm peak}\theta_G/(2-c_0) \leq 1-\kappa_c$; $k_1 = \min(M, \tau_1^{1/a})$, $r_1 = \sigma^2 + k_1^{1-b}$. Then $I^{\rm last}(T) \leq \Delta^{\rm last}_{\rm DE}(T) \leq I^{\rm last}(T)/(1-\Gamma_c)$ with $\Gamma_c \leq C\varrho_{\rm peak}(\theta_G + q_{\rm bw})$ uniformly in the duration, and: (i) the noise injected before the cooldown contributes $\asymp \varrho_{\rm peak}\iota_{\rm bw}r_1k_1\Upsilon(\phi)$, $\Upsilon(\phi) = (1+\phi)^{1/a} - \phi^{1/a}$ (for $\tau_1 + \tau_c \leq c_{\rm win}M^a$): a cooldown as long as the run removes only an $O(1)$ fraction of it ($\Upsilon(1) = 0.41$ at $a=2$), whereas the pre-cooldown SGD noise, aligned and entering through the faster kernel S_2, is reduced by k_c^{1-a}, $k_c = \min(M, \tau_c^{1/a})$; (ii) for a linear cooldown to $\eta_{\rm fin} = \zeta\eta_{\rm peak}$ at $\tau_1 = M^a$, $\phi \geq 1$,*

$$I^{\rm last}(T) \asymp_\sigma \varrho_{\rm peak}q_{\rm bw}\sigma^2\sqrt{\zeta^2 + (1-\zeta^2)/\phi}\,: \tag{79}$$

the noise injected while the rate is still large decays like $\phi^{-1/2}$, down to a floor $\propto \eta_{\rm fin}$; (iii) as $g_j = j^{-a}$, $\tau_1/M^a \to 1$, $\tau_c/M^a \to \phi$ and $\varrho_{\rm peak} \to 0$ at fixed ζ, the retained fraction $\Delta^{\rm last}_{\rm DE}(T)/\Delta^{\rm last}_{\rm DE}(T_1)$ converges to an explicit width-independent curve of (ϕ, ζ) ((83); tested in Appendix G).

Correspondence with the main text. The main text (§4.1) states (79) of (ii) informally as the law (14); the sandwich $I^{\rm last}(T) \le \Delta^{\rm last}_{\rm DE}(T) \le I^{\rm last}(T)/(1-\Gamma_c)$ with $\Gamma_c \le C\varrho_{\rm peak}(\theta_G + q_{\rm bw})$ is the "$\asymp_\sigma$" of that equation, (i) gives the two retention factors quoted after it, and (iii) is the width-independent curve tested in Appendix G. Backward rounding only; $G_c := 2\varrho_{\rm peak}\theta_G/(2-c_0) \le 1-\kappa_c$. The telescoping identity $1-\pi_j(t,0)^2 = \sum_{s<t}\eta_s g_j(2-\eta_s g_j)\pi_j(t,s+1)^2$ with $g_j^2 e_j \le g_j \sum_l g_l e_l$ and $\eta_s^2 \le \eta_{\rm peak}\eta_s$ shows that the exact-arithmetic risk equation has kernel sum $\le G_c$, so $\sigma^2 \le \mathsf{R}^0_s \le R^c_* := (\sigma^2 + \sum_j g_j w_j)/\kappa_c$, and the stable phase has $\beta_0 \le (1-\kappa_c)/(1+\kappa_c)$. The source is $b_s = \iota_{\rm bw}\mathfrak{R}_{\rm bw}(s)$, $\mathfrak{R}_{\rm bw}(s) = \mathsf{R}^0_s + (2\bar\alpha_{\rm bw}/\iota_{\rm bw})\mathsf{S}^0_s \in [\sigma^2, (\beta/\iota_{\rm bw})R^c_*]$.

(i) Exact decomposition and transfer. Unrolling the pure response with $\eta_s^2 = \eta_{\rm fin}\eta_s + \eta_s(\eta_s - \eta_{\rm fin})$ during the cooldown,

$$I^{\rm last}(T) = I_{\rm pre} + I_{\rm fin} + I_{\rm dec}, \qquad I_X = \frac{\iota_{\rm bw}}{2B}\sum_s w_X(s)\,\mathfrak{R}_{\rm bw}(s)\sum_j g_j \pi_j(T, s+1)^2 = \frac{\iota_{\rm bw}}{2B}\,\bar{\mathfrak{R}}_X\, Z_X, \tag{80}$$

with $w_{\rm pre} = \eta^2_{\rm peak}\mathbf{1}_{s<T_1}$, $w_{\rm fin} = \eta_{\rm fin}\eta_s\mathbf{1}_{s\ge T_1}$, $w_{\rm dec} = \eta_s(\eta_s - \eta_{\rm fin})\mathbf{1}_{s\ge T_1}$, $Z_X = \sum_s w_X(s)\sum_j g_j\pi_j(T,s+1)^2$ and $\bar{\mathfrak{R}}_X$ the corresponding weighted mean of $\mathfrak{R}_{\rm bw}$. With $\psi(t) = 2I^{\rm last}(t)$,

$$(K_c x)(t) = \sum_{s<t}\frac{\eta_s^2}{B}\sum_j g_j\pi_j(t,s+1)^2(2g_j+\beta)x(s), \qquad \Gamma_c(T) = \max_{t\le T,\,\psi(t)>0}\frac{(K_c\psi)(t)}{\psi(t)}\,: \tag{81}$$

with $f = m = 0$, (65) reads $(\mathsf{H}\delta)_j \le (2g_j+\beta)\sum_l g_l\delta_l$, so the risk moment obeys $x \le \psi + K_c x$ and the induction of Proposition E.5 gives the sandwich. On $\tau_T \le \Lambda M^a$, composing the kernel once (fix the injection time, sum over the intermediate time) and using Lemma E.2 with $\int_0^u h_2(u-v)h_1(v)dv \le Ch_1(u)$, $\int_0^u h_1(u-v)h_1(v)dv \le C(1+u)^{\varpi}h_1(u)$ for $h_1 = (1+v)^{\varpi-1}$, $h_2 = (1+v)^{\varpi-2}$ (split at $u/2$), gives $\Gamma_c(T) \le C_\Lambda\varrho_{\rm peak}(\theta_G + \beta M) \le C'_\Lambda\varrho_{\rm peak}(\theta_G + q_{\rm bw})$ with σ^2-free constants.

(ii) Pre-cooldown part. The stable-phase geometric sum in (80) is exactly $I_{\rm pre} = \frac{\varrho_{\rm peak}\iota_{\rm bw}}{2}\sum_j \frac{1-(1-\eta_{\rm peak}g_j)^{2T_1}}{2-\eta_{\rm peak}g_j}\widetilde{\mathfrak{R}}_j(T_1)\pi_j(T,T_1)^2$ with the geometric-history means $\widetilde{\mathfrak{R}}_j(T_1) \asymp r_1$ of (64) (valid as $\beta_0 \le (1-\kappa_c)/(1+\kappa_c)$). The identity $(1-e^{-2\tau_1 g})e^{-2\tau_c g} = 2\int_0^{\tau_1} g e^{-2g(u+\tau_c)}du$ and Lemma E.2(i) give, with $K_0 = (1-c_0)^{-1}$, $c\int_0^{\tau_1} S^{\rm c}_1(K_0(u+\tau_c))du \le I_{\rm pre}/(\varrho_{\rm peak}\iota_{\rm bw}r_1) \le C\int_0^{\tau_1} S^{\rm c}_1(u+\tau_c)du$. On $\tau_1 + \tau_c \le c_{\rm win}M^a$, Lemma E.3 makes both integrals $\asymp \int_0^{\tau_1}(1+u+\tau_c)^{\varpi-1}du \asymp \tau_1^{\varpi}\Upsilon_\varpi(\phi)$. At $\tau_1 = M^a$, $\pi_j(T,T_1)^2 \le e^{-2\tau_c g_M} \le e^{-2c\phi}$ and the M remaining terms are bounded. Since $\Upsilon_\varpi(1) = 2^\varpi - 1 > 0$, the profile $1-\phi^\varpi$ is comparable to Υ_ϖ only on $\phi \le \phi_0 < 1$.

(iii) Linear cooldown. Enumerate the cooldown steps backwards, $n = 0, \dots, L_c - 1$: rates $a_n = \eta_{\rm fin} + d_\eta n$, remaining times $v_n = n\eta_{\rm fin} + d_\eta n(n-1)/2$, $v_{n+1} - v_n = a_n$, $\tau_c = L_c(\eta_{\rm peak} + \eta_{\rm fin})/2$. Let $F_\zeta(x) = \sqrt{\zeta^2 + (1-\zeta^2)x} - \zeta$ (increasing, concave). The identity $a_n^2 = \eta^2_{\rm fin} + 2d_\eta v_n + d_\eta(a_n - \eta_{\rm fin})$ and the cell bounds $d_\eta(n-1) \le \sqrt{\eta^2_{\rm fin} + 2d_\eta v} - \eta_{\rm fin} \le d_\eta(n+1)$ on $[v_n, v_{n+1}]$, $n \ge 2$ (cells $n = 0, 1$ merged), show that the excess weight $a_n(a_n - \eta_{\rm fin})$ is comparable to $\eta_{\rm peak}F_\zeta(v/\tau_c)$ integrated over the cell ($2d_\eta\tau_c = (\eta^2_{\rm peak} - \eta^2_{\rm fin})L_c/(L_c - 1)$). Lemma E.2 converts $\sum_s \eta_s(\cdot)\sum_j g_j\pi_j(T,s+1)^2$ into integrals at a fixed exponential rate (a rate change $v \mapsto cv$ costs constants, as $F_\zeta(x)/c \le F_\zeta(x/c) \le F_\zeta(x)$ for $c \ge 1$), giving, with the exact residual means of (80) and $\tau_c \ge \tau_*$,

$$I_{\rm fin} \asymp \varrho_{\rm peak}\iota_{\rm bw}\bar{\mathfrak{R}}_{\rm fin}\,\zeta\int_0^{\tau_c} S^{\rm c}_1(v)dv, \qquad I_{\rm dec} \asymp \varrho_{\rm peak}\iota_{\rm bw}\bar{\mathfrak{R}}_{\rm dec}\int_0^{\tau_c} F_\zeta(v/\tau_c)S^{\rm c}_1(v)dv. \tag{82}$$

Since $\int_0^u S^{\rm c}_1 = \frac{1}{2}\sum_j(1-e^{-2ug_j}) \asymp \min(M, u^\varpi)$, $F_\zeta(v/u) \le 1-\zeta$ and $F_\zeta(v/u) \ge (1-\zeta)/4$ on $[u/4, u/2]$: $I_{\rm fin} \asymp \varrho_{\rm peak}\iota_{\rm bw}\bar{\mathfrak{R}}_{\rm fin}\zeta\min(M, \tau_c^\varpi)$ and $I_{\rm dec} \asymp \varrho_{\rm peak}\iota_{\rm bw}\bar{\mathfrak{R}}_{\rm dec}(1-\zeta)\tau_c^\varpi$ for $\tau_c \le M^a$. For $u = \tau_c \ge M^a$, concavity gives $F_\zeta(v/u) \le \max(1, v/M^a)F_\zeta(M^a/u)$; $\int_0^\infty S^{\rm c}_1 = M/2$ and

$\int_0^\infty v S_1^{\mathrm{c}} = \frac{1}{4}\sum_j g_j^{-1} \asymp M^{a+1}$ give $I_{\mathrm{dec}} \lesssim \varrho_{\mathrm{peak}}\iota_{\mathrm{bw}}\bar{\mathfrak{R}}_{\mathrm{dec}} M F_\zeta(M^a/\tau_c)$, and $v \in [M^a/4, M^a/2]$ with the last half of the modes gives the matching lower bound. Adding the fin and dec weights before averaging gives (79) at $\tau_1 = M^a$, $\phi \geq 1$. For fixed noise, every residual mean lies in $[\sigma^2, (\beta/\iota_{\mathrm{bw}})R_*^c]$, $\sqrt{\zeta^2 + (1-\zeta^2)/\phi} \geq \phi^{-1/2}$ absorbs $I_{\mathrm{pre}} \lesssim M e^{-c\phi}$, and for $\phi \leq 1$ the last half of the modes retains a fixed fraction of M while the cooldown injection is at most its peak stationary bound (telescoping).

Gain uniform in the duration (outline). For $\tau_t \leq H_M := 16(1+\eta_{\max})M^a$ use (i). For $\tau_t \geq H_M$, solving the quadratic identity shows that the reverse-time local rate seen from t is $\leq C h_t(1+\sqrt{v/H_M})$ and $\geq c h_t$ on $[H_M/4, H_M/2]$, $h_t = \min\{\eta_{\mathrm{peak}}, \eta_{t-1} + \sqrt{d_\eta H_M}\}$, so the unit-source response $J(t) = \frac{1}{B}\sum_{s<t}\eta_s^2 \sum_j g_j \pi_j(t, s+1)^2$ is $\asymp M h_t/B$ (Lemma E.2 with $\int_0^\infty S_1^{\mathrm{c}} = M/2$, $\int_0^\infty \sqrt{v} S_1^{\mathrm{c}} \leq CM\sqrt{H_M}$). The source floor $\iota_{\mathrm{bw}}\sigma^2 J \leq \psi \leq \beta R_*^c J$ gives $\psi(s) \leq C(\sigma)(1+\sqrt{(\tau_t - \tau_s)/H_M})\psi(t)$, the only step using $\sigma^2 > 0$; inserted into (81) it yields $\Gamma_c \leq C(\sigma)\varrho_{\mathrm{peak}}\int_0^\infty (1+\sqrt{v/H_M})[2S_2^{\mathrm{c}} + \beta S_1^{\mathrm{c}}]dv \leq C_{\mathrm{lin}}(\sigma)\varrho_{\mathrm{peak}}(\theta_G + q_{\mathrm{bw}})$; $C_{\mathrm{lin}}(\sigma)$ may depend on $(\beta/\iota_{\mathrm{bw}})R_*^c/\sigma^2$ and is not tracked.

The leading curve. If $g_j = j^{-a}$, $\tau_1/M^a \to 1$, $\tau_c/M^a \to \phi$, $\eta_{\mathrm{fin}}/\eta_{\mathrm{peak}} \to \zeta$ and $\varrho_{\mathrm{peak}} \to 0$ with $\eta_{\mathrm{peak}} g_1 \leq c_0$, at fixed $\sigma^2 > 0$ and $q_{\mathrm{bw}} > 0$, then

$$\frac{\Delta_{\mathrm{DE}}^{\mathrm{last}}(T)}{\Delta_{\mathrm{DE}}^{\mathrm{last}}(T_1)} \to \frac{\mathcal{P}_a(\phi) + \mathcal{C}_a(\phi,\zeta)}{\mathcal{P}_a(0)}, \qquad \mathcal{P}_a(\phi) = \int_0^1 (1 - e^{-2x^{-a}})e^{-2\phi x^{-a}}dx, \tag{83}$$

$\mathcal{C}_a(\phi,\zeta) = 2\int_0^\phi \sqrt{\zeta^2 + (1-\zeta^2)y/\phi}\int_0^1 x^{-a}e^{-2yx^{-a}}dx\,dy$: the curve plotted in Figure 5(b).

(iv) Leading curve. For a frozen source $b_s \equiv b_*$ and pure response, on $x = j/M \in [\delta, 1]$: $|\log \pi_j(t,u)^2 + 2g_j\sum_{v=u}^{t-1}\eta_v| \leq \eta_{\max} g_j^2(\tau_t - \tau_u)/(1-c_0) \to 0$, the reverse cell mesh $\eta_{\max}/M^a \to 0$, and the quadratic identity gives the limiting reverse rate $\eta_{\mathrm{peak}}\sqrt{\zeta^2 + (1-\zeta^2)y/\phi}$; the Riemann sums converge and $I^{\mathrm{last}}(T)/(\varrho_{\mathrm{peak}} b_* M/4) \to \mathcal{P}_a(\phi) + \mathcal{C}_a(\phi,\zeta)$ (the stable geometric sum supplies the $\frac{1}{4}$; $\mathcal{N}_a := \mathcal{P}_a(0) > 0$ is the normalization at T_1); modes $j < \delta M$ contribute $\leq C\delta$ (telescoping); let $\delta \downarrow 0$. For the actual source, Lemma E.1 and the risk-kernel bound $G_c \to 0$ give $\mathfrak{R}_{\mathrm{bw}}(s)/\sigma^2 \to 1$ outside the first εM^a of optimization time, whose normalized contribution is $\leq C(\sigma)\varepsilon\delta^{-a}$; hence $b_* = \iota_{\mathrm{bw}}\sigma^2$. Finally $\Gamma_c \leq C_{\mathrm{lin}}(\sigma)\varrho_{\mathrm{peak}}(\theta_G + q_{\mathrm{bw}}) \to 0$ transfers the limit to $\Delta_{\mathrm{DE}}^{\mathrm{last}}$ at T and T_1. At fixed nonzero ϱ_{peak} no equality to the curve is claimed.

The exact-arithmetic risk of the cooled last iterate. Proposition E.10, the cooled row of Corollary F.1 and the cooled rows of Theorem F.8 compare the rounding penalty with the exact-arithmetic risk $\mathcal{E}_0^{\mathrm{cooled}}(T)$ of the cooled last iterate. The following lemma gives its order. Write

$$D_a(M) = M^{1-a}\ (a<2), \qquad D_a(M) = M^{-1}\log M\ (a=2), \qquad D_a(M) = M^{-a/2}\ (a>2),$$

so that $D_a(M) = M^{-d_a}\ell_a(M)$ with $d_a = \min\{a-1, a/2\}$, $\ell_a(M) = \log M$ for $a = 2$ and $\ell_a(M) = 1$ otherwise, and $\mathsf{n}_{\mathrm{peak}} = \varrho_{\mathrm{peak}}\sigma^2 M^{b-1}$ for the noise ratio at the peak rate.

Lemma E.11 (exact-arithmetic risk of the cooled last iterate)**.** *Let Assumption 3(a) hold, and run the deterministic-equivalent dynamics in exact arithmetic at the rate η_{peak} for optimization time $\tau_1 = M^a$ and then along a linear cooldown to $\eta_{\mathrm{fin}} = \zeta\eta_{\mathrm{peak}}$, $0 \leq \zeta \leq 1$, of optimization time $\tau_c = \phi\tau_1$, $\phi \geq 1$, with $\eta_{\mathrm{peak}} g_1 \leq c_0$ and $2\varrho_{\mathrm{peak}}\theta_G/(2-c_0) \leq 1 - \kappa_c$. Then*

$$\begin{aligned} \mathcal{E}_0^{\mathrm{cooled}}(T) &\geq c\,e^{-C\phi}M^{1-b} + c\,\varrho_{\mathrm{peak}}\sigma^2\big[\zeta + \phi^{-1/2}D_a(M)\big], \\ \mathcal{E}_0^{\mathrm{cooled}}(T) &\leq C\,M^{1-b} + C\,\frac{R_*^c}{\sigma^2}\,\varrho_{\mathrm{peak}}\sigma^2\big[\zeta + \phi^{-1/2}D_a(M)\big], \end{aligned} \tag{84}$$

with $R_*^c = (\sigma^2 + \sum_j g_j w_j)/\kappa_c$ and constants c, C depending only on $(a, b, c, C, c_0, \theta_G)$. In particular, at fixed ϕ, $\mathcal{E}_0^{\mathrm{cooled}}(T) \asymp_\sigma M^{1-b}$ if and only if $\zeta\,\mathsf{n}_{\mathrm{peak}} \lesssim 1$ and $\mathsf{n}_{\mathrm{peak}} D_a(M) \lesssim 1$. At fixed controls and noise the second condition holds for all large M if and only if $b \leq 1 + a/2$, since $M^{b-1}D_a(M) = M^{b-1-d_a}\ell_a(M)$.

Proof. Write $e_j(t)$ for the exact-arithmetic mode energies, $\mathsf{R}_s = \sigma^2 + \sum_l g_l e_l(s)$, $\chi_s = \tau_T - \tau_s$ and $K_0 = (1-c_0)^{-1}$. The recursion (10) with all rounding coefficients zero unrolls to

$$e_j(T) = \pi_j(T)^2 w_j + \sum_{s<T} \pi_j(T, s+1)^2 \, \frac{\eta_s^2}{B}\big(\mathsf{R}_s g_j + g_j^2 e_j(s)\big),$$

with nonnegative terms. The injection lies between $\frac{\eta_s^2}{B}\sigma^2 g_j$ and $\frac{\eta_s^2}{B}\, 2g_j R_*^c$, because $g_j^2 e_j(s) \le g_j \mathsf{R}_s$ and $\sigma^2 \le \mathsf{R}_s \le R_*^c$ under the stability condition (the kernel-mass bound $G_c \le 1 - \kappa_c$ stated after Proposition E.10). Hence, with $\mathcal{E}_0^{\mathrm{cooled}}(T) = \frac{1}{2}\sum_j g_j e_j(T)$,

$$\mathrm{Bias} + \frac{\sigma^2}{2B} N \;\le\; \mathcal{E}_0^{\mathrm{cooled}}(T) \;\le\; \mathrm{Bias} + \frac{R_*^c}{B} N,$$
$$\mathrm{Bias} = \tfrac{1}{2}\sum_j g_j w_j \pi_j(T)^2, \qquad N = \sum_{s<T} \eta_s^2 \sum_j g_j^2 \pi_j(T, s+1)^2.$$

Bias. By Lemma E.2(i), $e^{-2K_0 g_j \tau_T} \le \pi_j(T)^2 \le e^{-2g_j\tau_T}$ with $\tau_T = (1+\phi)M^a$, and $g_j w_j \asymp j^{-b}$, $g_j \asymp j^{-a}$ for $j \le M$ by Assumption 3(a). For the upper bound, $g_j\tau_T \ge c(1+\phi)(M/j)^a \ge 2c(M/j)^a$, so $\mathrm{Bias} \le C\sum_{j\le M} j^{-b} e^{-4c(M/j)^a} \le C' M^{1-b}\int_0^1 x^{-b} e^{-4cx^{-a}} dx$, uniformly in $\phi \ge 1$. For the lower bound, the modes $j \in [M/2, M]$ have $g_j\tau_T \le C 2^a(1+\phi)$ and contribute $\sum_{M/2\le j\le M} j^{-b}\, e^{-2K_0 C 2^a(1+\phi)} \ge c\, e^{-C'\phi} M^{1-b}$.

Noise. By Lemma E.2(i), $\pi_j(T, s+1)^2$ lies between $e^{-2K_0 g_j \chi_{s+1}}$ and $e^{-2g_j\chi_{s+1}}$, so the inner sum lies between $S_2^{\mathrm{c}}(K_0\chi_{s+1})$ and $S_2^{\mathrm{c}}(\chi_{s+1})$. As in part (iii) of the proof of Proposition E.10 (the cell bounds of the linear ramp and Lemma E.2(ii); a change of rate $v \mapsto K_0 v$ costs constants), the sum over s is comparable to the integral in the remaining time,

$$N \asymp \int_0^{\tau_c} \eta(v)\, S_2^{\mathrm{c}}(v)\, dv + \eta_{\mathrm{peak}} \int_{\tau_c}^{\tau_T} S_2^{\mathrm{c}}(v)\, dv, \qquad \eta(v) = \eta_{\mathrm{peak}}\sqrt{\zeta^2 + (1-\zeta^2)v/\tau_c},$$

where $\eta(v)$ is the rate of the linear cooldown at remaining time v (the quadratic identity of part (iii)). Three integrals of $S_2^{\mathrm{c}}(v) = \sum_j g_j^2 e^{-2g_j v}$ are needed: $\int_0^\infty S_2^{\mathrm{c}} = \frac{1}{2}\theta_G$; $\int_0^\infty v^{1/2} S_2^{\mathrm{c}}(v) dv = \frac{\sqrt{\pi}}{2^{5/2}}\sum_j g_j^{1/2}$, and $\sum_{j\le M} g_j^{1/2} \asymp \sum_{j\le M} j^{-a/2} \asymp M^{a/2} D_a(M)$; and $\int_{\tau_c}^\infty S_2^{\mathrm{c}} = \frac{1}{2}\sum_j g_j e^{-2g_j\tau_c} \le C M^{1-a} e^{-c\phi}$, since $g_j\tau_c \ge c\phi(M/j)^a$ for $j \le M$. *Upper bound.* $\eta(v) \le \eta_{\mathrm{peak}}(\zeta + \sqrt{v/\tau_c})$ gives $\int_0^{\tau_c} \eta S_2^{\mathrm{c}} \le \eta_{\mathrm{peak}}[\frac{1}{2}\zeta\theta_G + C\tau_c^{-1/2} M^{a/2} D_a(M)] = \eta_{\mathrm{peak}}[\frac{1}{2}\zeta\theta_G + C\phi^{-1/2} D_a(M)]$, and the pre-cooldown part is at most $C\eta_{\mathrm{peak}} M^{1-a} e^{-c\phi} \le C'\eta_{\mathrm{peak}}\phi^{-1/2} D_a(M)$, because $M^{1-a} \le D_a(M)$ for every $a > 1$. *Lower bound.* On $[0,1]$, $S_2^{\mathrm{c}} \ge e^{-2g_1}\sum_j g_j^2 \ge c$, so $\int_0^{\tau_c} \eta S_2^{\mathrm{c}} \ge c\,\zeta\eta_{\mathrm{peak}}$. Moreover $\eta(v) \ge \eta_{\mathrm{peak}}\sqrt{1-\zeta^2}\sqrt{v/\tau_c}$, and for $j \le M$ the integral of $v^{1/2} e^{-2g_j v}$ over $[0, \tau_c]$ is at least a fixed fraction of its integral over $[0, \infty)$, since $g_j\tau_c \ge c\phi \ge c$; so $\int_0^{\tau_c} \eta S_2^{\mathrm{c}} \ge c\sqrt{1-\zeta^2}\,\eta_{\mathrm{peak}}\phi^{-1/2} D_a(M)$. If $\zeta \le \frac{1}{2}$ this is at least $c'\eta_{\mathrm{peak}}\phi^{-1/2} D_a(M)$; if $\zeta > \frac{1}{2}$, then $\zeta \ge \frac{1}{2} \ge c''\phi^{-1/2} D_a(M)$ because D_a is bounded. Hence $N \asymp \eta_{\mathrm{peak}}[\zeta + \phi^{-1/2} D_a(M)]$, and $\eta_{\mathrm{peak}}/B = \varrho_{\mathrm{peak}}$ gives (84).

The range. By (84), $\mathcal{E}_0^{\mathrm{cooled}}(T) \asymp_\sigma M^{1-b}$ at fixed ϕ if and only if $\varrho_{\mathrm{peak}}\sigma^2[\zeta + \phi^{-1/2} D_a(M)] \lesssim M^{1-b}$, that is, $\zeta \mathsf{n}_{\mathrm{peak}} \lesssim 1$ and $\mathsf{n}_{\mathrm{peak}} D_a(M) \lesssim 1$. With $\mathsf{n}_{\mathrm{peak}} = \varrho_{\mathrm{peak}}\sigma^2 M^{b-1}$, the second condition reads $\varrho_{\mathrm{peak}}\sigma^2 M^{b-1-d_a}\ell_a(M) \lesssim 1$: for $a < 2$ the exponent is $b - a < 0$, for $a = 2$ it is $b - 2 < 0$, and for $a > 2$ it is $b - 1 - a/2$, which is ≤ 0 if and only if $b \le 1 + a/2$. □

Lemma E.11 compares with the constant-rate floor as follows: a cooldown replaces the SGD-noise floor $\varrho_{\mathrm{peak}}\sigma^2$ of Lemma E.1 by the stationary level $\zeta\varrho_{\mathrm{peak}}\sigma^2$ of the terminal rate plus the contribution $\varrho_{\mathrm{peak}}\sigma^2\phi^{-1/2} D_a(M)$ of the noise added along the ramp, which vanishes as $M \to \infty$; the noise added before the cooldown contributes at most $\varrho_{\mathrm{peak}}\sigma^2 M^{1-a} e^{-c\phi}$, the factor k_c^{1-a} of Proposition E.10(i) with $k_c = M$.

E.8 Proof of Theorem E.12: backward rounding, comparison for the exact recursion

The main text states the following theorem in one sentence (§3); this is its full form.

Theorem E.12 (backward rounding: the law holds for the algorithm; full statement)**.** *Let only the backward copy be rounded ($p_{\mathrm{fw},t} = \infty$, exact input), with the Gaussian feature map of §2, $d \gtrsim \log M$, constant controls, and Assumption 3(a),(b),(d). Then, with probability at least $1 - M^{-D}$ over S and simultaneously for every precision p_{bw}, on the window and for both output rules,*

$$\mathcal{E}^{\mathrm{out}}(t) = \mathcal{E}_0^{\mathrm{out}}(t) + \Delta^{\mathrm{out}}(t), \qquad c\,(\log M)^{1-a} \;\le\; \frac{\Delta^{\mathrm{out}}(t)}{\varrho\, q_{\mathrm{bw}}\, r_t k_t / M} \;\le\; C \log M, \tag{85}$$

the upper bound under the small-gain condition (e) for the comparison dynamics of the proof.

The mechanism is a comparison. Write B_t and p_t for the batch size and the backward precision at step t, and put $\beta_{-,t} = \min_k m_{p_t}(G_{kk})$ and $\beta_{+,t} = \frac{16}{5} \max_k m_{p_t}(G_{kk})$, the smallest and (a multiple of) the largest mean rounding variance over the feature coordinates. Let $C_t^{\pm}$ solve

$$C_{t+1}^{\pm} = H_t C_t^{\pm} H_t + \tfrac{\eta_t^2}{B_t}\big[G C_t^{\pm} G + \big(\sigma_{\mathrm{eff}}^2 + \mathrm{tr}(G C_t^{\pm})\big)(G + \beta_{\pm,t} I)\big], \qquad C_0^{\pm} = C_0. \tag{86}$$

In the eigenbasis of G, (86) is the deterministic-equivalent dynamics (10) with the backward coefficient ι_{bw} replaced by $\beta_{\pm,t}$ and all other rounding coefficients zero. For every deterministic schedule (η_t, B_t, p_t) and every fixed-weight output $\sum_t w_t v_t$, that is, the last iterate, the average and any cooldown, the second moment of the algorithm is sandwiched, $C_t^- \preceq C_t \preceq C_t^+$, and so is its rounding penalty, with no stability condition (Theorem E.16). The comparison is made between second moments of whole trajectories in the positive-semidefinite order, which is why signed weights and sign-changing propagators are allowed. For the Gaussian feature map, $c\,4^{-p}/[M(\log M)^{a-1}] \le \beta_- \le \beta_+ \le C\,4^{-p} \log M/M$ on one event for all p (Theorem E.18), and Theorem 4 applied to the two comparison dynamics gives (85) (Corollary E.17 and §E.8.6). The backward row of Theorem 4 and the backward rows of the precision requirements of §4.1 are therefore statements about the rounded algorithm, while the forward and interaction terms go through Conjecture 2.

This subsection proves Theorem E.12 (both its sandwich mechanism and the law (85)). Notation: $h = \eta^2/B$, $s = \sigma_{\mathrm{eff}}^2$, D_t the perturbation of the second moment, $\mathcal{B}$ the backward rounding map; Theorems E.16 and E.18 below are the sandwich and the coefficient bounds behind Theorem E.12, and the law (85) is proved at the end. The stability criterion of Proposition E.19 carries the hypothesis $\eta(1 + 1/B)g_1 < 2$ and states the existence of the averaged limit explicitly.

E.8.1 Definitions, conditioning, and the rounding-error moments

Fix a full-row-rank sketch $S \in \mathbb{R}^{M \times d}$ and positive definite $\Sigma \in \mathbb{R}^{d \times d}$, and put $G = S\Sigma S^\top \succ 0$. First condition on both S and a fixed teacher θ. Fresh Gaussian features and labels have the representation

$$z \sim N(0, G), \qquad y = z^\top v_* + \epsilon, \qquad \epsilon \perp z, \quad \mathbb{E}\epsilon = 0, \quad \mathbb{E}\epsilon^2 = s \ge 0,$$

where $v_* = G^{-1} S\Sigma\theta$ and $s = \sigma^2 + \theta^\top(\Sigma - \Sigma S^\top G^{-1} S\Sigma)\theta$. Let $A_S = (s - \sigma^2)/2$. Test excess risk is $A_S + \frac{1}{2}\|v - v_*\|_G^2$. For deterministic finite real steps η_t, batch sizes $B_t \in \mathbb{N}$, and precisions $p_t \in \mathbb{N}_0 \cup \{\infty\}$, the actual update is

$$v_{t+1} = v_t - \frac{\eta_t}{B_t} \sum_{\ell=1}^{B_t} Q_{p_t}(z_{t\ell})\,(z_{t\ell}^\top v_t - y_{t\ell}). \tag{87}$$

All samples and coordinate rounding decisions are fresh and independent of past samples and rounding decisions. Conditional on the fixed teacher, the current feature vector, and the past, the rounding errors are centered, independent across coordinates, and independent of the current label noise. Consequently they are conditionally independent of the current label residual. Input, forward operand, and all other arithmetic are exact. Initialization is deterministic.

The ideal binary SR variance is $\nu_p(x)$; write

$$m_p(g) = \mathbb{E}\nu_p(\sqrt{g}Z), \qquad \alpha_p(g) = \tfrac{1}{2}\mathbb{E}[\nu_p(\sqrt{g}Z)(Z^2 - 1)], \qquad Z \sim N(0,1).$$

We use the following all-precision consequences of Lemma B.1(2) and (4):

$$\frac{4^{-p} g}{192} \le m_p(g) \le \frac{4^{-p} g}{4}, \qquad 0 \le \alpha_p(g) \le \frac{11}{10} m_p(g). \tag{88}$$

The lower constant follows from $[24(1+2^{-p})^3]^{-1} \geq 1/192$. At $p=\infty$ all three functions vanish and $4^{-\infty}=0$. The sharper ratio 1.03 of Lemma B.1(4) could replace 1.1; it is not needed here.

Set $\delta_t = v_t - v_*$, $C_t = \mathbb{E}[\delta_t \delta_t^\top \mid S, \theta]$, $H_t = I - \eta_t G$, and $\varepsilon_t = \eta_t^2 / B_t$. Define

$$M_t = \operatorname{diag}_k m_{p_t}(G_{kk}), \qquad \Lambda_t(X) = \operatorname{diag}_k \left(\frac{\alpha_{p_t}(G_{kk})}{G_{kk}} X_{kk} \right), \tag{89}$$

$$\mathcal{B}_t(X) = \operatorname{tr}(GX) M_t + 2\Lambda_t(GXG), \tag{90}$$

$$\mathcal{V}_0(X) = GXG + \operatorname{tr}(GX)G, \qquad \mathcal{L}_t^0(X) = H_t X H_t^\top + \varepsilon_t \mathcal{V}_0(X). \tag{91}$$

Every expectation in the first part of this subsection is over training and rounding, with S, θ fixed. Finiteness at each finite horizon follows by induction from the Gaussian fourth moments, the finite SR variance envelope, and the recursions proved below. There is no asymptotic-in-time assumption.

E.8.2 Exact innovations, positivity, and a local sandwich

Lemma E.13 (Exact innovation covariance)**.** *With $\mathcal{F}_t$ containing the past and the fixed teacher, $\xi_{t+1} := \delta_{t+1} - H_t \delta_t$ satisfies*

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0,$$

and

$$\mathbb{E}[\xi_{t+1}\xi_{t+1}^\top \mid \mathcal{F}_t] = \varepsilon_t \{ \mathcal{V}_0(\delta_t \delta_t^\top) + sG + \mathcal{B}_t(\delta_t \delta_t^\top) + sM_t \}. \tag{92}$$

Consequently

$$C_{t+1} = \mathcal{L}_t^0(C_t) + \varepsilon_t sG + \varepsilon_t \{ \mathcal{B}_t(C_t) + sM_t \}. \tag{93}$$

The maps $\mathcal{V}_0, \mathcal{L}_t^0, \mathcal{B}_t$ are completely positive on real matrices.

Proof. For one sample put $r = z^\top \delta - \epsilon$ and $g = Q_p(z) r$. Conditional centering gives $\mathbb{E}[g \mid \delta] = G\delta$ and

$$\mathbb{E}[gg^\top \mid \delta] = \mathbb{E}[(zz^\top + D_p(z))((z^\top \delta)^2 + s)], \qquad D_p(z) = \operatorname{diag}_k \nu_p(z_k).$$

Gaussian fourth moments give $\mathbb{E}[zz^\top (z^\top \delta)^2] = \operatorname{tr}(G\delta\delta^\top) G + 2G\delta\delta^\top G$. Conditioning the Gaussian vector on its kth coordinate gives

$$\mathbb{E}[\nu_p(z_k)(z^\top \delta)^2] = m_p(G_{kk}) \operatorname{tr}(G\delta\delta^\top) + \frac{2\alpha_p(G_{kk})}{G_{kk}} (G\delta\delta^\top G)_{kk}.$$

Subtract $(G\delta)(G\delta)^\top$, and use the independence of the B_t samples to obtain (92). Martingale centering eliminates the cross terms with $H_t \delta_t$, proving (93).

Complete positivity of $X \mapsto AXA^\top$ is immediate. A map $X \mapsto \operatorname{tr}(PX)Q$ is completely positive for $P, Q \succeq 0$: spectral decompositions $P = \sum_i p_i u_i u_i^\top$, $Q = \sum_j q_j v_j v_j^\top$ give Kraus matrices $\sqrt{p_i q_j} v_j u_i^\top$. This proves the claims for $\mathcal{V}_0, \mathcal{L}_t^0$. Independently of any sign of α_p,

$$\mathcal{B}_t(X) = \mathbb{E}_z[D_{p_t}(z)(z^\top X z)] = \sum_k \mathbb{E}_z[A_k(z) X A_k(z)^\top], \quad A_k(z) = \sqrt{\nu_{p_t}(z_k)} e_k z^\top.$$

This integrable Kraus representation proves the last claim. □

Lemma E.14 (Backward arithmetic sandwich)**.** *For every $X \succeq 0$, with $r = \operatorname{tr}(GX)$,*

$$rM_t \preceq \mathcal{B}_t(X) \preceq \frac{16}{5} rM_t. \tag{94}$$

Put

$$\beta_{-,t} = \min_k m_{p_t}(G_{kk}), \qquad \beta_{+,t} = \frac{16}{5} \max_k m_{p_t}(G_{kk}). \tag{95}$$

Then

$$\beta_{-,t}(s+r) I \preceq sM_t + \mathcal{B}_t(X) \preceq \beta_{+,t}(s+r) I. \tag{96}$$

The same statements hold with any smaller nonnegative lower coefficient and any larger upper coefficient.

Proof. Set $Y = G^{1/2}XG^{1/2}$. Since $Y \succeq 0$, $Y \preceq (\operatorname{tr} Y)I$. Congruence by $G^{1/2}$ gives

$$GXG \preceq rG, \qquad (GXG)_{kk} \le rG_{kk}, \qquad 0 \le [2\Lambda_t(GXG)]_{kk} \le \frac{11}{5} m_{p_t}(G_{kk})r.$$

Adding rM_t proves (94). For the noise term use the scalar inequality

$$s + \frac{16}{5}r \le \frac{16}{5}(s + r).$$

Together with the extrema of the diagonal of M_t, this proves (96). All differences here are diagonal PSD matrices. □

Proposition E.15 (Exact positive perturbation)**.** *Let C_t^0 be the matched exact-arithmetic recursion, including its sampling noise:*

$$C_{t+1}^0 = \mathcal{L}_t^0(C_t^0) + \varepsilon_t sG, \qquad C_0^0 = C_0.$$

Then $D_t = C_t - C_t^0$ satisfies

$$D_{t+1} = \mathcal{L}_t^0(D_t) + \varepsilon_t \mathcal{B}_t(D_t) + f_t, \quad f_t = \varepsilon_t\{\mathcal{B}_t(C_t^0) + sM_t\} \succeq 0, \quad D_0 = 0. \tag{97}$$

Thus $D_t \succeq 0$. Defining the exact-arithmetic causal response in the usual ordered way gives the finite-time insertion expansion $D^{[1]} = \mathcal{R}_0 f$, $D^{[n+1]} = \mathcal{R}_0(\varepsilon \mathcal{B} D^{[n]})$, $D = \sum_{n\ge1} D^{[n]}$, with PSD terms and only finitely many nonzero terms at each time.

Proof. Subtract the exact-arithmetic recursion from (93). Positivity and induction give the claims. Each additional insertion requires an additional past time, so the expansion terminates on every finite time triangle. □

E.8.3 The actual algorithm is bracketed, also after averaging

Theorem E.16 (Finite-horizon comparison for all linear outputs)**.** *Under the preceding conditional model define $C_t^{\pm}$ by*

$$C_{t+1}^{\pm} = \mathcal{L}_t^0(C_t^{\pm}) + \varepsilon_t sG + \varepsilon_t \beta_{\pm,t}\{s + \operatorname{tr}(GC_t^{\pm})\}I, \qquad C_0^{\pm} = C_0. \tag{98}$$

For every finite T and $0 \le t \le T$,

$$C_t^0 \preceq C_t^- \preceq C_t \preceq C_t^+. \tag{99}$$

There are comparison processes $\delta_t^{\pm}$ with these second moments, the same initialization and conditional mean transitions H_t, such that their stacked trajectory second moments *satisfy*

$$\mathbb{E}[Z_T^0(Z_T^0)^\top] \preceq \mathbb{E}[Z_T^-(Z_T^-)^\top] \preceq \mathbb{E}[Z_T^{\mathrm{SR}}(Z_T^{\mathrm{SR}})^\top] \preceq \mathbb{E}[Z_T^+(Z_T^+)^\top], \quad Z_T = (\delta_0^\top, \dots, \delta_T^\top)^\top. \tag{100}$$

In particular, for deterministic real weights $w_0, \dots, w_T$, set

$$L^\diamond(w) = \frac{1}{2}\mathbb{E}\left\|\sum_{t=0}^{T} w_t \delta_t^\diamond\right\|_G^2, \qquad \Delta^\diamond(w) = L^\diamond(w) - L^0(w).$$

Then

$$0 \le \Delta^-(w) \le \Delta^{\mathrm{SR}}(w) \le \Delta^+(w). \tag{101}$$

For weights summing to one these are excess-risk premiums of the parameter outputs $\sum_t w_t v_t$. This includes $w_T = 1$ and $w_1 = \cdots = w_T = 1/T$, $w_0 = 0$. ***No stability condition or bound on $\eta_t g_1$ is required.***

Proof. The maps $X \mapsto \mathcal{L}_t^0(X) + \varepsilon_t\beta_{\pm,t}\operatorname{tr}(GX)I$ and the actual homogeneous map are positive. The source brackets in Lemma E.14, induction, and the common initial condition prove (99).

For the stronger assertion define

$$V_t^{\pm}(X) = \varepsilon_t\{\mathcal{V}_0(X) + sG + \beta_{\pm,t}(s + \operatorname{tr}(GX))I\}, \quad V_t^{\mathrm{SR}}(X) = \varepsilon_t\{\mathcal{V}_0(X) + sG + \mathcal{B}_t(X) + sM_t\}.$$

These maps take PSD matrices to PSD matrices and are order preserving in X. Construct a comparison process by $\delta^{\pm}_{t+1} = H_t\delta^{\pm}_t + [V^{\pm}_t(\delta^{\pm}_t(\delta^{\pm}_t)^\top)]^{1/2}\omega_{t+1}$, with independent standard Gaussian ω_{t+1}. It has finite second moments by induction and satisfies (98). The actual process has the corresponding martingale innovation by Lemma E.13.

Let F_t append $H_t\delta_t$ to a stacked vector Z_t, and let J_t insert an M-vector into the final block of Z_{t+1}. If K_t is its second-moment matrix and $[K_t]_{tt}$ its final diagonal block, then

$$K^{\diamond}_{t+1} = F_tK^{\diamond}_tF_t^\top + J_tV^{\diamond}_t([K^{\diamond}_t]_{tt})J_t^\top.$$

Cross terms vanish because the innovation is centered conditional on the past. Suppose $K^+_t \succeq K^{\mathrm{SR}}_t$. The deterministic terms preserve this order, and

$$\begin{aligned}V^+_t([K^+_t]_{tt}) - V^{\mathrm{SR}}_t([K^{\mathrm{SR}}_t]_{tt}) = \{&V^+_t([K^+_t]_{tt}) - V^+_t([K^{\mathrm{SR}}_t]_{tt})\}\\ &+ \{V^+_t([K^{\mathrm{SR}}_t]_{tt}) - V^{\mathrm{SR}}_t([K^{\mathrm{SR}}_t]_{tt})\} \succeq 0.\end{aligned}$$

The first brace uses positivity of the homogeneous part; the second uses (96). The lower and exact-arithmetic comparisons follow identically. Their initial stacked matrices agree. This proves (100) by induction. Congruence by $(w_0I, \dots, w_TI)$ and the PSD risk readout $G/2$ give (101). This proof never assumes that H_t is PSD. □

Corollary E.17 (Scalar computation and teacher averaging)**.** *Let $G = \sum_j g_j\phi_j\phi_j^\top$ and $e^{\pm}_j(t) = \phi_j^\top C^{\pm}_t\phi_j$. Their exact recursion is*

$$e^{\pm}_j(t+1) = (1-\eta_tg_j)^2e^{\pm}_j(t) + \frac{\eta_t^2}{B_t}\left[g_j^2e^{\pm}_j(t) + (g_j + \beta_{\pm,t})\left(s + \sum_\ell g_\ell e^{\pm}_\ell(t)\right)\right]. \tag{102}$$

For $t \geq s$, all three processes obey $\mathbb{E}[\delta_t\delta_s^\top] = H_{t-1}\cdots H_sC_s$. Thus their averaged risks are computed exactly from (102)*, or from $X_1 = C_1$, $X_{t+1} = H_tX_t + C_{t+1}$ and*

$$L(\mathrm{av}, T) = \frac{1}{2T^2}\,\mathrm{tr}\left[G\sum_{t=1}^{T}(X_t + X_t^\top - C_t)\right]. \tag{103}$$

The theorem also holds after averaging a centered teacher of covariance C_θ, conditional on S, with

$$C_0 = v_0v_0^\top + V_*, \quad V_* = G^{-1}S\Sigma C_\theta\Sigma S^\top G^{-1}, \quad s = \sigma^2 + \mathrm{tr}[C_\theta(\Sigma - \Sigma S^\top G^{-1}S\Sigma)].$$

No commutativity of V_* and G is assumed.

Proof. Projection of (98) gives (102); its off-diagonal initial entries cannot enter this recursion. Conditional mean propagation gives the two-time identity by iteration and the tower property, and summing the two-time second moments proves (103). Conditional PSD inequalities can be averaged over the teacher. The comparison recursions are linear in (C_0, s), so their averages are exactly the stated systems. □

E.8.4 Gaussian sketch: proved logarithmic coefficient bounds

Theorem E.18 (One event valid for every precision)**.** *Assume $S = M^{-1/2}Z$ with independent standard Gaussian entries, $\Sigma = \mathrm{diag}(\lambda_j)_{j\leq d}$ and $c_\lambda j^{-a} \leq \lambda_j \leq C_\lambda j^{-a}$, $a > 1$. Fix $D > 0$, $M \geq 2$, put*

$$x = (D+2)\log M, \quad K = \lceil 16(D+2)\log M\rceil, \quad L_1 = C_\lambda\big(1 + \tfrac{1}{a-1}\big), \quad L_2 = C_\lambda^2\big(1 + \tfrac{1}{2a-1}\big),$$

and assume $d \geq K$. With probability at least $1 - M^{-D}$, simultaneously for every $p \in \mathbb{N}_0 \cup \{\infty\}$,

$$\frac{c_\lambda 4^{-p}}{384MK^{a-1}} \leq \beta_-(p), \qquad \beta_+(p) \leq \frac{4\,4^{-p}}{5M}[L_1 + 2\sqrt{L_2x} + 2C_\lambda x]. \tag{104}$$

In particular these have orders $4^{-p}/[M(\log M)^{a-1}]$ and $4^{-p}\log M/M$, with constants depending only on $a, D, c_\lambda, C_\lambda$. This is a coefficient event, not a proof of an eigenvalue or teacher-source deterministic equivalent.

Proof. For a row write $Q = MG_{kk} = \sum_j \lambda_j Z_j^2$. Retaining its first K coordinates gives $Q \geq c_\lambda K^{-a}\chi_K^2$. Markov's inequality applied to $\exp(-\chi_K^2/2)$ gives

$$\Pr(\chi_K^2 \leq K/2) \leq e^{K/4}2^{-K/2} \leq e^{-K/16}.$$

For the upper tail, if $\theta = \sum_j \lambda_j$, $v = \sum_j \lambda_j^2$, and $L = \max_j \lambda_j$, then, for $0 < t < (2L)^{-1}$,

$$\log \mathbb{E}e^{t(Q-\theta)} = \frac{1}{2}\sum_j [-\log(1-2t\lambda_j) - 2t\lambda_j] \leq \frac{vt^2}{1-2Lt}.$$

The inequality follows by expanding $-\log(1-u)-u$ into its positive power series and bounding it by $u^2/[2(1-u)]$. Chernoff's inequality with $t = \sqrt{x}/(\sqrt{v}+2L\sqrt{x})$ gives $\Pr(Q > \theta+2\sqrt{vx}+2Lx) \leq e^{-x}$. Since $\theta \leq L_1, v \leq L_2, L \leq C_\lambda$, a union bound over the rows and both tails costs at most $2Me^{-x} = 2M^{-D-1} \leq M^{-D}$. On this event,

$$\min_k G_{kk} \geq \frac{c_\lambda}{2MK^{a-1}}, \qquad \max_k G_{kk} \leq \frac{L_1 + 2\sqrt{L_2 x} + 2C_\lambda x}{M}.$$

Now apply (88) and (95). The event involves only G, so it holds for all precisions at once, without another union bound. □

E.8.5 Optional fixed-width stationary consequences, with full proofs

Throughout this section $G, \eta > 0, B, p$ are fixed and the horizon tends to infinity *before* any width limit. This is a separate consequence, not a finite-window claim at $T \asymp M^a/\eta$.

Proposition E.19 (Exact comparison stability and stationary risk)**.** *Assume $\eta > 0$ and $\eta(1 + 1/B)g_1 < 2$ (so that $h_j > 0$ below; if $\eta(1+1/B)g_1 \geq 2$ then $\rho(A_\beta) \geq a_1^0 \geq 1$). Put $\varepsilon = \eta^2/B$,*

$$h_j = 2 - \eta(1+1/B)g_j, \quad a_j^0 = (1-\eta g_j)^2 + \varepsilon g_j^2, \quad \chi = \frac{\eta}{B}\sum_j \frac{g_j}{h_j}, \quad \zeta_\beta = \frac{\eta}{B}\sum_j \frac{g_j+\beta}{h_j}.$$

For the isotropic comparison at coefficient $\beta \geq 0$, the diagonal homogeneous matrix $A_\beta = \operatorname{diag}(a_j^0) + \varepsilon(g+\beta\mathbf{1})g^\top$ has spectral radius less than one if and only if $\zeta_\beta < 1$. Under this condition its stationary energies and last-iterate premium over the matched exact-arithmetic system are

$$e_{j,\infty}^\beta = \frac{\eta s}{B(1-\zeta_\beta)}\frac{g_j+\beta}{g_j h_j}, \tag{105}$$

$$\Delta_\infty^\beta(\text{last}) = \frac{s\eta\beta}{2B}\frac{\sum_j h_j^{-1}}{(1-\chi)(1-\zeta_\beta)}. \tag{106}$$

Moreover

$$\lim_{T\to\infty} T\Delta_T^\beta(\text{av}) = \frac{s}{2B}\sum_j \frac{2-\eta g_j}{h_j}\left[\frac{\beta}{g_j(1-\zeta_\beta)} + \frac{\zeta_\beta - \chi}{(1-\chi)(1-\zeta_\beta)}\right] =: \mathcal{A}_\beta. \tag{107}$$

If $\zeta_{\beta_+} < 1$, the actual SR system is stable and its stationary last-iterate premium and its limiting T times average premium are bounded between (106) *and* (107)*, respectively, evaluated at β_- and β_+. (The averaged limit exists for the actual system as well: the two-time identity of Corollary E.17 holds for it, $C_r^{\mathrm{SR}} \to C_\infty^{\mathrm{SR}}$ exponentially, and the Cesàro argument below applies verbatim with $e_{j,\infty}^\beta$ replaced by $\phi_j^\top C_\infty^{\mathrm{SR}}\phi_j$, so $\lim_T T\,L^{\mathrm{SR}}(\text{av}, T) = \frac{1}{2\eta}\sum_j(2-\eta g_j)\phi_j^\top C_\infty^{\mathrm{SR}}\phi_j$.)*

Proof. Let $D = \operatorname{diag}(1-a_j^0) = \operatorname{diag}(\eta g_j h_j) \succ 0$ and $u = \varepsilon(g+\beta\mathbf{1}) > 0$. For $w = D^{-1}u > 0$, $g^\top w = \zeta_\beta$ and $A_\beta w = w - (1-\zeta_\beta)u$. The positive-vector characterization of the spectral radius of a nonnegative matrix gives $\rho(A_\beta) < 1$ when $\zeta_\beta < 1$ and $\rho(A_\beta) \geq 1$ when $\zeta_\beta \geq 1$. For completeness the first implication follows in the weighted maximum norm from $A_\beta w \leq rw$ for some $r < 1$; the second follows by iterating $A_\beta w \geq w$ and the spectral-radius formula.

Stationarity in (102) gives $e_j = \eta(s+r)(g_j+\beta)/(Bg_j h_j)$ and $r = \sum_j g_j e_j = \zeta_\beta(s+r)$. Solving proves (105) and $r = s\zeta_\beta/(1-\zeta_\beta)$. Subtract its exact-arithmetic value at $\beta = 0$ and divide by two to obtain (106).

The comparison matrix map is positive and its homogeneous powers tend to zero exponentially in finite dimension: on PSD inputs its trace is governed by the stable diagonal matrix A_β, and every symmetric input is a difference of PSD inputs. Consequently C_t converges exponentially to its unique stationary second moment. Also $h_j > 0$ implies $|1 - \eta g_j| < 1$. Using (103), a finite-dimensional geometric-series bound and Cesaro convergence give

$$\lim_{T\to\infty} TL^\beta(\mathrm{av}, T) = \frac{1}{2}\sum_j g_j e^\beta_{j,\infty} \frac{2 - \eta g_j}{\eta g_j}.$$

Indeed the summand at time s is multiplied by $1 + 2\sum_{\ell=1}^{T-s}(1 - \eta g_j)^\ell$; this is uniformly bounded, converges away from the final boundary, and the transient initial moments have vanishing Cesaro contribution. Substituting (105) and subtracting the exact-arithmetic value gives (107).

The actual homogeneous map is bounded above on the PSD cone by the upper comparison homogeneous map. Its powers therefore decay as well, by induction and the preceding PSD trace argument. It has a unique stationary moment and the same mean propagator H. Apply the finite-horizon output comparison and take the indicated limits. The additive A_S cancels from every premium. □

Corollary E.20 (What width dependence is already proved)**.** *Suppose $h_j \geq h_* > 0$ and $\zeta_{\beta_+} \leq 1 - \kappa$ for a specified $\kappa > 0$. For the actual algorithm,*

$$\frac{s\eta M \beta_-}{4B} \leq \Delta^{\mathrm{SR}}_\infty(\mathrm{last}) \leq \frac{s\eta M \beta_+}{2Bh_*\kappa^2}. \tag{108}$$

On the event of Theorem E.18, this gives a stationary last-iterate premium between constant multiples of $(s\eta/B)4^{-p}(\log M)^{1-a}$ and $(s\eta/B)4^{-p}\log M$. Writing $J_G = \sum_j g_j^{-1}$, the average satisfies

$$\frac{s\beta_-}{2B} J_G \leq \lim_{T\to\infty} T\Delta^{\mathrm{SR}}_T(\mathrm{av}) \leq \frac{s\beta_+}{B}\left(\frac{1}{\kappa h_*} + \frac{2}{\kappa^2 h_*^2}\right) J_G. \tag{109}$$

No assertion about the random width order of J_G is needed for this result.

Proof. For positive η, $h_j \leq 2$. In (106) the lower denominator factors are at most one; for the upper bound both are at least κ and $\sum_j h_j^{-1} \leq M/h_*$. This proves (108), and (104) gives the stated logs. In (107), $1 \leq (2 - \eta g_j)/h_j \leq 2/h_*$. The first term is therefore between $s\beta J_G/(2B)$ and $s\beta J_G/(B\kappa h_*)$. The second is nonnegative and at most $s\eta\beta M^2/(B^2\kappa^2 h_*^2)$. Cauchy–Schwarz gives $M^2 \leq (\mathrm{tr}\, G) J_G$, while $\chi \geq \eta\,\mathrm{tr}\, G/(2B)$ and $\chi < 1$, so this second term is at most $2s\beta J_G/(B\kappa^2 h_*^2)$. Use the two comparison coefficients to conclude. □

E.8.6 Proof of the law (85) of Theorem E.12

Project (86) on the eigenbasis of G: with $e_j^\pm(t) = \phi_j^\top C_t^\pm \phi_j$,

$$e_j^\pm(t+1) = (1 - \eta g_j)^2 e_j^\pm(t) + \frac{\eta^2}{B}\Big[g_j^2 e_j^\pm(t) + (g_j + \beta_\pm)\big(\sigma^2_{\mathrm{eff}} + \sum_l g_l e_l^\pm(t)\big)\Big],$$

which is (10) with $\iota_{\mathrm{bw}} = \beta_\pm$, $\iota_{\mathrm{fw}} = \bar\alpha_{\mathrm{fw}} = \bar\alpha_{\mathrm{bw}} = \gamma_{\mathrm{fwbw}} = 0$; the off-diagonal entries of $C_t^\pm$ in the eigenbasis do not enter, and the outputs are the same positive functionals as in Corollary B.8. The backward first-order response of Proposition E.4 is linear in ι_{bw} ((64) holds for any $\iota_{\mathrm{bw}} > 0$ with the same geometric-history means $\bar R_j(t) \asymp r_t$, whose proof uses only the exact-arithmetic solution), so under Assumption 3 for the spectrum and target of the realised feature map and $\beta_0 \leq 1 - \kappa_0$, $I^{\mathrm{out}}_\pm(t) \asymp \varrho\beta_\pm r_t k_t$ on the regime for both outputs. The lower bound $\Delta^{\mathrm{out}}_- \geq I^{\mathrm{out}}_-$ is (60) (positivity alone), and the sandwich of Theorem E.16 gives $\Delta^{\mathrm{out}} \geq \Delta^{\mathrm{out}}_-$. For the upper bound, Proposition E.5 applied to the upper system gives $\Delta^{\mathrm{out}}_+ \leq I^{\mathrm{out}}_+/(1 - \gamma_*)$ when its feedback gain $\Gamma^+(T) \leq \gamma_*$ (Lemma E.6 with $\beta = \beta_+$ and $f = m = 0$ gives the sufficient condition $C_{\mathrm{sg}}\varrho[\theta_G + \beta_+ k_T] \leq \gamma_*$), and the sandwich gives $\Delta^{\mathrm{out}} \leq \Delta^{\mathrm{out}}_+$. Inserting the coefficient bounds of Theorem E.18 gives the logarithmic bracket. □

Remark E.21 (scope of the results of this appendix). Four things are not proved here and nothing above uses them. A cooldown with forward rounding active: a frozen-source calculation gives a principal decaying component $\varrho_{\mathrm{peak}} \frac{q_{\mathrm{fw}} U}{M\sqrt{\tau_c}} \sum_j g_j^{1/2}$ for $\tau_c \geq M^a$, but the evolving source U^0_s with

both copies active through an arbitrarily long cooldown is not controlled. A timing theorem with the full feedback, beyond the first-order response of Proposition F.5. A cooldown gain that is uniform in the duration with a constant bounded as $\sigma^2 \downarrow 0$; the sketch of the duration-uniform gain in the proof of Proposition E.10 uses the source floor. The constants C_{sg}, $C_{\mathrm{nat}}(\sigma)$, $C_{\mathrm{lin}}(\sigma)$ are not tracked, so the explicit sufficient conditions are order-level, and the constant-free hypotheses are the exact tests (68) and (81).

F Full statements and proofs for Section 4

Section 4 states its rules informally, and we now present the full statements and the proofs. Every statement of this appendix is a consequence of Theorem 4 (for the algorithm through Conjecture 2, and for the backward rounding terms through Theorem E.12), read at the compute-optimal horizon or along a schedule. §F.1 proves the bit requirements behind Table 1 (Corollary F.1); §F.2 proves the regime map, Proposition 5; §F.3 derives the response weights and the allocation (15) of §4.2 (Proposition F.6) and its closed forms at a constant rate (Corollary F.7); §F.4 states and derives the compute scaling law, Theorem F.8, in full (Theorem F.8); §F.5 derives the tolerance under a wall-clock budget quoted in §4.3 (Corollary F.9); §F.6 states Proposition F.10 in full (Proposition F.10) with its lemmas; §F.7 gives the precision schedule of §4.3 as Algorithm 1, and §F.8 the joint optimisation of learning-rate and precision schedules.

F.1 Bit requirements

The main text states the bit requirements in one sentence and Table 1 (§4.1: (13) holds if $p \geq p^{\vartheta} + c_1$ for every requirement of the output rule and fails if $p \leq p^{\vartheta} - c_2$ for one of them); Corollary F.1 is the full statement, and Corollary F.2 and Proposition F.4 below are the steps of its proof.

Corollary F.1 (precision requirements: where the bits are needed)**.** *Under Theorem 4, the penalty stays below a fraction ϑ of the exact-arithmetic risk, $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) \leq \vartheta\, \mathcal{E}_0^{\mathrm{out}}(t)$, iff (up to constant factors at the threshold) each term of* (12) *is at most $\vartheta\, \mathcal{E}_0^{\mathrm{out}}(t)$. In mantissa bits: write each term as $q_r K_r(t)$ (the interaction term as $q_{\mathrm{fw}} q_{\mathrm{bw}} K_{\times}(t)$) and put $p_r^{\vartheta} = \frac{1}{2}\log_2\big(K_r/(\vartheta\, \mathcal{E}_0^{\mathrm{out}})\big)$ (the joint requirement being on $p_{\mathrm{fw}} + p_{\mathrm{bw}}$); there are constants $c_1, c_2 > 0$, depending only on those of Assumption 3 (and, for the cooled row, on σ^2 and ϕ, under the hypotheses of Lemma E.11), such that $p_r \geq p_r^{\vartheta} + c_1$ for every term implies $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) \leq \vartheta\, \mathcal{E}_0^{\mathrm{out}}(t)$, and $p_r \leq p_r^{\vartheta} - c_2$ for one term implies $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) > \vartheta\, \mathcal{E}_0^{\mathrm{out}}(t)$. At the end of a compute-optimal run ($\tau_T = M^a$, $k_T = M$) the requirements are*

$$
\begin{aligned}
&\textit{last iterate:} && q_{\mathrm{fw}} \lesssim \frac{\vartheta\,(M^{1-b} + \varrho\sigma^2)}{\varrho M^{a-b}}, && q_{\mathrm{bw}} \lesssim \frac{\vartheta\,(M^{1-b} + \varrho\sigma^2)}{\varrho\,(\sigma^2 + M^{1-b})};\\
&\textit{average:} && q_{\mathrm{fw}} \lesssim \frac{\vartheta}{\varrho}, && q_{\mathrm{bw}} \lesssim \frac{\vartheta}{\varrho}\,\frac{M^{1-b}}{\sigma^2 + M^{1-b}},\\
& && && q_{\mathrm{fw}} q_{\mathrm{bw}} \lesssim \frac{\vartheta}{\varrho} M^{1-a} \ \ (\textit{joint});\\
&\textit{cooled last iterate:} && q_{\mathrm{bw}} \lesssim \frac{\vartheta\, M^{1-b}}{\varrho\sigma^2\, c(\phi,\zeta)}; && \textit{input, shared copy: } 16^{-p_x} \lesssim \vartheta M^{1-b},
\end{aligned}
\tag{110}
$$

so, at fixed controls and label noise and with the noise ratio $\mathsf{n} = \varrho\sigma^2 M^{b-1}$, the bits grow with the width like $\frac{a-1}{2}\log_2 M$ for the joint requirement of the average and for the forward copy of the last iterate while $\mathsf{n} \lesssim 1$ (like $\frac{a-b}{2}\log_2 M$ once $\mathsf{n} \gg 1$), like $\frac{b-1}{2}\log_2 M$ for the backward copy of the cooled last iterate and, once $\sigma^2 \gg M^{1-b}$, of the average, and like $\frac{b-1}{4}\log_2 M$ for the input, whereas the forward bits of the average and the backward bits of the last iterate stay bounded. Table 1 lists the bits needed and their values at $M = 1024$. Against training time at fixed width the slopes $p_{\mathrm{bw}} \geq \frac{b}{2a}\log_2 \tau_t$, $p_{\mathrm{fw}} \geq \frac{a-b+1}{2a}\log_2 \tau_t$, $p_{\mathrm{fw}} + p_{\mathrm{bw}} \geq \frac{a+1}{2a}\log_2 \tau_t$ are sufficient but not necessary.

Proof of Corollary F.1. The equivalence and the bit form are Corollary F.2. At $\tau_T = M^a$ we have $k_T = M$ and $r_T = \sigma^2 + M^{1-b}$, and Lemma E.1 gives $\mathcal{E}_0^{\mathrm{last}}(T) \asymp M^{1-b} + \varrho\sigma^2$ and $\mathcal{E}_0^{\mathrm{avg}}(T) \asymp M^{1-b} + \varrho\sigma^2 M^{1-a} \asymp M^{1-b}$, the last step because $\varrho\sigma^2 M^{b-a} \leq \eta_{\max}\Sigma_{\max}$ on the window. The terms of (12) at $k_T = M$ are $\varrho q_{\mathrm{fw}} M^{a-b}$ and $\varrho q_{\mathrm{bw}}(\sigma^2 + M^{1-b})$ for the last iterate, and $\varrho q_{\mathrm{fw}} M^{1-b}$, $\varrho q_{\mathrm{bw}}(\sigma^2 + M^{1-b})$ and $\varrho q_{\mathrm{fw}} q_{\mathrm{bw}} M^{a-b}$ for the average; dividing each by ϑ times

the corresponding exact-arithmetic risk gives the last-iterate and average rows of (110). The cooled row is the law $\Delta_{\mathrm{DE}}^{\mathrm{cooled}}(T) \asymp_\sigma \varrho_{\mathrm{peak}} q_{\mathrm{bw}} \sigma^2 c(\phi,\zeta)$ of (14) (Proposition E.10(ii)) against the reference scale M^{1-b}; it is the requirement $\Delta_{\mathrm{DE}}^{\mathrm{cooled}}(T) \le \vartheta\, \mathcal{E}_0^{\mathrm{cooled}}(T)$ because $\mathcal{E}_0^{\mathrm{cooled}}(T) \asymp_\sigma M^{1-b}$ on the range $\zeta \mathsf{n}_{\mathrm{peak}} \lesssim 1$, $\mathsf{n}_{\mathrm{peak}} D_a(M) \lesssim 1$ of Lemma E.11. The input row is Proposition C.1(ii): the shared rounded input leaves a bias floor $\asymp 16^{-p_x}\|v_*^S\|_G^2$ with $\|v_*^S\|_G^2 \asymp 1$. *Bits.* With $q = 4^{-p}$ and $\mathsf{n} = \varrho\sigma^2 M^{b-1}$, the rows of (110) read $p \ge p^\vartheta$ with the entries of Table 1: for instance the forward row of the last iterate is $4^{-p_{\mathrm{fw}}} \le \vartheta(M^{1-b} + \varrho\sigma^2)/(\varrho M^{a-b}) = \vartheta(1+\mathsf{n})/(\varrho M^{a-1})$, and the backward rows use $(\sigma^2 + M^{1-b})/M^{1-b} = (\mathsf{n}+\varrho)/\varrho$. *Two-sided form.* By Theorem 4 and Lemma E.1 there are constants $0 < c_\Delta \le C_\Delta$ and $0 < c_E \le C_E$ with $c_\Delta \sum_i q_i K_i \le \Delta_{\mathrm{DE}}^{\mathrm{out}} \le C_\Delta \sum_i q_i K_i$ and $c_E \bar{\mathcal{E}} \le \mathcal{E}_0^{\mathrm{out}} \le C_E \bar{\mathcal{E}}$, where the sums run over the at most three terms, K_i and $\bar{\mathcal{E}}$ are the unit-prefactor forms used in Table 1, and q_i is q_{fw}, q_{bw} or $q_{\mathrm{fw}} q_{\mathrm{bw}}$; for the cooled row, Proposition E.10 and Lemma E.11 give the same with constants depending on σ^2 and ϕ. If every $q_i K_i \le 4^{-c_1}\vartheta\bar{\mathcal{E}}$ with $3C_\Delta 4^{-c_1} \le c_E$, then $\Delta_{\mathrm{DE}}^{\mathrm{out}} \le 3C_\Delta 4^{-c_1}\vartheta\bar{\mathcal{E}} \le \vartheta\, \mathcal{E}_0^{\mathrm{out}}$. If one $q_i K_i \ge 4^{c_2}\vartheta\bar{\mathcal{E}}$ with $c_\Delta 4^{c_2} > C_E$, then $\Delta_{\mathrm{DE}}^{\mathrm{out}} \ge c_\Delta q_i K_i > C_E\vartheta\bar{\mathcal{E}} \ge \vartheta\, \mathcal{E}_0^{\mathrm{out}}$. In bits these are $p_i \ge p_i^\vartheta + c_1$ for all i and $p_i \le p_i^\vartheta - c_2$ for one i. *Growth.* At fixed ϱ, σ^2 and ϑ, $\mathsf{n} \propto M^{b-1}$; the entries of Table 1 then grow as stated: $\frac{1}{2}\log_2(\varrho M^{a-1}/(\vartheta(1+\mathsf{n})))$ grows like $\frac{a-1}{2}\log_2 M$ while $\mathsf{n} \lesssim 1$ and like $\frac{a-b}{2}\log_2 M$ once $\mathsf{n} \gg 1$; $\frac{1}{2}\log_2((\mathsf{n}+\varrho)/(\vartheta(1+\mathsf{n}))) \le \frac{1}{2}\log_2((1+\varrho)/\vartheta)$ is bounded; $\frac{1}{2}\log_2((\mathsf{n}+\varrho)/\vartheta)$ grows like $\frac{b-1}{2}\log_2 M$ once $\mathsf{n} \gg \varrho$, i.e. $\sigma^2 \gg M^{1-b}$; the cooled, joint and input entries grow like $\frac{b-1}{2}$, $\frac{a-1}{2}$ and $\frac{b-1}{4}$ times $\log_2 M$; and $\frac{1}{2}\log_2(\varrho/\vartheta)$ is constant. The slopes in τ_t are Proposition F.4. □

In Table 1, "joint" is the interaction term, whose requirement is on $p_{\mathrm{fw}} + p_{\mathrm{bw}}$, and "cooled" the law (14), with n at the peak rate; a negative entry means that no bit is needed. The column "theory" evaluates p^ϑ at $M = 1024$ with the constant of the rounding variance, $m_p(g) \approx 0.09 \cdot 4^{-p} g$, and "measured" is the number of bits at which the penalty of the exact recursion equals ϑ times the exact-arithmetic risk (Appendix G); the two agree within 0.4 bit for the last iterate, and the formula is conservative by 1.4 to 1.9 bits for the average. In the cooled row, the SGD noise injected before the cooldown is reduced by the factor k_c^{1-a}, $k_c = \min(M, \tau_c^{1/a})$, whereas the backward rounding noise injected before it is retained by the factor $\Upsilon(\phi) = (1+\phi)^{1/a} - \phi^{1/a}$ of Proposition E.10.

Corollary F.2 (precision budgets)**.** *Under Theorem 4, $\Delta_{\mathrm{DE}}^{\mathrm{out}}(t) \le \vartheta\, \mathcal{E}_0^{\mathrm{out}}(t)$ holds, up to constant factors at the threshold, iff each row of* (12) *is at most ϑ times the exact-arithmetic scale $\mathcal{E}_0^{\mathrm{out}}(t) \asymp k_t^{1-b} + \varrho\sigma^2$ (last) or $k_t^{1-b} + \varrho\sigma^2 k_t/\tau_t$ (average); in bits, $p \ge \frac{1}{2}\log_2(1/\mathit{bound}) + O(1)$. At $\tau_T \asymp M^a$ the averaged budgets read $q_{\mathrm{fw}} \lesssim \vartheta/\varrho$, $q_{\mathrm{bw}} \lesssim (\vartheta/\varrho)M^{1-b}/(\sigma^2 + M^{1-b})$ and $q_{\mathrm{fw}} q_{\mathrm{bw}} \lesssim (\vartheta/\varrho)M^{1-a}$, the joint one binding exactly when $q_{\mathrm{fw}} \operatorname{tr}\Sigma_d M^{a-b} \gtrsim \sigma^2 + M^{1-b}$; for the last iterate the joint row never binds. The slopes in $\log_2 \tau_t$ (Appendix F.1) are $\frac{1}{2a}$ (averaged forward), $\frac{a+1}{2a}$ (joint), $\frac{1}{2a}$ or $\frac{b}{2a}$ (backward, according as $\sigma^2 \lesssim k_t^{1-b}$ or not) and $\frac{1}{2}$ (last-iterate forward, bias regime).*

Corollary F.3 (backward rounding is hidden by SGD noise while the rate is constant)**.** *At constant step the backward first-order response of the last iterate is $\frac{\varrho\iota_{\mathrm{bw}}}{2}\sum_j \frac{1-r_j^{2t}}{2-\eta g_j}\bar{R}_j(t)$ ((64)) and the SGD-noise term of the exact-arithmetic risk is $\frac{\varrho}{2}\sum_j \frac{g_j(1-r_j^{2t})}{2-\eta g_j}\bar{R}_j(t)$, with the same residual means; their ratio is at most $\iota_{\mathrm{bw}} k_t/\theta_G \approx 0.09\, q_{\mathrm{bw}} k_t/M \le 0.09 \cdot 4^{-p_{\mathrm{bw}}}$. Backward SR at p bits thus costs at most $\approx 0.1 \cdot 4^{-p}$ of what SGD noise costs (0.15% at FP8-E4M3) while that noise is present, and becomes visible when it is removed by averaging (the ratio becomes $\approx 0.09\, q_{\mathrm{bw}} \operatorname{tr}\Sigma_d k_t^a/M$, growing with τ_t) or by a cooldown (Proposition E.10).*

Proof. Both sums are (64) and the SGD-noise part of the exact-arithmetic risk with the same residual means $\bar{R}_j(t)$; the weights differ by ι_{bw} against g_j, and $\sum_j (1-r_j^{2t})/(2-\eta g_j) \lesssim k_t$ while $\sum_j g_j(1-r_j^{2t})/(2-\eta g_j) \gtrsim \theta_G$ once $\tau_t \ge \tau_*$. For the average the isotropic filter Q_{-1} replaces k_t by $\asymp \operatorname{tr}\Sigma_d k_t^a$ (Lemma E.3). □

Proof of Corollary F.2. $\mathcal{E}_0^{\mathrm{out}}(t) \asymp \mathcal{L}_t^{\mathrm{out}}$ by Lemma E.1. Dividing each row of (12) by the exact-arithmetic scale, a bound on a sum of nonnegative terms implies it for each term, and bounds by a third of the target on each term imply it for the sum: the budgets in both directions up to constants; $q = 4^{-p}$ gives the bit form. The last-iterate joint row is implied by the forward one, its premium row being $q_{\mathrm{bw}} k_t/M \le 1$ times the forward row. The slopes are read off from $\mathcal{L}_t^{\mathrm{avg}} \asymp$

k_t^{1-b} on the window ($\varrho\sigma^2 k_t^{b-a} \le \eta_{\max}\Sigma_{\max}$) and from the three last-iterate regimes ($\mathcal{L}_t^{\text{last}}, r_t$) $\asymp$ (k^{1-b}, k^{1-b}), (k^{1-b}, σ^2), ($\varrho\sigma^2, \sigma^2$), in which $q_{\text{fw}} \lesssim (M/\varrho)k^{-a}, (M/\varrho)k^{-a}, M\sigma^2 k^{-(a-b+1)}$ and $q_{\text{bw}} \lesssim (M/\varrho)k^{-1}, M(\varrho\sigma^2)^{-1}k^{-b}, M/k$ (the last ≥ 1). At $\tau_T \asymp M^a$, $\mathcal{L}_T^{\text{avg}} = M^{1-b} + \varrho\sigma^2 M^{1-a} \asymp M^{1-b}$, and the ratio of the backward bound on q_{bw} to the bound $\vartheta M^{1-a}/(\varrho q_{\text{fw}})$ from the joint budget is $q_{\text{fw}}\theta M^{a-b}/(\sigma^2 + M^{1-b})$ once $\iota_{\text{bw}} \asymp q_{\text{bw}}\theta/M$, $m \asymp q_{\text{fw}}q_{\text{bw}}\theta^2/M^2$ are restored, so the joint budget binds iff this is $\gtrsim 1$. (Under the natural gain alone, Proposition E.9 adds $q_{\text{fw}}q_{\text{bw}} \lesssim \vartheta M\mathcal{L}_t^{\text{last}}/(\varrho^2 r_t \tau_t)$.) □

Proposition F.4 (the three slopes are sufficient only)**.** *For fixed $\sigma^2 > 0$ under Theorem 4 there are additive constants (depending on σ^{-2}) such that $p_{\text{bw}} \ge \frac{b}{2a}\log_2\tau_t + C_{\text{bw}}$, $p_{\text{fw}} \ge \frac{a-b+1}{2a}\log_2\tau_t + C_{\text{fw}}$, $p_{\text{fw}} + p_{\text{bw}} \ge \frac{a+1}{2a}\log_2\tau_t + C_{\text{fwbw}}$ imply $\Delta_{\text{DE}}^{\text{last}}(t) \le \vartheta\mathcal{E}_0^{\text{last}}(t)$ on the window; they are not necessary. Indeed they give $q_{\text{bw}} \le c_b k_t^{-b}$, $q_{\text{fw}} \le c_a k_t^{-(a-b+1)}$, $q_{\text{fw}}q_{\text{bw}} \le c_{ab}k_t^{-a-1}$ with constants as small as desired, and by $\mathcal{L}_t^{\text{last}} \ge \max(\varrho\sigma^2, k_t^{1-b})$ the three premium ratios are $\le c_a/(M\sigma^2)$, $\varrho c_b r_t/M$, $\varrho c_{ab}/M^2$: of order M^{-1}, M^{-2} instead of one, an over-provision of $\frac{1}{2}\log_2 M$ bits per location ($\log_2 M$ for the sum).*

F.2 The regime map: derivation of the ratios behind §4.1

Proof of Proposition 5. The paragraphs *Last iterate* and *Average* below prove parts (i) and (ii); every ratio is a ratio of the terms of (12) up to the constants of Theorem 4. Write $D = M^{1-b}$, $r = \sigma^2 + D$ (so $r_T = r$ at $k_T = M$) and $\mathsf{n} = \varrho\sigma^2/D$, the noise ratio of §2 and Proposition 5. Proposition 5 states the last-iterate and average comparisons below as its parts (i) and (ii); the cooled last iterate is the last paragraph of this subsection, where its threshold is stated but not derived. At $\tau_T = M^a$ and equal bits $q_{\text{fw}} = q_{\text{bw}} = q = 4^{-p}$ the first-insertion entries of Proposition E.4 are

$$I_{\text{fw}}^{\text{last}} \asymp \varrho\, q\, M^{a-b}, \qquad I_{\text{bw}}^{\text{last}} \asymp I_{\text{bw}}^{\text{avg}} \asymp \varrho\, q\, r, \qquad I_{\text{fw}}^{\text{avg}} \asymp \varrho\, q\, D, \qquad I_{\text{fwbw}}^{\text{avg}} \asymp \varrho\, q^2 M^{a-b}, \quad (111)$$

with σ^2-free constants, and $\mathcal{L}_T^{\text{last}} \asymp D(1+\mathsf{n})$, $\mathcal{L}_T^{\text{avg}} \asymp D$ (Lemma E.1; $\varrho\sigma^2 M^{b-a} \le \eta_{\max}\Sigma_{\max}$). Every ratio below is a ratio of entries up to the constants of Theorem 4; the total penalty of any set of active locations is within $(1-\gamma_*)^{-1}$ of the sum of its entries under that system's gain condition (Proposition E.5).

Last iterate. $I_{\text{fw}}^{\text{last}}/I_{\text{bw}}^{\text{last}} \asymp M^{a-b}/(\sigma^2 + D)$: ϱ and q cancel, so the ratio is the same for every batch size and precision, and it exceeds 1 for all $M \ge M_0$, a constant-dependent width ($M_0 = (\Sigma_{\max} + 1)^{1/(a-b)}$ at unit prefactor) — the constant-step last iterate is forward-limited for every admissible n. For backward rounding alone, $\Delta_{\text{DE}}^{\text{last}} \le I_{\text{bw}}^{\text{last}}/(1-\gamma_*) \asymp q_{\text{bw}}\,\text{tr}\,\Sigma_d\,(1 + D/\sigma^2)\,\varrho\sigma^2/(1-\gamma_*) \le Cq_{\text{bw}}\,\varrho\sigma^2/(1-\gamma_*)$, since $\sigma^2 = \sigma_{\text{label}}^2 + 2\bar{A}_S \ge 2\bar{A}_S \gtrsim D$: at most a constant times $4^{-p_{\text{bw}}}$ of the SGD-noise floor. The marginal backward penalty of a two-location run adds the last-iterate interaction entry, at most q_{bw} times the forward entry, and obeys the same bound iff $q_{\text{fw}} \lesssim \sigma^2 M^{b-a}$.

On the constant of Corollary F.3. Its ratio "at most $\iota_{\text{bw}}k_t/\theta_G$" is an order statement. The backward entry averages the history $\mathfrak{R}_{\text{bw}}(s) = \mathsf{R}_s^0 + (2\bar{\alpha}_{\text{bw}}/\iota_{\text{bw}})\mathsf{S}_s^0$, while the SGD-noise term of the exact-arithmetic risk averages $\mathsf{R}_s^0 + g_j e_j^0(s)$ (the injection $\varrho\eta[g_j\mathsf{R}_s^0 + g_j^2 e_j^0(s)]$ in the proof of Lemma E.1); the two agree only up to factors in $[\frac{1}{2}, 1+C']$ ($g_j e_j^0 \le \mathsf{R}_s^0 - \sigma^2$; $\mathsf{S}_s^0 \le g_1\mathsf{R}_s^0/M$ and $\bar{\alpha}_{\text{bw}}/\iota_{\text{bw}} \asymp M/\,\text{tr}\,\Sigma_d$). The two spectral sums are bounded with $\lesssim$ and $\gtrsim$; for ηg_j small their ratio is $\approx (1 + \frac{2}{a-1})\,k_t/\theta_G$. The 0.09 is the coefficient's (Lemma B.1) and is exact for ι_{bw}; the ratio of the backward entry to the exact-arithmetic SGD-noise term is $\asymp q_{\text{bw}}(\text{tr}\,\Sigma_d/\theta_G)\,k_t/M$ with an order-one constant (0.3 on the reduced dynamics at $M = 128$).

Average. From (111),

$$\frac{I_{\text{bw}}^{\text{avg}}}{I_{\text{fw}}^{\text{avg}}} \asymp 1 + \frac{\sigma^2}{D}, \qquad \frac{I_{\text{fwbw}}^{\text{avg}}}{I_{\text{bw}}^{\text{avg}}} \asymp 4^{-p}\,\frac{\text{tr}\,\Sigma_d M^{a-b}}{\sigma^2 + D}, \qquad \frac{I_{\text{fwbw}}^{\text{avg}}}{I_{\text{fw}}^{\text{avg}}} \asymp 4^{-p}\,\text{tr}\,\Sigma_d M^{a-1}. \quad (112)$$

The forward entry never binds; forward and backward are comparable iff $\sigma^2 \asymp D$ (label noise at most of the order of the approximation error). The interaction entry is quadratic in q, so one factor 4^{-p} survives against the linear entries, and the backward entry binds iff $\sigma^2 + D \gtrsim 4^{-p}\,\text{tr}\,\Sigma_d M^{a-b}$ — the condition of Corollary F.2. This is a condition on σ^2: all three entries carry the prefactor ϱ, so their ratios do not depend on the batch size, and the noise ratio n does not classify them. Indeed, for $a = 2$, $b = \frac{3}{2}$, $g_j = j^{-2}$, $w_j = j^{1/2}$, $\sigma^2 = 1$, $\eta = \frac{1}{2}$, $M = 2^{16m}$ with $B = M^{3/8}$ and $p = 3m$ one has

$\mathsf{n} = \frac{1}{2}M^{1/8} \to \infty$ while $I^{\mathrm{avg}}_{\mathrm{fwbw}}/I^{\mathrm{avg}}_{\mathrm{bw}} \asymp M^{1/8} \to \infty$; with $B = M$ and $p = 8m$, $\mathsf{n} = \frac{1}{2}M^{-1/2} \to 0$ while $I^{\mathrm{avg}}_{\mathrm{fwbw}}/I^{\mathrm{avg}}_{\mathrm{bw}} \asymp M^{-1/2} \to 0$; in both families $\beta_0 \to 0$ and the left side of (69) is $O(M^{-1/8})$, resp. $O(M^{-1})$. What does move with the batch is the *budget* comparison of Corollary F.2: at equal bits and tolerance ϑ the joint budget $q \lesssim (\vartheta/\varrho)^{1/2}M^{(1-a)/2}$ is tighter than the backward budget $q \lesssim \vartheta/(\varrho + \mathsf{n})$ iff $\mathsf{n} < (\vartheta\varrho M^{a-1})^{1/2} - \varrho$ at unit prefactor (at $\vartheta = 0.1$, $\varrho = 1/64$, $M = 1024$: $\mathsf{n}_\times \approx 1.25$ against $\mathsf{n} = 0.05$, the joint budget of Table 1; at $\varrho = 1/8$, $\sigma^2 = 1$: $\mathsf{n}_\times \approx 3.45$ against $\mathsf{n} = 4$).

Cooled last iterate. With $\mathsf{n}_{\mathrm{peak}} = \varrho_{\mathrm{peak}}\sigma^2/D$ and D as the reference scale (the bias term at $k = M$; the exact-arithmetic cooled risk is not stated here), Proposition E.10(ii) gives $\Delta^{\mathrm{last}}_{\mathrm{DE}}(T)/D \asymp_\sigma 4^{-p_{\mathrm{bw}}}\mathsf{n}_{\mathrm{peak}}\, c(\phi, \zeta)$: the cooled backward penalty stays below a fraction ϑ of D iff $4^{-p_{\mathrm{bw}}}\mathsf{n}_{\mathrm{peak}}c(\phi, \zeta) \lesssim \vartheta$ (the cooled row of Table 1). Its comparison with a forward penalty after the cooldown is open (Remark E.21); the backward-only cooled law transfers to the algorithm by the sandwich of Theorem E.12, and Conjecture 2 (constant controls, last iterate and average) does not cover the cooled output.

Numerical illustration. Reduced dynamics at $M = 128$, $d = 512$, $\eta = 0.5/g_1$, $\tau_T = M^2$, single-location and inclusion–exclusion premiums as in Appendix G: the forward-over-backward ratio of the last iterate is 80.8 at $p = 2$ and at $p = 4$ ($B = 32$, $\sigma^2_{\mathrm{label}} = 0.1$), 68.6 at $B = 4$, 17.7 at $\sigma^2_{\mathrm{label}} = 1$; the interaction-over-backward ratio of the average is 0.154 at $p = 2$ and 0.0097 at $p = 4$ (a factor 15.9 for 4^2), 0.160 at $B = 4$; the backward-over-forward ratio of the average is 2.39 and 9.82 at $\sigma^2_{\mathrm{label}} = 0.1$ and 1. The pairs quoted in §4.1 differ in σ^2 as well as in B; the ratios above move through σ^2 only.

The last iterate requires enough forward precision, and whatever removes the SGD noise (a cooldown, an average, an EMA) enough backward precision, so that rounding at that location does not become the dominant source of error: aligned forward noise is re-contracted during the decay, backward noise sits on unlearned modes, so single-location ablations of a constant-rate run find backward rounding harmless and are wrong for the cooled model. An interior optimum of the training time at fixed precision would need $\varrho q_{\mathrm{bw}}\sigma^2_{\mathrm{eff}}M^{b-1} \gtrsim 6$, a condition we state without derivation: more steps never hurt.

At FP8-E4M3 ($p = 3$) the backward penalty of (i) is 0.15% of the SGD noise, and the ratio in (i) contains neither the batch size nor the precision; for the average the floor is removed, and the label noise, not the batch size, decides which term dominates.

F.3 When: the response weights, the allocation (15) and the closed-form schedules

Sensitivity to the injection time.

Proposition F.5 (sensitivity of the output to noise injected at step s)**.** *Propagated through the contraction alone, a unit isotropic noise impulse $M^{-1}\mathbf{1}$ injected at step s contributes to the average over iterates $1, \dots, T$ an amount $\asymp \frac{\tau_T^{1/a-1}}{M}(\frac{T-s}{T})^{1+1/a}$, decreasing in s, and to the cooled last iterate of Proposition E.10 an amount $\asymp \max(k_{T_1-s}, k_c)^{1-a}/M$ with $k_c = \min(M, \tau_c^{1/a})$, increasing in $s < T_1$ (for $\tau_c \le M^a$): late noise is least harmful for the average, early noise for a cooled last iterate.*

By (59), $W_{\mathrm{avg}}(s) = \frac{1}{2T^2}\sum_j g_j \frac{(1-r_j^{T-s})^2}{M\eta^2 g_j^2} = \frac{\mathsf{Q}_{-1}(T-s)}{2M\tau_T^2}$, with $\mathsf{Q}_{-1}(h) \asymp (\eta h)^{1+\varpi}$ on the window and $\asymp (\eta h)^2\theta_G$ below it (Lemma E.3); each filter $(1 - r_j^{T-s})^2$ decreases in s. $W_{\mathrm{last}}(s) = \frac{1}{2M}\sum_j g_j(1 - \eta_{\mathrm{peak}}g_j)^{2(T_1-1-s)}\pi_j(T, T_1)^2$ is the impulse propagated by its actual contractions; Lemma E.2(i) and the S_1 window at reverse time $u_s = \eta_{\mathrm{peak}}(T_1 - 1 - s) + \tau_c$ give $\asymp (1+u_s)^{\varpi-1}/M$, and $\max(1, x, y) \le 1 + x + y \le 3\max(1, x, y)$ with $x = (\eta_{\mathrm{peak}}(T_1 - 1 - s))^{\varpi}$, $y = \tau_c^{\varpi}$ gives $\asymp \max\{1, x, y\}^{1-a}/M$, the maximum-width form away from $s = T_1 - 1$; removing stable-phase contraction factors increases every summand, so it increases in s. The weight of an injection through the full resolvent $\prod_s(\mathsf{D}_s + \varrho\eta_s\mathsf{H})$ is a different object, and a variable-step averaging filter is not a geometric sum (hence the constant-step averaging phase).

Isotropic noise is remembered longer than aligned noise, since $S_1(\chi) \asymp \chi^{1/a-1}$ decays more slowly than $S_2(\chi) \asymp \chi^{1/a-2}$; the weights (113) are computed in $O(MT)$ operations from the exact-arithmetic trajectory. In the closed forms of Corollary F.7, a step whose weight is four times the geometric mean gets one more bit, the gain over a constant schedule is large when the weights spread over many orders of magnitude, the forward precision of an average peaks at $s/T = (a-b+1)/(a-b+2)$ and is nearly constant over most of the run, and the backward allocation depends on the step only through the remaining time.

The allocation (15) **in full, and its proof.** §4.2 defines the response weights in one sentence and states the allocation in (15); Proposition F.6 is the full statement.

Proposition F.6 (when: response weights and the optimal allocation of a bit budget (first-order response))**.** *Let $\chi_s = \tau_T - \tau_s$ be the remaining optimization time after step s, $\pi_j = \prod_{u>s}(1-\eta_u g_j)$, and $S_2(\chi) = \sum_j g_j^2\pi_j^2 \asymp \chi^{1/a-2}$, $S_1(\chi) = \sum_j g_j\pi_j^2 \asymp \chi^{1/a-1}$ the spectral sums of the remaining time for aligned and for isotropic noise (Lemma E.3); and, at a constant rate ($\pi_j = r_j^h$, $r_j = 1-\eta g_j$, h the number of remaining steps), $\mathsf{Q}_\ell(h) = \sum_j g_j^\ell (1-r_j^h)^2$ the filters through which an impulse at step s reaches the average over the remaining $h = T-s$ iterates (an injection at s enters every later iterate, with total weight $((1-r_j^h)/(\eta g_j))^2$ in mode j; (59) of Appendix D.2); on the window $\mathsf{Q}_0(h) \asymp (\eta h)^{1/a}$ and $\mathsf{Q}_{-1}(h) \asymp (\eta h)^{1+1/a}$ (Lemma E.3). In the first-order-response model the penalty of a precision schedule (p_s) at one location is $\sum_s W(s)\,4^{-p_s}$ with the* response weights

$$\begin{aligned} W_{\mathrm{fw}}^{\mathrm{last}}(s) &= \eta_s^2 \mathsf{U}_s S_2(\chi_s), & W_{\mathrm{bw}}^{\mathrm{last}}(s) &= \eta_s^2 \mathsf{R}_s S_1(\chi_s), \\ W_{\mathrm{fw}}^{\mathrm{avg}}(s) &= \eta_s^2 \mathsf{U}_s \frac{\mathsf{Q}_0(T-s)}{T^2}, & W_{\mathrm{bw}}^{\mathrm{avg}}(s) &= \eta_s^2 \mathsf{R}_s \frac{\mathsf{Q}_{-1}(T-s)}{T^2}, \end{aligned} \tag{113}$$

computable in $O(MT)$ from the exact-arithmetic trajectory. Among schedules with mean $\bar p$ bits, $\sum_s W(s)4^{-p_s}$ is minimized by

$$p_s = \bar p + \tfrac{1}{2}\log_2\big(W(s)/\,\mathrm{GM}(W)\big), \qquad \frac{\textit{penalty}(p_s)}{\textit{penalty}(\bar p)} = \frac{\mathrm{GM}(W)}{\mathrm{mean}(W)}, \tag{114}$$

GM *the geometric mean over s; under Assumption 3(e) the full penalty of any schedule is within the factor $(1-\gamma_*)^{-1}$ of its first-order response.*

Proof of Proposition F.6. The weights. In the first-order response the rounding terms of (10) are evaluated on the exact-arithmetic trajectory ($\bar{\mathsf{R}}_t \to \mathsf{R}_t^0$, $\mathsf{U}_t \to \mathsf{U}_t^0$), so the penalty is linear in the rounding coefficients of the individual steps, which are proportional to $q_{r,s} = 4^{-p_{r,s}}$ up to the profile of Lemma B.1; hence it is $\sum_s W(s)4^{-p_s}$ with nonnegative weights. At step s the forward terms add $\eta_s^2\mathsf{U}_s$ times an aligned profile and the backward terms $\eta_s^2\mathsf{R}_s$ times an isotropic one, up to factors that do not depend on s; the response of the last iterate to such an addition is the first-insertion entry of Proposition E.4 resolved in the insertion time, which gives $S_2(\chi_s)$ for the aligned and $S_1(\chi_s)$ for the isotropic profile, and the response of the average is (59), which gives the filters Q_0 and Q_{-1} (Proposition F.5). Factors independent of s do not change the minimizer below; they are the constants c_r of Algorithm 1.

The allocation. Minimize $\Phi(p) = \sum_{s<T} W(s)4^{-p_s}$ over real schedules with $\sum_s p_s = T\bar p$ (the constraint $p_s \ge 0$ is not imposed), for weights $W(s) > 0$. Each term is strictly convex in p_s, so a stationary point of the Lagrangian $\Phi(p) - \lambda(\sum_s p_s - T\bar p)$ is the unique minimizer: $-\ln 4\, W(s)4^{-p_s} = \lambda$ for every s, that is, $W(s)4^{-p_s}$ does not depend on s, and $p_s = \frac{1}{2}\log_2 W(s) + c$. Averaging over s gives $\bar p = \frac{1}{2}\log_2 \mathrm{GM}(W) + c$, which is (114). At the optimum every term equals $W(s)4^{-p_s} = \mathrm{GM}(W)\,4^{-\bar p}$, so the penalty is $T\ \mathrm{GM}(W)\,4^{-\bar p}$, against $4^{-\bar p}\sum_s W(s) = T\,\mathrm{mean}(W)\,4^{-\bar p}$ for the constant schedule; the ratio is $\mathrm{GM}(W)/\mathrm{mean}(W) \le 1$ by the inequality of the arithmetic and geometric means.

The full penalty. Under Assumption 3(e), Proposition E.5 gives $I^{\mathrm{out}} \le \Delta_{\mathrm{DE}}^{\mathrm{out}} \le I^{\mathrm{out}}/(1-\gamma_*)$ for the penalty of the schedule. □

The closed-form schedules of §4.2 are the following corollary.

Corollary F.7 (closed-form schedules at a constant rate)**.** *At constant η and on the window,* (15) *gives, up to additive constants,*

$$p_{\mathrm{fw}}^{\mathrm{last}}(s) \approx \tfrac{1}{2}\log_2 \mathsf{U}_s + \tfrac{2a-1}{2a}\log_2 \frac{1}{\chi_s}, \qquad p_{\mathrm{bw}}^{\mathrm{last}}(s) \approx \tfrac{a-1}{2a}\log_2 \frac{1}{\chi_s},$$
$$p_{\mathrm{bw}}^{\mathrm{avg}}(s) \approx \tfrac{a+1}{2a}\log_2 \frac{T-s}{T}, \qquad p_{\mathrm{fw}}^{\mathrm{avg}}(s) \approx \tfrac{a-b+1}{2a}\log_2 \frac{s}{T} + \tfrac{1}{2a}\log_2 \frac{T-s}{T}: \tag{115}$$

forward bits belong at the end of training for a last iterate and near $s/T = (a-b+1)/(a-b+2)$ for an average, backward bits late for a last iterate and early for an average, and the backward allocations depend on the step only through the remaining time χ_s, not through the elapsed time τ_s.

Proof of Corollary F.7. At constant η and on the window, Lemma E.3 gives $S_2(\chi) \asymp \chi^{1/a-2}$, $S_1(\chi) \asymp \chi^{1/a-1}$, $\mathsf{Q}_0(h) \asymp (\eta h)^{1/a}$ and $\mathsf{Q}_{-1}(h) \asymp (\eta h)^{1+1/a}$, and (114) reads $p_s = \frac{1}{2}\log_2 W(s) + c$ with c fixed by the mean; constant factors in W only shift c. (a) Forward copy, last iterate: $W = \eta^2 \mathsf{U}_s S_2(\chi_s)$ gives $p_s = c' + \frac{1}{2}\log_2 \mathsf{U}_s + \frac{2a-1}{2a}\log_2(1/\chi_s)$. (b) Backward copy, last iterate: $W = \eta^2 \mathsf{R}_s S_1(\chi_s)$, and on the exact-arithmetic trajectory $\mathsf{R}_s \asymp r_s = \sigma^2 + k_s^{1-b} \in [\sigma^2, \sigma^2+1]$ (Lemma E.1), so $\frac{1}{2}\log_2 \mathsf{R}_s$ is bounded by constants depending on σ^2 and $p_s = c' + \frac{a-1}{2a}\log_2(1/\chi_s)$. (c) Backward copy, average: $W = \eta^2 \mathsf{R}_s \mathsf{Q}_{-1}(T-s)/T^2$ gives $p_s = c' + \frac{a+1}{2a}\log_2 \frac{T-s}{T}$. (d) Forward copy, average: $W = \eta^2 \mathsf{U}_s \mathsf{Q}_0(T-s)/T^2$ with $\mathsf{U}_s \asymp k_s^{a-b+1}$ and $k_s = (\eta s)^{1/a}$ on the window (Lemma E.1) gives

$$p_s = c' + \tfrac{a-b+1}{2a}\log_2 \tfrac{s}{T} + \tfrac{1}{2a}\log_2 \tfrac{T-s}{T}.$$

This is the fourth row of (115). The two terms pull in opposite directions, and the right-hand side is concave in s/T with its maximum at $s/T = (a-b+1)/(a-b+2)$, where its derivative $\frac{a-b+1}{2a\ln 2}\frac{1}{s} - \frac{1}{2a\ln 2}\frac{1}{T-s}$ vanishes. At $(a,b) = (2, \frac{3}{2})$ it stays within 0.7 bit of its maximum for $s \in [T/10, 9T/10]$ and within 1.1 bits for $s \le 0.99\,T$, against a range of eleven bits for the forward allocation of the last iterate in the time-allocation test (Appendix G). *The clock.* In (b) and (c) the schedule depends on s only through the remaining time ($\frac{T-s}{T} = \chi_s/\tau_T$ at constant η); in (a) and (d) the elapsed time enters through U_s.

F.4 The compute scaling law: derivation of (116) and of the fixed-precision frontier

The main text (§4.3) states the compute scaling law in two sentences; Theorem F.8 is its full form, and the rest of this subsection derives it.

Theorem F.8 (compute scaling law for precision (first-order response))**.** *Charge a matrix product at the width of its wider operand, $w(p) = (p+p_0)/(p_{\mathrm{ref}}+p_0)$, with p_0 the sign-and-exponent bits of the format ($p_0 = 9$ for an 8-bit exponent) and p_{ref} the mantissa bits of the reference format, so that a run costs $f = BT\,M\,[w(p_{\mathrm{fw}}) + w(p_{\mathrm{bw}})]$ when the partners are cast to the operands' widths (a copy that is not rounded is charged at the reference width, $w = 1$, and its arithmetic is exact, so that "backward rounding only" below means exact forward arithmetic; the products that form the features are not charged); let $M_*(f) \asymp (f/\bar{\kappa})^{1/(a+1)}$ be the compute-optimal width with $\tau_T = M_*^a$, so that without rounding $\mathcal{E}_*(f) \asymp f^{-(b-1)/(a+1)}$ (Paquette et al., 2024). Then, up to constants — for the average without restriction, for the constant-rate last iterate while $\varrho\sigma^2 \lesssim M_*^{1-b}$, and for the cooled last iterate while $\zeta\varrho\sigma^2 \lesssim M_*^{1-b}$ (Proposition E.10's explicit cooled law):*

(i) Optimal bits grow logarithmically in compute. *Keeping the penalty below a fixed fraction of the loss along the compute-optimal frontier requires*

$$p_{\mathrm{fw}}^* \approx \tfrac{a-1}{2(a+1)}\log_2 f \ (\textit{last iterate}), \qquad (p_{\mathrm{fw}}+p_{\mathrm{bw}})^* \approx \tfrac{a-1}{2(a+1)}\log_2 f \ (\textit{average}),$$
$$p_{\mathrm{bw}}^* \approx \tfrac{b-1}{2(a+1)}\log_2 f \ (\textit{all output rules}), \qquad p_x^* \approx \tfrac{b-1}{4(a+1)}\log_2 f, \tag{116}$$

where for the average the backward count is attainable only together with $p_{\mathrm{fw}}^* \approx \frac{a-b}{2(a+1)}\log_2 f$ forward bits, the two adding up to the joint budget; i.e. at $(a,b) = (2, \frac{3}{2})$ one forward bit per six doublings of compute (last iterate; joint budget of the average), one backward bit per twelve, one input bit per twenty-four.

(ii) Loss at optimal precision. *With bits scaled as in (i), $\mathcal{E}^*(f) \asymp (f/\log f)^{-(b-1)/(a+1)}$: the exact-arithmetic exponent is kept up to a logarithm.*

(iii) Fixed precision bends the frontier. *For the average output: at fixed $(p_x, p_{\rm fw}, p_{\rm bw})$ and matched width the loss saturates at the floor $\Phi(p) = 16^{-p_x} + \varrho\, q_{\rm bw}\sigma^2 + (\varrho\, q_{\rm fw} q_{\rm bw})^{(b-1)/(a-1)}$, in which the input bias floor 16^{-p_x} adds to the two rounding terms asymptotically; whereas, with exact input, an under-trained wider model spreads the isotropic noise over more coordinates and the compute-optimal frontier bends to the exponent $(b-1)/(a+b)$ (backward rounding only) or, asymptotically, $(b-1)/(2a)$ (both operands); the latter is stability-limited: an over-widened run amplifies its unlearned modes by a factor exponential in $\varrho q_{\rm fw} q_{\rm bw} \tau_T/M$, so the width must grow at least like $\varrho q_{\rm fw} q_{\rm bw} k_T^a/\log f$, and the frontier is $(f \log f)^{-(b-1)/(2a)}$ up to constants; while $\varrho q_{\rm fw} q_{\rm bw} \tau_T/M \lesssim 1$ the first-order response gives the intermediate exponent $(b-1)/(a+\max\{b, \frac{a+1}{2}\})$. For the constant-rate last iterate, with exact input, the floor is $\varrho\sigma^2 + (\varrho\, q_{\rm fw})^{(b-1)/(a-1)}$, backward rounding alone produces no bend, and both operands give $(b-1)/(2a)$. A bend, not the parallel shift of the fitted $N_{\rm eff}$ law of Kumar et al. (2025).*

In the cost model, $p_0 = 9$ for an 8-bit exponent, and a copy that is not rounded is computed in the reference format and charged $w = 1$: bf16 with $p_{\rm ref} = 7$ in our runs plays the role of exact arithmetic, as in practice. Part (iii) says that a fixed precision changes the slope of the compute-optimal loss on a log–log plot, not only its offset: intuitively, a wider model trained for less time spreads the isotropic rounding noise over more coordinates, so widening no longer pays off at the exact-arithmetic rate.

Constants are dropped throughout ($\asymp$, $\lesssim$ hide factors depending only on $(a, b, c, C, c_0, \theta, \Sigma_{\max}, \kappa_0, \gamma_*)$); a bit count "$\approx$" is up to an additive $O(1)$ plus the $O(\log\log f)$ of $\log_2 \bar\kappa$ below; $\vartheta, \varrho, \sigma^2, p_0, p_{\rm ref}$ are fixed; the input is exact except in the input row. Write $\mathsf{n} = \varrho\sigma^2/M^{1-b}$, $\mathsf{D}_{\rm fw} = \varrho q_{\rm fw}$, $\mathsf{D}_{\rm bw} = \varrho q_{\rm bw}\sigma^2$, $\mathsf{D}_\times = \varrho q_{\rm fw} q_{\rm bw}$ and $\mathsf{G} = \mathsf{D}_\times \tau_T/M$ as in Proposition E.7.

Cost and the matched-horizon width. A constant-rate run of optimisation time τ_T costs $f = BTMW = M\tau_T W/\varrho$, $W = w(p_{\rm fw}) + w(p_{\rm bw})$, with $w = 1$ for a copy that is not rounded (its arithmetic exact); a cooled run (rate $\eta_{\rm peak}$ for $\tau_1 = M^a$, then linear to $\zeta\eta_{\rm peak}$ over $\tau_c = \phi M^a$) has $T = (M^a/\eta_{\rm peak})(1 + 2\phi/(1+\zeta))$ steps, a constant factor in f. At the matched horizon $\tau_T = M^a$,

$$f = \bar\kappa\, M^{a+1}, \qquad \bar\kappa := \frac{W}{\varrho}, \qquad M_*(f) = \Big(\frac{f}{\bar\kappa}\Big)^{\frac{1}{a+1}}, \qquad \log_2 M_* = \frac{\log_2 f - \log_2 \bar\kappa}{a+1}, \quad (117)$$

and $\log_2 \bar\kappa = O(\log\log f)$ whenever $p_{\rm fw} + p_{\rm bw} = O(\log f)$; that the matched policy is compute-optimal to constants is the hypothesis of the law, not a consequence of (117). By Lemma E.1 at $k_T = M$, $\mathcal{E}_0^{\rm avg}(T) \asymp M^{1-b}$ unconditionally and $\mathcal{E}_0^{\rm last}(T) \asymp M^{1-b}(1+\mathsf{n})$, so the exact-arithmetic frontier $\mathcal{E}_*(f) \asymp M_*^{1-b} \asymp f^{-(b-1)/(a+1)}$ holds for the average for every f and for the constant-rate last iterate while $\mathsf{n} \lesssim 1$, i.e. $\varrho\sigma^2 \lesssim M_*^{1-b}$, a finite range of f at fixed controls (beyond it the last iterate saturates at $\varrho\sigma^2$). For the cooled last iterate the same order, $\mathcal{E}_0^{\rm cooled}(T) \asymp M^{1-b}$, is Lemma E.11: the terminal rate injects $\asymp \zeta\varrho\sigma^2$ (the stationary level of the rate $\zeta\eta_{\rm peak}$ on the top modes), the ramp $\asymp \varrho\sigma^2 M^{-d_a}\ell_a(M)$ with $d_a = \min\{a-1, a/2\}$ and $\ell_a(M) = \log M$ for $a = 2$, 1 otherwise (through the reverse rate $\eta_{\rm peak}\sqrt{\zeta^2 + (1-\zeta^2)\chi/\tau_c}$ at remaining time χ and $S_2(\chi) \asymp \chi^{1/a-2}$), and the pre-cooldown noise is reduced by k_c^{1-a}; hence the cooled law holds while $\zeta\mathsf{n} \lesssim 1$ and $\varrho\sigma^2 M^{b-1-d_a}\ell_a(M) \lesssim 1$, the second condition being automatic for large M iff $b \le 1 + a/2$.

Part (i): the bit counts. "Penalty below a fraction ϑ of the loss" at $M = M_*(f)$ is the conjunction of all rows of Corollary F.2 for the output rule at hand (Theorem 4, with its certificate assumed along the frontier); each row is solved with $q_r = 4^{-p_r}$ and clipped to $p \ge 0$. The loss is the total exact-arithmetic excess risk $A_S + \mathcal{E}_0^{\rm out}$, which differs from Corollary F.2's scale $\mathcal{E}_0^{\rm out}$ by a constant factor at matched width, i.e. by an additive $O(1)$ in every bit count. At $k_T = M$, $r_T = \sigma^2 + M^{1-b}$,

and for σ^2 fixed with $M \geq \sigma^{-2/(b-1)}$, the rows are

$$\begin{aligned}
&\text{last iterate: } q_{\rm fw} \lesssim \frac{\vartheta(1+\mathsf{n})}{\varrho} M^{1-a}, \quad q_{\rm bw} \lesssim \frac{\vartheta}{\mathsf{n}} \ (\mathsf{n} \lesssim 1) \\
&\Rightarrow\ p_{\rm fw} \geq \tfrac{a-1}{2} \log_2 M + O(1), \quad p_{\rm bw} \geq \tfrac{b-1}{2} \log_2 M + O(1); \\
&\text{average: } q_{\rm fw} \lesssim \frac{\vartheta}{\varrho}, \quad q_{\rm bw} \lesssim \frac{\vartheta}{\mathsf{n}}, \quad q_{\rm fw} q_{\rm bw} \lesssim \frac{\vartheta}{\varrho} M^{1-a} \\
&\Rightarrow\ p_{\rm bw} \geq \tfrac{b-1}{2} \log_2 M + O(1), \quad p_{\rm fw} + p_{\rm bw} \geq \tfrac{a-1}{2} \log_2 M + O(1); \\
&\text{cooled: } q_{\rm bw} \lesssim \frac{\vartheta M^{1-b}}{\varrho \sigma^2 c(\phi,\zeta)} \ \Rightarrow\ p_{\rm bw} \geq \tfrac{b-1}{2} \log_2 M + O(1); \\
&\text{input: } 16^{-p_x} \lesssim \vartheta M^{1-b} \ \Rightarrow\ p_x \geq \tfrac{b-1}{4} \log_2 M + O(1)
\end{aligned} \tag{118}$$

(last iterate: the forward and backward rows of (12) with $\mathcal{E}_0^{\rm last} \asymp M^{1-b}(1+\mathsf{n})$, imposed simultaneously; average: the forward, backward and interaction rows of (12), the interaction row from $\varrho q_{\rm fw} q_{\rm bw} M^{a-b} \leq \vartheta M^{1-b}$, a sum of bit counts; cooled: $c(\phi,\zeta)\varrho_{\rm peak} q_{\rm bw} \sigma^2 \leq \vartheta \mathcal{E}_0^{\rm cooled}$ by Proposition E.10, the forward copy at the reference width being assumed harmless after the cooldown; input: the asymptotic ridge bias $16^{-p_x}\|v_*^S\|_G^2$ of Proposition C.1(ii) against the exact-arithmetic order, a bias-only requirement, uncharged in W). Inserting (117): for the last iterate $(p_{\rm fw}^*, p_{\rm bw}^*) = \big(\frac{a-1}{2(a+1)}, \frac{b-1}{2(a+1)}\big)(\log_2 f - \log_2 \bar\kappa) + O(1)$; for the average the feasible set is $\{p_{\rm bw} \geq \frac{b-1}{2}\log_2 M,\ p_{\rm fw} + p_{\rm bw} \geq \frac{a-1}{2}\log_2 M\}$ up to $O(1)$, so $(p_{\rm fw}+p_{\rm bw})^* \approx \frac{a-1}{2(a+1)}\log_2 f$ with any split between the corners $(\frac{a-b}{2}, \frac{b-1}{2})\log_2 M$ and $(0, \frac{a-1}{2})\log_2 M$: the backward count $p_{\rm bw}^* \approx \frac{b-1}{2(a+1)}\log_2 f$ is attainable only together with $p_{\rm fw}^* \approx \frac{a-b}{2(a+1)}\log_2 f$; for the cooled last iterate $p_{\rm bw}^* \approx \frac{b-1}{2(a+1)}\log_2 f$ (the step factor of the cooldown is a constant in f); for the input $p_x^* \approx \frac{b-1}{4(a+1)}\log_2 f$. This is (116); at $(a,b) = (2, \frac{3}{2})$ the slopes are $\frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{24}$. The last-iterate rows are finite-range laws: for $\mathsf{n} \gg 1$ at fixed controls they give $p_{\rm fw} \gtrsim \frac{a-b}{2}\log_2 M$ and $q_{\rm bw} \lesssim \vartheta$, and the total loss $\asymp \varrho\sigma^2$ is no longer a frontier.

Part (ii). With all rows of (i) imposed, $\mathcal{E}_0^{\rm out} \leq \mathcal{E}_{\rm DE}^{\rm out} = \mathcal{E}_0^{\rm out} + \Delta_{\rm DE}^{\rm out} \leq (1+\vartheta)\mathcal{E}_0^{\rm out}$ at $M = M_*(f)$ (positivity below; the sandwich of Proposition E.5 above, its constant inside the $O(1)$ of the bit counts). In every minimal-budget allocation of (i) the charged sum grows like $\log_2 f$, so $W \asymp \log_2 f$, $\bar\kappa(f) \asymp \log f/\varrho$, and

$$M_*(f) \asymp \Big(\frac{f}{\log f}\Big)^{\frac{1}{a+1}}, \qquad \mathcal{E}^*(f) \asymp M_*(f)^{1-b} \asymp \Big(\frac{f}{\log f}\Big)^{-\frac{b-1}{a+1}} \tag{119}$$

on the ranges of the exact-arithmetic law above. This order is also the optimum over bit choices at fixed controls (η, B): for any run $f = \bar\kappa k^a M \geq \bar\kappa k^{a+1}$, so $\mathcal{E} \geq \mathcal{E}_0 \gtrsim k^{1-b} \geq (f/\bar\kappa)^{-(b-1)/(a+1)}$; a charged copy with $p \geq \varepsilon \log_2 f$ makes $\bar\kappa \gtrsim \varepsilon \log f/\varrho$, while bits $\leq \varepsilon \log_2 f$ at every charged copy leave, by positivity alone, a penalty term $(\mathsf{D}_{\rm fw} k^{a-b+1}/M$ with $q_{\rm fw} \geq f^{-2\varepsilon}$ for the last iterate, $\mathsf{D}_\times k^{a-b+2}/M^2$ with $q_{\rm fw} q_{\rm bw} \geq f^{-4\varepsilon}$ for the average) whose minimum over (k, M) at fixed f is polynomially larger than $f^{-(b-1)/(a+1)}$ when $\varepsilon < \frac{a-1}{2(a+1)}$ (last iterate) or $\varepsilon < \frac{a-1}{4(a+1)}$ (average).

Part (iii), the floor at matched width. Exact input throughout (iii). For the average with both operands rounded at fixed bits, at $k_T = M$,

$$\mathcal{E}_0^{\rm avg} + I^{\rm avg} \asymp M^{1-b} + \mathsf{D}_{\rm bw} + \mathsf{D}_\times M^{a-b}, \qquad \inf_M \big[M^{1-b} + \mathsf{D}_{\rm bw} + \mathsf{D}_\times M^{a-b}\big] \asymp \mathsf{D}_{\rm bw} + \mathsf{D}_\times^{\frac{b-1}{a-1}} =: \Phi_0 \tag{120}$$

(the forward term $\varrho q_{\rm fw} M^{1-b}$ and $\varrho\sigma^2 M^{1-a}$ are absorbed into the bias), the minimum of $M^{1-b} + \mathsf{D}_\times M^{a-b}$ being reached at $M_{\rm cap} \asymp \mathsf{D}_\times^{-1/(a-1)}$; put $M_{\rm cap}^{\rm bw} = \mathsf{D}_{\rm bw}^{-1/(b-1)}$. [lower bound] $\mathcal{E}_{\rm DE}^{\rm avg}(T; M) \gtrsim \Phi_0$ for every M, by positivity. [certified] $\mathcal{E}_{\rm DE}^{\rm avg}(T; M_\star) \lesssim \Phi_0$ at $M_\star = c_\star \min\{M_{\rm cap}, M_{\rm cap}^{\rm bw}\}$ with $c_\star$ small: there the bias is $\asymp \Phi_0$, the interaction term is $\leq c_\star^{a-b}\Phi_0$, and the width-dependent terms of (69) are $\mathsf{G} \leq c_\star^{a-1}\frac{b-1}{a-b}$ and $\varrho q_{\rm bw}(1 + \sigma^2 M_\star^{b-1}) \leq \varrho q_{\rm bw} + c_\star^{b-1}$, so the certificate holds once $\varrho(\theta_G + q_{\rm fw} + q_{\rm bw})$ is small. Hence Φ_0 is the lower envelope of the matched-width family, reached at $f_{\rm sat} \asymp \bar\kappa \min\{M_{\rm cap}, M_{\rm cap}^{\rm bw}\}^{a+1}$ ("saturates"); beyond $M_{\rm cap}$ the interaction term, and by Proposition E.7 the amplification of the unlearned modes, turn the matched-width curve up. With a shared rounded input the asymptotic risk is also $\geq 16^{-p_x}\|v_*^S\|_G^2 \asymp 16^{-p_x}$

(Proposition C.1(ii)): the input bias floor adds to Φ_0 asymptotically, the two being lower bounds of different kinds (exact input at finite time; rounded input as $t \to \infty$). $M_\star$ is a matched-width run inside the window of Conjecture 2, so the floor is the one statement of (iii) that transfers to the algorithm through it. For the constant-rate last iterate, $\mathcal{E}_0^{\rm last} + I^{\rm last} \asymp M^{1-b} + \varrho\sigma^2 + \mathsf{D}_{\rm fw} M^{a-b}$ (the backward term $\varrho q_{\rm bw}(\sigma^2 + M^{1-b})$ is absorbed), so its floor is $\varrho\sigma^2 + \mathsf{D}_{\rm fw}^{(b-1)/(a-1)}$, reached at $c_\star M_{\rm cap}^{\rm last}$, $M_{\rm cap}^{\rm last} \asymp \mathsf{D}_{\rm fw}^{-1/(a-1)}$, [certified] on $\mathsf{n} \lesssim 1$: the SGD floor $\varrho\sigma^2$ stands where Φ_0 has $\mathsf{D}_{\rm bw}$, and the cap is set by the forward term alone.

Part (iii), the bends. Over-widened runs have $k := k_T = \tau_T^{1/a} < M$, $f = \bar\kappa k^a M$ and $\mathsf{G} = \mathsf{D}_\times k^a/M$. At fixed positive bits the last term of (69) is $\geq \varrho \min\{q_{\rm bw}\sigma^2 k^{b-1}, q_{\rm fw} k^{a-1}\}$ and grows with k whatever M is, so the certificate holds only for $k \lesssim k_{\rm cert} := \min\{(\gamma_*/(C_{\rm sg}\mathsf{D}_{\rm bw}))^{1/(b-1)}, (\gamma_*/(C_{\rm sg}\varrho q_{\rm fw}))^{1/(a-1)}\}$, the backward crossover scale, and never covers an over-widened branch beyond it. *Backward rounding only* (exact forward arithmetic, $q_{\rm fw} = 0$): $\mathcal{E}_0^{\rm avg} + I^{\rm avg} \asymp k^{1-b} + \mathsf{D}_{\rm bw} k/M$; at fixed f the optimum balances the two, $M \asymp \mathsf{D}_{\rm bw} k^b$, $f \asymp \bar\kappa \mathsf{D}_{\rm bw} k^{a+b}$, and

$$\mathcal{E}_{\rm DE}^{\rm avg} \asymp k^{1-b} \asymp \Big(\frac{f}{\bar\kappa\, \mathsf{D}_{\rm bw}}\Big)^{-\frac{b-1}{a+b}} \qquad \big(f \gtrsim f_{\rm sat}^{\rm bw} \asymp \bar\kappa\, \mathsf{D}_{\rm bw}^{-\frac{a+1}{b-1}}\big), \tag{121}$$

[lower bound] for every policy and [certified] along this one: with $q_{\rm fw} = 0$ the last term of (69) vanishes, $\mathsf{G} = 0$, and the bracket is $\leq \varrho(\theta_G + q_{\rm bw})$ on the whole window (Remark E.8(d)) — the one bend that is fully certified; a forward copy run in the reference format instead is a both-operands run with $\mathsf{D}_\times = \varrho 4^{-p_{\rm ref}} q_{\rm bw}$, harmless only up to the backward crossover. The constant-rate last iterate has no such bend: $\mathsf{D}_{\rm bw} k/M \leq \varrho\sigma^2$ at every (k, M), so on $\mathsf{n} \lesssim 1$ its frontier is the exact-arithmetic one and beyond it saturates at $\varrho\sigma^2$. *Both operands, first-order regime.* $\mathcal{E}_0^{\rm avg} + I^{\rm avg} \asymp k^{1-b} + \mathsf{D}_{\rm bw} k/M + \mathsf{D}_\times k^{a-b+2}/M^2$ (the interaction entry of Proposition E.4 carries $1/M^2$; the forward entry $\varrho q_{\rm fw} k^{2-b}/M \leq \varrho q_{\rm fw} k^{1-b}$ is absorbed), minimised at fixed f by $M \asymp \max\{k, \mathsf{D}_{\rm bw} k^b, \mathsf{D}_\times^{1/2} k^{(a+1)/2}\}$ with exponent $(b-1)/(a + \max\{b, \frac{a+1}{2}\})$; along that policy $\mathsf{G} \asymp \min\{(\mathsf{D}_\times/\mathsf{D}_{\rm bw}) k^{a-b}, \mathsf{D}_\times^{1/2} k^{(a-1)/2}\}$ reaches order one at $k_\times \asymp \max\{(\mathsf{D}_{\rm bw}/\mathsf{D}_\times)^{1/(a-b)}, \mathsf{D}_\times^{-1/(a-1)}\}$. The surrogate is a lower bound for every policy, the loss [certified] for $k \lesssim k_{\rm cert}$, and for $k \gtrsim k_\times$ the amplification of Proposition E.7 sets in: this is the intermediate exponent "while $\mathsf{G} \lesssim 1$" ($\frac{1}{7}$ at $(2, \frac{3}{2})$; at the defaults of the experiments $k_\times$ lies beyond every run). *Both operands, asymptotically.* By Proposition E.7(v),

$$\inf_{\bar\kappa k^a M = f,\ k \leq M} \mathcal{E}_{\rm DE}^{\rm avg}(T) \asymp \Big(\frac{f \log f}{\bar\kappa\, \mathsf{D}_\times}\Big)^{-\frac{b-1}{2a}}, \qquad \text{attained by } M \asymp \frac{\mathsf{D}_\times k^a}{\log f}\ \ (\mathsf{G} = \varepsilon \log f), \tag{122}$$

[lower bound] from the amplification of its part (ii) and positivity, [upper bound] from the growth bound of its parts (iii)–(iv), no certificate: an over-widened run amplifies its unlearned modes by a factor between $e^{c'\mathsf{G}}$ and $e^{L\mathsf{G}}$, so the width must grow at least like $\mathsf{D}_\times k^a/\log f$, the exponent is $(b-1)/(2a)$, and the order is better than $f^{-(b-1)/(2a)}$ by $(\log f)^{-(b-1)/(2a)}$. Along the policy the weak-mode feedback exceeds the contraction ($\mathsf{D}_\times M^{a-1} \to \infty$); what it bounds is the amplification accumulated over the run. For the constant-rate last iterate on $\mathsf{n} \lesssim 1$, $\mathcal{E}_0^{\rm last} + I^{\rm last} \asymp k^{1-b} + \mathsf{D}_{\rm fw} k^{a-b+1}/M$ (the backward term $\leq q_{\rm bw}(\mathsf{n}_k + \varrho) k^{1-b}$, $\mathsf{n}_k = \varrho\sigma^2 k^{b-1}$, is absorbed), balanced by $M \asymp \mathsf{D}_{\rm fw} k^a$, $f \asymp \bar\kappa \mathsf{D}_{\rm fw} k^{2a}$: $\mathcal{E}_{\rm DE}^{\rm last} \asymp (f/(\bar\kappa \mathsf{D}_{\rm fw}))^{-(b-1)/(2a)}$, [certified] along the policy ($\mathsf{G} = c q_{\rm bw}$ bounded, the last term of (69) at most $q_{\rm bw}(\varrho + \mathsf{n}_k)$) on $\bar\kappa \mathsf{D}_{\rm fw}^{-(a+1)/(a-1)} \lesssim f \lesssim \bar\kappa \mathsf{D}_{\rm fw} (\varrho\sigma^2)^{-2a/(b-1)}$, a range that is nonempty iff $\mathsf{D}_{\rm fw}^{(b-1)/(a-1)} \gtrsim \varrho\sigma^2$: "both operands give $(b-1)/(2a)$". *Scope of (iii).* Exact input; fixed bits, controls and noise (the onset scales $k_{\rm cert}$, $k_\times$, $M_{\rm amp}$ depend on them); uniform average; the window. The bends use $\tau_T = k^a \ll M^a$, outside the window of Conjecture 2, so of (iii) only the floor transfers to the algorithm; the backward rows of (i) transfer by Theorem E.12 up to its $\log M$ factors, with exact input and exact forward arithmetic.

Numerical illustration. Deterministic-equivalent dynamics with the realised coefficients at $(a, b) = (2, \frac{3}{2})$, $\sigma^2 = 0.1$, $B = 32$, $d = 4M$, seed 1, penalties at $p = 1$ (210 trajectories; ratios are taken with the loss $A_S + \mathcal{E}_0^{\rm out}$ as denominator, with the excess-only ratios alongside; fits are $\log_2$–$\log_2$ over $M \in \{256, \dots, 2048\}$ with the residual standard error of the fit). *Rows of* (118). Width slopes of the penalty-to-loss ratios: forward, last iterate 0.997 ± 0.014 (0.850 against the excess alone) for the predicted $a - 1 = 1$; interaction, average 1.009 ± 0.010 for 1; forward, average -0.017 ± 0.016 for

0; the three backward rows 0.28, 0.27, 0.30 raw and 0.470 ± 0.012, 0.456 ± 0.014, 0.485 ± 0.013 once divided by the effective noise $\sigma^2_{\text{eff}}(M) = \sigma^2 + 2A_S$ (which falls from 0.20 to 0.14 across the grid), for the predicted $b - 1 = \frac{1}{2}$. The ratio $L/\mathcal{E}_0$ is constant for the average (8.4, slope 0.004) and drifts for the last iterate ($15 \to 11$: its excess is noise-dominated, slope -0.34, while L has slope -0.488 for $-\frac{1}{2}$). Over-widened runs $M = k^\gamma$, ratio slopes in $\log_2 k$: interaction 0.95, 0.06, -0.96 at $\gamma = 1, 1.5, 2$ for $1, 0, -1$; forward, last iterate 0.96, 0.50, 0.02 for $1, \frac{1}{2}, 0$; backward, average, divided by $r_k = \sigma^2_{\text{eff}} + k^{1-b}$, 0.52, 0.12, -0.46 for $\frac{1}{2}, 0, -\frac{1}{2}$. At $(3, \frac{3}{2})$ with $\sigma^2 = 1$ the backward-average ratio along $\gamma = \frac{3}{2}$ has slope -0.09 (-0.04 normalised) for the predicted 0. *Bits against compute.* With $p^* = p + \frac{1}{2}\log_2(\rho/\vartheta)$, $\vartheta = 0.1$, $f = BTM\,W$ and the $\bar{\kappa}$ correction subtracted: forward, last iterate 0.162 ± 0.001 for 0.162; joint budget of the average 0.164 ± 0.001 for 0.163; the backward counts 0.045–0.049 for 0.081, the shortfall being that of the raw backward ratios (and on this grid the backward thresholds are negative, i.e. the backward constraint is inactive at $\vartheta = 0.1$: the clipping of (i)). Reconstructed risks at the fitted bits (interpolated in q, largely outside the recorded precisions) have compute slopes -0.165 to -0.173 against $\log_2(f/\log_2 f)$ for the predicted $-\frac{1}{6}$. *Floors, caps and envelopes.* On the scale of the total loss both caps lie beyond $M = 2048$ for $p \geq 1$ ($A_S \approx 0.8M^{-1/2}$ dominates the excess bias by two orders of magnitude); relative to the excess alone the last-iterate cap at $p = 1$ is interior ($M = 192$) and moves $\times 4$ in width per added bit; the lower envelopes are boundary-limited at $M = 2048$ (upper-decade exact-arithmetic exponents 0.14–0.18 for $\frac{1}{6}$) and read off no bend. *Cooled exact-arithmetic law.* $\mathcal{E}_0^{\text{cooled}}(T)/M^{1-b} = 0.0077, 0.0079, 0.0078, 0.0080$ at $M = 256, \ldots, 2048$ ($\phi = 1$, $\zeta = 0.02$; two-point slopes $0.04, -0.01, 0.02$ for 0; $\mathcal{E}_0^{\text{cooled}}/A_S \approx 0.0096$ throughout): the hypothesis behind the cooled rows holds on this range. *Growth bound.* In the rescaled system of Remark E.8(c) ($M = 256$, mixed coefficient $\times 391$, $p = 0$), the maxima of $e^{-\lambda\tau_t}\bar{\mathsf{R}}_t$ and $e^{-\lambda\tau_t} q_{\text{fw}} \mathsf{U}_t/M$ with $\lambda = L\varrho q_{\text{fw}} q_{\text{bw}}/M$ over horizons $\tau_T \in \{1, 2, 4\}M^a$ are unchanged from the shortest horizon for $L \in \{4, 8\}$ and grow with the horizon for $L \leq 2$ (by $2.5 \cdot 10^{15}$ at $L = 0$, $4.7 \cdot 10^3$ at $L = 2$): the weighted bound of Proposition E.7(iii) holds there with a threshold between 2 and 4 (a finite-horizon check, not a proof). Amplification itself: Remark E.8(c).

F.5 The tolerance under a wall-clock budget: derivation behind §4.3

The main text (§4.3) states the tolerance under a wall-clock budget in one sentence; Corollary F.9 is the full statement, and the rest of this subsection derives it. One more bit costs about the fraction $1/(p + p_0)$ of the steps and divides the penalty by four. The statement concerns the deterministic-equivalent dynamics; Theorems 4 and E.12 transfer the values of the penalty, not the location of the optimum; in the power-law regime $e_0 = (b - 1)/a$, and one bit trades against about 8% of the steps at $p^* = 3$, $p_0 = 9$.

Corollary F.9 (the tolerance under a wall-clock budget)**.** *Fix the width, batch and step, round both copies to p bits, and charge each product at $w(p)$ (Theorem F.8), so that a wall-clock limit fixes $f = 2BM\,w(p)\,T$ and $\tau_T \propto f/w(p)$. Let $e_0 := -d\log\mathcal{E}_0^{\text{last}}/d\log\tau_T$ be the loss–time elasticity of the exact-arithmetic excess risk at the operating point ($e_0 = (b-1)/a$ for the power law of Lemma E.1 when $k_T \ll M$ and the SGD floor is negligible; larger at fixed width on the window). If the tolerance $\vartheta = \Delta/\mathcal{E}_0$ scales like 4^{-p} and the optimum over p is interior, then, to first order in $e_\Delta/(\ln 4\,(p^* + p_0))$, with e_Δ the elasticity of the penalty in τ_T, the tolerance at the optimum is*

$$\vartheta^* \approx \frac{e_0}{\ln 4\,(p^* + p_0)} = \frac{b-1}{a\ln 4\,(p^* + p_0)} \approx 1.5\% \quad \text{at } (a,b) = (2, \tfrac{3}{2}),\ p^* = 3,\ p_0 = 9 \qquad (123)$$

in the power-law regime (about 3% under the energy proxy $w(p_1)w(p_2)$): at the optimum the retained penalty equals the loss gain of about $1/(\ln 4\,(p^* + p_0)) \approx 6\%$ more optimisation time, and one bit trades against $1/(p^* + p_0) \approx 8\%$ of the steps. So the fraction of the excess risk that should be spent on rounding is derived, not chosen. On the deterministic-equivalent dynamics at fixed width on the window, e_0 is 0.3–1.0 at $(a,b) = (2, \frac{3}{2})$ rather than 0.25, and ϑ^* is correspondingly larger.

Setting. Fix S, M, B and η (hence ϱ and A_S); round both copies to the same p bits, constant in time; charge each product at $w(p) = (p + p_0)/(p_{\text{ref}} + p_0)$ with the partners cast to the operands' widths (§4.3), so that a run costs $f = 2BM\,w(p)\,T$ and a wall-clock limit on fixed hardware is a fixed f: $\tau_T = \eta T = C_f/P^m$ with $P = p + p_0$, $C_f = \eta f(p_{\text{ref}} + p_0)^m/(2BM)$, $m = 1$ for this time proxy and $m = 2$ for the energy proxy $w(p_1)w(p_2)$. The risk of the run is $L(p) = A_S + \mathcal{E}_0(\tau_T(p)) + \Delta(p, \tau_T(p))$

and the tolerance is $\vartheta = \Delta/\mathcal{E}_0$ (Corollary F.2), both in excess of A_S. Everything below is about the smooth surrogate in which the scaling forms of Lemma E.1 and Theorem 4 are power laws with fixed prefactors, $\mathcal{E}_0 = C_0\tau^{-\alpha} + N$ ($\alpha = \frac{b-1}{a}$, $N \propto \varrho\sigma^2$ the SGD floor) and $\Delta = \sum_i K_i 4^{-s_i p}\tau^{\beta_i}$ over the rows of Theorem 4 ($s_i = 1$ for the forward and backward rows, 2 for the interaction row; $\beta = \frac{a-b+1}{a}$ for the forward last-iterate row, $\frac{2-b}{a}$ forward average, $\frac{1}{a}$ and $\frac{2-b}{a}$ for the two backward monomials $\sigma^2 k$ and k^{2-b}, $\frac{a-b+2}{a}$ interaction), with p relaxed to a real variable. An $\asymp$ fixes no derivative, so nothing here is a theorem about the dynamics: it is a first-order-response calculation in the sense of §4.3, and it says nothing about the rounded algorithm (Theorem E.12 and Conjecture 2 transfer values, not derivatives or optima).

Elasticities. $e_0 := -\partial \log \mathcal{E}_0/\partial \log \tau = \alpha s$ with $s = C_0\tau^{-\alpha}/(C_0\tau^{-\alpha} + N)$ the bias share ($s \to 1$ when the floor is negligible); $e_\Delta := \partial \log \Delta/\partial \log \tau|_p$, the Δ_i-weighted average of the β_i ($= \beta_{\mathrm{fw}}$ within about 10% for the last iterate at unit prefactors, since $\Delta_{\mathrm{fw}}/\Delta_{\mathrm{bw}} \asymp k^{a-b}/(\sigma^2 + k^{1-b}) \geq 11$ for $k \geq 16$ at $(a, b, \sigma^2) = (2, \frac{3}{2}, 0.1)$, Appendix F.2); $\ell := -\partial \log \Delta/\partial p|_\tau = \ln 4 \sum_i s_i\Delta_i/\Delta$, equal to $\ln 4$ up to the dominated interaction contribution of the last iterate (Proposition E.9).

Stationary identity. Since $d \log \tau_T/dp = -m/P$ and A_S is fixed,

$$\frac{dL}{dp} = \frac{m\, e_0\, \mathcal{E}_0}{P} - \ell\, \Delta - \frac{m\, e_\Delta\, \Delta}{P} = \frac{m\, e_0\, \mathcal{E}_0}{P}\big[1 - \rho(p)\big], \qquad \rho(p) := \frac{(\ell P + m e_\Delta)\, \Delta}{m\, e_0\, \mathcal{E}_0}, \quad (124)$$

so at an interior stationary point

$$\vartheta_{\mathrm{stat}} = \frac{\Delta}{\mathcal{E}_0} = \frac{m\, e_0}{\ell P + m\, e_\Delta} \qquad \Big(\text{several monomials: } m\, e_0\, \mathcal{E}_0 = \sum_i (s_i \ln 4\, P + m\beta_i)\, \Delta_i\Big). \quad (125)$$

The first term of (124) is the marginal loss of one more bit (it costs the fraction m/P of the optimisation time, hence me_0/P of the exact-arithmetic risk), the second its marginal gain (it divides the penalty by four), the third the smaller change of the injected noise (of either sign when $\beta < 0$, which needs $b > 2$). With $m = 1$, $\ell = \ln 4$, $e_0 = \alpha$ and the factor $(1 + e_\Delta/(\ln 4\, P))^{-1}$ dropped, (125) becomes $\vartheta^* \approx e_0/(\ln 4\, (p^* + p_0)) = (b-1)/(a \ln 4\, (p^* + p_0))$ in §4.3. At $(a, b, p^*, p_0) = (2, \frac{3}{2}, 3, 9)$ the stationary value is 1.438% and the leading term 1.503% (4.5% of the former; imposing the leading value leaves $dL/dp = -e_\Delta \vartheta \mathcal{E}_0/P \neq 0$); under the energy proxy $2e_0/(\ln 4\, P + 2e_\Delta) = 2.76\%$ (leading 3.01%); when the interaction row dominates the sum (average output, both ratios of (112) large) $\ell = \ln 16$ and the tolerance halves, $e_0/(2 \ln 4\, P + \frac{a-b+2}{a}) = 0.72\%$.

Existence. With constant elasticities $d \log \rho/dp = -\ell + [1 - m(e_0 + e_\Delta)]/P - me_\Delta/[P(\ell P + me_\Delta)] < 0$, so on the precision interval $[p_L, p_U]$ cut out by the window ($\tau_T \leq M^a$ at the lower end, the burn-in $\tau_T \geq \tau_*$ at the upper end) dL/dp changes sign at most once: an interior optimum exists iff $\rho(p_L) > 1 > \rho(p_U)$, it is then unique and obeys (125); otherwise the minimiser is an end of the interval, with $\vartheta \leq \vartheta_{\mathrm{stat}}$ at the lower end (the first bit does not pay) and $\vartheta \geq \vartheta_{\mathrm{stat}}$ at the upper end. Integer precisions are compared through neighbours: in the forward surrogate an optimum at $p = 3$ is consistent with any $\vartheta(3) \in [0.66\%, 2.64\%]$.

Exchange rates. At the optimum one bit trades against m/P of the optimisation time (8.3% at $P = 12$; 9.1% more steps for one bit fewer, 7.7% fewer for one bit more), and the loss balance is $e_0/P = 2.08\%$ of $\mathcal{E}_0$ on each side of (124). The retained penalty $\vartheta_{\mathrm{stat}}$ equals the exact-arithmetic-risk gain of about $\vartheta_{\mathrm{stat}}/e_0 = 5.8\%$ more optimisation time ($1/(\ln 4\, P) = 6.0\%$ when the changing penalty is counted as well): this, and not the marginal loss of a bit, is the "6%" of §4.3.

Fixed width. $e_0 = \alpha$ is the $k_T \ll M$, bias-dominated value. At fixed width the bias part of the excess risk is the truncated mode sum, $\mathrm{Bias} = \frac{1}{2}\sum_{j \leq M} g_j w_j \pi_j(T)^2 \approx \frac{(2\tau)^{-\alpha}}{2a}\Gamma(\alpha, z)$ with $z = 2\tau_T/M^a$, whose elasticity $\alpha + z^\alpha e^{-z}/\Gamma(\alpha, z)$ rises from 0.29 at $k_T/M = 0.01$ to 2.8 at $k_T = M$ for $(2, \frac{3}{2})$, and $e_0 = s\, e_{\mathrm{B}} + (1 - s)e_{\mathrm{N}}$ with the non-bias part's own elasticity (e_{N} of either sign). On the reduced dynamics of one realised feature map (seed 1; $\sigma^2_{\mathrm{label}} = 0.1$, $B = 32$, $\eta = 0.5/g_1$, exact SR cell moments) the exact-arithmetic last-iterate elasticity is 0.29–0.30, 0.33, 0.42–0.43, 0.60–0.64 and 0.73–0.96 at $k_T/M = 0.01, 0.03, 0.1, 0.3, 1$ for $M = 256$–2048: above $\alpha = \frac{1}{4}$ everywhere on the window, and the slope of $\log_2 \vartheta$ against p on $p = 2..6$ is -2.00 to 0.1% ($\ell = \ln 4$). Since (125) is linear in e_0, the stationary tolerance at fixed width exceeds the power-law value by $(e_0/\alpha)(1 + e_\Delta/(\ln 4\, P))^{-1}$: on the six interior in-window budgets at $M = 512, 1024, 2048$ the interpolated optimum agrees with (125) to 0.7% and lies 1.7–2.7 times above the leading term;

on the idealised dynamics ($g_j = j^{-a}$, $w_j = j^{a-b}$, $\iota_{\mathrm{fw}} = 0.09\, q\ \mathrm{tr}\,\Sigma_d/M$) at the budget whose $p = 0$ run reaches $\tau_T = M^a$, the optimum is $p^* = 3$ ($M = 256$) or 4 ($M = 1024$) at $\vartheta = 6.7\%$ and 5.3%, 4.5 and 3.8 times the leading term. This is why the corollary's e_0 is the elasticity at the operating point and "1.5%" its value in the power-law regime.

F.6 Precision and the learning-rate schedule: the lemmas and the derivation behind §4.4

The main text (§4.4) states the following proposition in two sentences; Proposition F.10 is its full form, and the rest of this subsection derives it. Write $\chi_s = \tau_T - \tau_s$ for the optimisation time remaining after step s, $S_\ell(\chi) = \sum_j g_j^\ell e^{-2g_j\chi}$ for the spectral sums, and

$$K(\chi) = S_2(\chi) + \iota_{\mathrm{bw}} S_1(\chi) \asymp \chi^{\frac{1}{a}-2}\big(1+\kappa\chi\big), \qquad \kappa \approx 0.09\, c_a\, q_{\mathrm{bw}}\, \frac{\mathrm{tr}\,\Sigma_d}{M}, \qquad \chi_\times = \frac{1}{\kappa}, \tag{126}$$

for the noise kernel of the last-iterate loss in the optimal-control formulation of Bordelon & Mori (2026) (bias plus a noise integral against a kernel of the remaining time; the residual frozen at σ^2_{eff}): SGD noise dominates the kernel while the remaining time is below the crossover $\chi_\times$, backward-rounding noise while it is above. Here $c_a = S_1(\chi_\times)/(\chi_\times S_2(\chi_\times))$ is a spectral constant, $2a/(a-1)$ in the bulk.

Proposition F.10 (precision and the learning-rate schedule (first-order response))**.** *Backward rounding at $q_{\mathrm{bw}} = 4^{-p_{\mathrm{bw}}}$, label noise $\sigma^2_{\mathrm{eff}} > 0$, step sizes below a cap under which the reduced system is stable (condition (H) below). Up to constants depending on $\sigma^{-2}_{\mathrm{eff}}$:*

(i) Shape. *The schedule minimising the last-iterate loss decays as $\eta \propto K(\chi)^{-1/2}$: as $(T-t)^{2a-1}$ in exact arithmetic, but only as $(T-t)^{(a-1)/(a+1)}$ while the remaining time exceeds $\chi_\times$. Low precision calls for a gentler and longer cooldown; the steep final phase occupies a share $\asymp (\kappa\tau_T)^{-(a+1)/(2a)}$ of the run.*

(ii) Peak rate and batch. *For runs in which rounding noise dominates the SGD noise ($\kappa\tau_T \gg 1$, i.e. $p_{\mathrm{bw}} \lesssim \frac{1}{2}\log_2(\tau_T/M) + O(1)$) and the cap is inactive, the optimal peak rate at fixed width and steps is*

$$\eta^* \asymp \Big(\frac{BM}{q_{\mathrm{bw}}\,\mathrm{tr}\,\Sigma_d\,\sigma^2_{\mathrm{eff}}}\Big)^{\frac{a}{a+b}} T^{-\frac{b}{a+b}}\ : \tag{127}$$

one bit less lowers it by $4^{a/(a+b)}$ (≈ 2.2 at $(a,b) = (2,\frac{3}{2})$), or raises the optimal batch by the same factor at a fixed sample budget BT.

(iii) Training-time exponent. *In the same regime, at fixed width and bits the schedule-optimal loss is $\asymp \big(q_{\mathrm{bw}}\,\mathrm{tr}\,\Sigma_d\,\sigma^2_{\mathrm{eff}}/(BMT)\big)^{(b-1)/(a+b)}$: it falls with training time at the rate $(b-1)/(a+b)$ instead of the $(b-1)/a$ of exact arithmetic. Growing the bits as $p_{\mathrm{bw}} \geq \frac{b}{2a}\log_2\tau_T - \frac{1}{2}\log_2 M + O(1)$ restores $(b-1)/a$.*

(iv) Precision schedule. *Through a linear cooldown, the allocation of Proposition F.6 lowers the backward precision by $\frac{1}{2a}$ bit per halving of the remaining optimisation time and raises the forward precision by at least $\frac{a-1}{2a}$ bit.*

Setting. Backward rounding only, with a deterministic step-size schedule $\eta_t \leq \eta_{\max}$ and precision schedule $\iota_{\mathrm{bw},t} \leq \iota_{\mathrm{bw}}$, $\bar\alpha_{\mathrm{bw},t} \leq \bar\alpha_{\mathrm{bw}}$; the reduced system (10) with $\iota_{\mathrm{fw}} = \bar\alpha_{\mathrm{fw}} = \gamma_{\mathrm{fwbw}} = 0$ and the time-dependent propagator $\pi_j(t,u) = \prod_{v=u}^{t-1}(1-\eta_v g_j)$ of Appendix E.7 reads

$$\bar e_j(t{+}1) = (1-\eta_t g_j)^2 \bar e_j(t) + \frac{\eta_t^2}{B}\Big[g_j\big(\bar{\mathsf{R}}_t + g_j\bar e_j(t)\big) + \iota_{\mathrm{bw},t}\bar{\mathsf{R}}_t + 2\bar\alpha_{\mathrm{bw},t}\bar{\mathsf{S}}_t\Big], \qquad \bar e_j(0) = w_j, \tag{128}$$

with $\bar{\mathsf{R}}_t = \sigma^2 + \sum_l g_l\bar e_l(t)$, $\bar{\mathsf{S}}_t = \frac{1}{M}\sum_l g_l^2\bar e_l(t)$ and $\sigma^2 = \sigma^2_{\mathrm{eff}} > 0$; exact-arithmetic quantities ($\iota = \bar\alpha = 0$) carry the superscript 0. Write $u = \tau_T = \sum_{s<T}\eta_s$ for the total optimisation time, $\chi_s = \tau_T - \tau_s$ for the time remaining after step s, $S_\ell^{(s)} = \sum_j g_j^\ell \pi_j(T, s+1)^2$ for the exact spectral sums and $S^{\mathrm{c}}_\ell(\chi) = \sum_j g_j^\ell e^{-2g_j\chi}$ for their continuous versions; by Lemma E.2, $S^{\mathrm{c}}_\ell(\chi_s^+/(1-c_0)) \leq S_\ell^{(s)} \leq S^{\mathrm{c}}_\ell(\chi_s^+)$ with $\chi_s^+ = \chi_s - \eta_s$, so the two agree up to constants and a one-step shift. The

functional of §4.4 is

$$J[\eta] = \mathrm{Bias}_T + \frac{\sigma^2}{2B}\sum_{s<T}\eta_s^2\big[S_2^{(s)} + \iota_{\mathrm{bw},s}S_1^{(s)}\big], \qquad \mathrm{Bias}_T = \tfrac{1}{2}\sum_j g_j w_j \pi_j(T)^2, \tag{129}$$

and its continuous version is $J^{\mathrm{c}}[\eta] = \mathrm{Bias}^{\mathrm{c}}(u) + \frac{\sigma^2}{2B}\int_0^T \eta(t)^2 K(\chi(t))\,dt$ with $\mathrm{Bias}^{\mathrm{c}}(u) = \frac{1}{2}\sum_j g_j w_j e^{-2g_j u}$, the kernel $K = S_2^{\mathrm{c}} + \iota_{\mathrm{bw}} S_1^{\mathrm{c}}$ of (126), $\dot\chi = -\eta$ and $\chi(T) = 0$; $J^{\mathrm{c}} \asymp J$ by the comparison above. Set $F_K(u) = \int_0^u \sqrt{K(v)}\,dv$. On the bulk window $1 \ll \chi \ll M^a$, $S_\ell^{\mathrm{c}}(\chi) \asymp \chi^{1/a-\ell}$ (Lemma E.3), so $\sqrt{K(\chi)} \asymp \chi^{\frac{1}{2a}-1}(1+\kappa\chi)^{1/2}$ with κ as in (126), and $\mathrm{Bias}^{\mathrm{c}}(u) \asymp u^{-(b-1)/a}$ for $1 \ll u \ll M^a$ (Lemma E.1).

Lemma F.11 (cap stability; the functional is the loss up to a constant)**.** *Let $\eta_{\max} g_1 \le c_0 < 1$, $\delta_0 = 2 - (1+B^{-1})c_0$, and assume*

$$\beta_{\mathrm{cap}} := \frac{\eta_{\max}}{B\,\delta_0}\big[\theta_G + M\iota_{\mathrm{bw}} + 2\bar\alpha_{\mathrm{bw}} g_1\big] \le 1 - \kappa_{\mathrm{cap}} \qquad \textit{(H)}. \tag{130}$$

Then for every schedule as above $\sup_{t\le T}\bar{\mathsf{R}}_t \le R_{\mathrm{cap}} := (\sigma^2 + \sum_j g_j w_j)/(1-\beta_{\mathrm{cap}})$, and

$$J[\eta] \;\le\; \mathcal{E}_{\mathrm{DE}}^{\mathrm{last}}(T) \;\le\; C_\sigma\, J[\eta], \qquad C_\sigma = \frac{R_{\mathrm{cap}}}{\sigma^2}\max\Big\{2,\; 1 + \frac{2\bar\alpha_{\mathrm{bw}} g_1}{M\iota_{\mathrm{bw}}}\Big\} \tag{131}$$

(the second entry of the maximum is absent when $\iota_{\mathrm{bw}} = 0$). In particular $\inf_\eta J \le \inf_\eta \mathcal{E}_{\mathrm{DE}}^{\mathrm{last}} \le C_\sigma \inf_\eta J$, and a minimiser of J is within the factor C_σ of optimal for the dynamics.

Proof. Put $d_j(\eta) = (1-\eta g_j)^2 + \eta^2 g_j^2/B$ and $P_j(t,s) = \prod_{v=s}^{t-1} d_j(\eta_v)$. Unrolling (128), $\bar e_j(t) = w_j P_j(t,0) + \sum_{s<t} P_j(t,s+1)\frac{\eta_s^2}{B}[(g_j+\iota_{\mathrm{bw},s})\bar{\mathsf{R}}_s + 2\bar\alpha_{\mathrm{bw},s}\bar{\mathsf{S}}_s]$. Multiplying by g_j, summing over j and using $\bar{\mathsf{S}}_s \le g_1\bar{\mathsf{R}}_s/M$ gives $\bar{\mathsf{R}}_t \le \sigma^2 + \sum_j g_j w_j + \sum_{s<t} k(t,s)\bar{\mathsf{R}}_s$ with $k(t,s) = \frac{\eta_s^2}{B}\sum_j g_j\big(g_j + \iota_{\mathrm{bw}} + 2\bar\alpha_{\mathrm{bw}} g_1/M\big)P_j(t,s+1)$. Since $1 - d_j(\eta) = \eta g_j[2 - (1+B^{-1})\eta g_j] \ge \delta_0 \eta g_j$ for $\eta g_j \le c_0$, we have $\eta_s^2 g_j \le \frac{\eta_{\max}}{\delta_0}(1-d_j(\eta_s))$, and the telescoping identity $\sum_{s<t}(1-d_j(\eta_s))P_j(t,s+1) = 1 - P_j(t,0) \le 1$ bounds the row mass by $\sum_{s<t} k(t,s) \le \beta_{\mathrm{cap}}$. Induction on t gives $\bar{\mathsf{R}}_t \le \sigma^2 + \sum_j g_j w_j + \beta_{\mathrm{cap}} R_{\mathrm{cap}} = R_{\mathrm{cap}}$. For the comparison we use the exact identity $\mathcal{E}_{\mathrm{DE}}^{\mathrm{last}}(T) = \mathrm{Bias}_T + \frac{1}{2B}\sum_{s<T}\eta_s^2 \sum_j g_j \pi_j(T,s+1)^2\,\Xi_j(s)$ with $\Xi_j(s) = g_j(\bar{\mathsf{R}}_s + g_j\bar e_j(s)) + \iota_{\mathrm{bw},s}\bar{\mathsf{R}}_s + 2\bar\alpha_{\mathrm{bw},s}\bar{\mathsf{S}}_s$ (Lemma D.2). Positivity (Proposition D.1) gives $\Xi_j(s) \ge \sigma^2(g_j + \iota_{\mathrm{bw},s})$, which is the lower bound; $g_j\bar e_j(s) \le \bar{\mathsf{R}}_s$ and $\bar{\mathsf{S}}_s \le g_1\bar{\mathsf{R}}_s/M$ give $\Xi_j(s) \le R_{\mathrm{cap}}\big[2g_j + (1 + 2\bar\alpha_{\mathrm{bw}} g_1/(M\iota_{\mathrm{bw}}))\iota_{\mathrm{bw},s}\big]$, which with $R_{\mathrm{cap}} \ge \sigma^2$ for the bias term is the upper bound. □

No gain condition (the small-gain hypothesis of Theorem 4) enters: (H) is the row-mass form of the cap-stability hypothesis with the rounding sources added. With $\eta_{\max} = 0.5/g_1$ and the realised $\theta_G \approx 1.7$, $\beta_{\mathrm{cap}} \approx 0.23$ at $B = 8$ and 0.06 at $B = 32$ for $p_{\mathrm{bw}} = 0$; it fails at $B = 1$. Minimisers, switch times and per-bit constants are properties of J (or of J^{c}) and do not transfer to the dynamics through (131); the optimal loss does.

Lemma F.12 (the minimiser at fixed total time; the arc)**.** *Fix $u > 0$ and $T \ge u/\eta_{\max}$. Among measurable $\eta : [0,T] \to (0,\eta_{\max}]$ with $\int_0^T \eta = u$, the noise term of J^{c} is minimised by $\eta = \min\{\eta_{\max}, \sqrt{\lambda/K(\chi)}\}$ for the $\lambda = \lambda^*$ fixed by the step count. The cap is inactive iff $F_K(u) \le \eta_{\max} T\sqrt{K(u)}$; then*

$$\eta(\chi) = \frac{F_K(u)}{T\sqrt{K(\chi)}}, \qquad \frac{F_K(\chi(t))}{F_K(u)} = 1 - \frac{t}{T}, \qquad \frac{\sigma^2}{2B}\int_0^T \eta^2 K\,dt = \frac{\sigma^2}{2BT}F_K(u)^2. \tag{132}$$

On the bulk window $\eta \propto (T-t)^{2a-1}$ for $\chi \ll \chi_\times$ and $\eta \propto (T-t)^{(a-1)/(a+1)}$ for $\chi \gg \chi_\times$, and when $\kappa u \gg 1$ the share of the run spent at $\chi < \chi_\times$ is $F_K(\chi_\times)/F_K(u) \asymp (\kappa u)^{-(a+1)/(2a)}$.

Proof. With $dt = d\chi/\eta$ the noise term is $\int_0^u \eta K\,d\chi$ under the constraint $\int_0^u d\chi/\eta = T$. For $\lambda > 0$ and every admissible η, $\int_0^u \eta K\,d\chi = \int_0^u [\eta K + \lambda/\eta]\,d\chi - \lambda T \ge \int_0^u [\eta_\lambda K + \lambda/\eta_\lambda]\,d\chi - \lambda T$, where

$\eta_\lambda = \min\{\eta_{\max}, \sqrt{\lambda/K}\}$ minimises the convex function $\eta \mapsto \eta K + \lambda/\eta$ on $(0, \eta_{\max}]$ pointwise; equality holds at $\eta = \eta_\lambda$ when $H(\lambda) := \int_0^u d\chi/\eta_\lambda = T$. H is continuous and nonincreasing, tends to infinity as $\lambda \downarrow 0$ and equals $u/\eta_{\max}$ for $\lambda \geq \eta_{\max}^2 K(0)$ ($K(0) = \sum_j g_j^2 + \iota_{\rm bw} \sum_j g_j < \infty$), so λ^* exists for $T \geq u/\eta_{\max}$ and η_{λ^*} is a global minimiser at fixed (u, T). Since K decreases, the cap binds on an initial segment $\{\chi \geq \chi_{\rm s}\}$ with $K(\chi_{\rm s}) = \lambda^*/\eta_{\max}^2$, and it is inactive iff $\lambda^* \leq \eta_{\max}^2 K(u)$. In that case $T = H(\lambda^*) = \lambda^{*-1/2} F_K(u)$, so $\sqrt{\lambda^*} = F_K(u)/T$; $t(\chi) = \int_\chi^u dv/\eta(v) = T(1 - F_K(\chi)/F_K(u))$; and $\int_0^u \eta K \, d\chi = \sqrt{\lambda^*} F_K(u) = F_K(u)^2/T$, which is (132). For $K \propto \chi^\gamma$ the arc integrates to $\chi \propto (T-t)^{2/(2+\gamma)}$, $\eta \propto (T-t)^{-\gamma/(2+\gamma)}$ (the Euler–Lagrange computation of the details at the end of this subsection, now as a statement about the global minimiser): $\gamma = \frac{1}{a} - 2$ gives $2a - 1$ and $\gamma = \frac{1}{a} - 1$ gives $\frac{a-1}{a+1}$. For the share, $F_K(\chi_\times) \asymp \chi_\times^{1/(2a)}$ (the integrand is $\asymp \chi^{1/(2a)-1}$ below $\chi_\times$) while $F_K(u) \asymp \sqrt{\iota_{\rm bw}}\, u^{(a+1)/(2a)}$ for $\kappa u \gg 1$, and $\kappa \asymp \iota_{\rm bw}$ up to the constant c_a. □

The exact slope of the arc on any finite kernel is $d \log \eta / d \log(T-t) = -K'(\chi) F_K(\chi)/(2K(\chi)^{3/2})$; the bulk exponents are its limits, and on the realised spectrum at $M = 256$ it rises from 2.1 at $\chi = 10$ to 2.7 at $\chi = 10^3$ (fitted 2.07–2.17 on $\chi \in [1, 390]$), so the curves of Figure 5(c) are to be compared with this formula, not with the exponent 3. When the cap is active the minimiser is a constant phase followed by the same arc; for the exact-arithmetic kernel on the bulk window the constant phase has length $u - \chi_{\rm s}$ with $\chi_{\rm s} \asymp (\varrho\sigma^2)^{a/(2a-1)} u^{(a+b-1)/(2a-1)} = o(u)$, the schedule of Bordelon & Mori (2026).

Lemma F.13 (the optimum over all schedules is a scalar problem)**.** *For every admissible schedule with total optimisation time u, $\int_0^T \eta^2 K \, dt \geq F_K(u)^2/T$, with equality iff $\eta \propto K(\chi)^{-1/2}$. Hence*

$$\min_\eta J^{\rm c}[\eta] \;\geq\; m(T) := \min_{0 < u \leq \eta_{\max} T} \Big[\mathrm{Bias}^{\rm c}(u) + \frac{\sigma^2}{2BT} F_K(u)^2 \Big], \tag{133}$$

with equality whenever the minimising u_* satisfies the cap test of Lemma F.12; the minimiser is then the arc (132) *at u_*, whose peak is $\eta^* = F_K(u_*)/(T\sqrt{K(u_*)})$. On the bulk window with $\kappa u_* \gg 1$,*

$$u_*^{\frac{a+b}{a}} \asymp \frac{BMT}{q_{\rm bw} \operatorname{tr} \Sigma_d \, \sigma^2}, \qquad m(T) \asymp \Big(\frac{q_{\rm bw} \operatorname{tr} \Sigma_d \, \sigma^2}{BMT} \Big)^{\frac{b-1}{a+b}}, \qquad \eta^* \asymp \frac{u_*}{T}, \tag{134}$$

the cap test reads $T \gtrsim T_{\min} := (BM/(q_{\rm bw} \operatorname{tr} \Sigma_d \sigma^2))^{a/b} \eta_{\max}^{-(a+b)/b}$, rounding dominance $\kappa u_* \gtrsim 1$ reads $T \gtrsim (\sigma^2/B) \iota_{\rm bw}^{-b/a}$, and $u_* \leq M^a$ reads $T \lesssim T_{\max} := \sigma^2 q_{\rm bw} \operatorname{tr} \Sigma_d M^{a+b-1}/B$.

Proof. Since $d\chi = \eta \, dt$, $F_K(u) = \int_0^T \eta \sqrt{K(\chi(t))}\, dt \leq (T \int_0^T \eta^2 K \, dt)^{1/2}$ by Cauchy–Schwarz, with equality iff $\eta\sqrt{K}$ is constant in t; the arc of Lemma F.12 has this property and is admissible when the cap test holds, which gives (133) and the equality case. On the bulk window $F_K(u)^2 \asymp u^{1/a} + \iota_{\rm bw} u^{1+1/a}$ (integrate $\sqrt{K} \asymp \chi^{\frac{1}{2a}-1}(1 + \kappa\chi)^{1/2}$ and square), $\mathrm{Bias}^{\rm c}(u) \asymp u^{-(b-1)/a}$ and $\iota_{\rm bw} \asymp q_{\rm bw} \operatorname{tr} \Sigma_d / M$ (Lemma B.1); when the rounding term dominates the SGD term, $\kappa u_* \gtrsim 1$, minimising $u^{-(b-1)/a} + \frac{\sigma^2 \iota_{\rm bw}}{BT} u^{1+1/a}$ gives (134); the peak of the arc is at the start of the run, $\eta^* = F_K(u_*)/(T\sqrt{K(u_*)}) \asymp \frac{2a}{a+1} u_*/T$, and the three conditions are the cap test, $\kappa u_* \gtrsim 1$ and $u_* \leq M^a$ with u_* from (134). □

For fixed positive $(q_{\rm bw}, B, \sigma^2, \eta_{\max})$ and large M the window $[T_{\min}, T_{\max}]$ is nonempty and lies above the rounding-dominance threshold; $q_{\rm bw} = 0$ is the exact-arithmetic case treated below. At finite width $m(T)$ is a one-line quadrature on the realised spectrum, and it is the prediction to use where the bulk exponents are not yet accurate: at $(a, b) = (2, \frac{3}{2})$ the local exponent of $\mathrm{Bias}^{\rm c}$ in u is -0.34 at $k_*/M = 0.044$ and -0.76 at 0.31 against the bulk value -0.25, so the power laws need $k_*/M \lesssim 0.03$.

Derivation of (i)–(iv) of the proposition of §4.4. (i) is Lemma F.12 for the surrogate, transferred to the optimal loss of the dynamics by Lemma F.11. (ii) At fixed (M, T), $\eta^* \asymp u_*/T$ with u_* from (134) is the law; one bit less multiplies $q_{\rm bw}$ by four and η^* by $4^{-a/(a+b)}$. At a fixed sample budget $N = BT$ and fixed η, $k = (\eta N/B)^{1/a}$ and the balance $k_*^{a+b} \asymp MN/(q_{\rm bw} \operatorname{tr} \Sigma_d \sigma^2)$

gives $B^* \asymp \eta N(q_{\rm bw}\,\mathrm{tr}\,\Sigma_d\sigma^2/(MN))^{a/(a+b)} \propto q_{\rm bw}^{a/(a+b)}$: the optimal batch rises by the same factor. The conditions are those of Lemma F.13: rounding dominant at u_*, i.e. $\kappa\tau_T \gtrsim 1$ or $p_{\rm bw} \lesssim \frac{1}{2}\log_2(\tau_T/M) + \frac{1}{2}\log_2(0.09c_a\,\mathrm{tr}\,\Sigma_d)$, and the cap inactive. (iii) The loss is $m(T)$ of (134) on $[T_{\min}, T_{\max}]$, a slope $(b-1)/(a+b)$ in $\log T$ on that finite range. In exact arithmetic ($\iota_{\rm bw} = 0$) the scalar problem has $u_* \propto T^{a/b} > \eta_{\max}T$ for large T, so the cap binds and $\min J_0^{\rm c} \asymp \mathrm{Bias}^{\rm c}(\eta_{\max}T) \asymp (\eta_{\max}T)^{-(b-1)/a}$ for $\eta_{\max}T \ll M^a$; the two ranges overlap for large M (at $(2, \frac{3}{2})$: $T_{\min} \asymp M^{4/3}$ against the exact-arithmetic window's $O(M^2)$), and the slopes are compared there. For the bit condition, the rounding term of any schedule under the cap is at most $\frac{\sigma^2\iota_{\rm bw}}{2B}\eta_{\max}\int_0^u S_1^{\rm c} \asymp \varrho_{\max}\iota_{\rm bw}\sigma^2 k_u$ (Lemma E.3); it is below the bias k_u^{1-b} at the cap iff $p_{\rm bw} \geq \frac{b}{2a}\log_2\tau_T - \frac{1}{2}\log_2 M + O(1)$, the $O(1)$ carrying $\varrho_{\max}\,\mathrm{tr}\,\Sigma_d\sigma^2$; this is sufficient, not necessary. (iv) The allocation (15) with the weights (113) computed from the exact-arithmetic trajectory, $W_{\rm bw}(s) = \eta_s^2\mathsf{R}_s^0 S_1^{(s)}$ and $W_{\rm fw}(s) = \eta_s^2\mathsf{U}_s^0 S_2^{(s)}$, gives on the bulk window $p_{\rm bw}(s) = c + \log_2\eta_s + \frac{1}{2}\log_2\mathsf{R}_s^0 - \frac{a-1}{2a}\log_2\chi_s + \epsilon_1(s)$ and $p_{\rm fw}(s) = c' + \log_2\eta_s + \frac{1}{2}\log_2\mathsf{U}_s^0 - \frac{2a-1}{2a}\log_2\chi_s + \epsilon_2(s)$, where ϵ_ℓ is the bounded deviation of $S_\ell^{(s)}$ from its power law. For a linear cooldown $\eta_t = \eta_{\rm fin} + d(T-t)$ the continuous clock satisfies $\eta^2 = \eta_{\rm fin}^2 + 2d\chi$ (up to one cell for the discrete sum), so $\log_2\eta_s = \frac{1}{2}\log_2\chi_s + O(1)$ for $\chi_s \gg \chi_{\rm fl} := \eta_{\rm fin}^2/(2d)$: with $\mathsf{R}_s^0 \approx \sigma^2$ the backward allocation falls by $\frac{1}{2a}$ bit per halving of χ_s, and the forward one rises by $\frac{a-1}{2a}$ bit per halving plus $\frac{1}{2}\Delta\log_2\mathsf{U}_s^0$. Writing $\mathsf{U}_s^0 = D_s + n^0(s)$ with $D_s = \sum_l w_l(1-\pi_l(s))^2$ nondecreasing and $n^0 \geq 0$ the exact-arithmetic noise energy (Lemma D.2), Lemma F.11 with $\iota = \bar\alpha = 0$ gives $n^0(s) \leq \frac{2R_{\rm cap}}{B}\sum_{v<s}\eta_v^2 S_1^{(v)} \lesssim \varrho_{\max}R_{\rm cap}k_s$ at every time under the cap, while $D_s \asymp k_s^{a-b+1}$; hence over any cooldown $[s_1, T]$, $\frac{1}{2}\Delta\log_2\mathsf{U}^0 \geq -n^0(s_1)/(2\ln 2\, D_{s_1}) \gtrsim -\varrho_{\max}R_{\rm cap}k_{s_1}^{b-a}$, and the forward increment is at least $\frac{a-1}{2a}$ per halving up to this total, plus the nonnegative contribution of D (up to $\frac{a-b+1}{2a}$ bit over the whole cooldown when $k < M$). The bound is one-sided.

Remarks. (a) *Matched width.* If the width is scaled with the run, $\tau_1 = M^a$ as in Proposition E.10, the same balance gives $\eta^* \propto q_{\rm bw}^{-a/(a+b-1)}$, a factor $4^{a/(a+b-1)} \approx 3.0$ per bit, and the time exponent $(b-1)/(a+b-1)$. (b) *Linear cooldowns to a terminal rate $\zeta\eta_{\rm peak}$.* The cooldown leaves an exact-arithmetic floor $N_{\rm fl} = \zeta\varrho_{\rm peak}\sigma^2\theta_G/4$, the small-step limit of the stationary excess risk $\sigma^2\beta_0/(2(1-\beta_0))$, $\beta_0(h) = \frac{h}{B}\sum_j g_j/(2-(1+B^{-1})hg_j)$, at the terminal rate; against the retained rounding term $P \asymp \varrho\iota_{\rm bw}\sigma^2 k_*$ this is $P/N_{\rm fl} \approx 0.36\,q_{\rm bw}k_*/(\zeta M)$. The $q_{\rm bw}$-elasticity of the optimal peak is then $-\gamma_P P/(\beta^2\mathrm{Bias} + \nu^2 N + \gamma_P^2 P)$ with the local exponents $\beta = \frac{b-1}{a}$, $\nu = 1$, $\gamma_P = 1 + \frac{1}{a}$ of the three terms in η at the optimum, which reaches $-a/(a+b)$ only for $P \gg N_{\rm fl}$; at $k_*/M \approx 0.05$ this needs $\zeta \lesssim 10^{-3}$. (c) *Algorithm.* Orders of the loss transfer through Theorem E.12 given (H) for both comparison copies; minimisers, switch times and per-bit constants do not, and nothing is claimed for the forward clause of (iv) beyond the reduced system. (d) *Numerical status.* On the realised feature map at $M = 256$: Lemma B.1's profile $\iota_{\rm bw}M/(q_{\rm bw}\theta_G) = 0.082$–$0.090$ for $p_{\rm bw} = 0$–4; the stationarity equation of the capped branch holds to 6%; the switch moves earlier ($t_{\rm s}/T$ from 0.914 to 0.888 at $p_{\rm bw} = 0$); the clock slopes on the bulk window are $+0.21$ (backward) and -0.265 (forward) per doubling of χ against ± 0.25. The peak-rate sweep of that run ($B = 1$, $k_*/M \approx 0.3$) lies outside (H) and outside the bulk and is reproduced by the elasticity formula of Remark (b) with the measured exponents to within 0.03. The balance (134) itself has not been checked in regime; a surrogate at large M or the dynamics at $B = 8$ would do it.

Further details for §4.4. *Stationarity.* For $\int \dot z^2 K(\chi_0 - z)\,dt$ with $\chi = \chi_0 - z$, $\dot z = \eta$, the Euler–Lagrange equation is $2\ddot z K - 2\dot z^2 K' = -\dot z^2 K'$, i.e. $\frac{d}{dt}\log\dot z = -\frac{1}{2}\frac{d}{dt}\log K(\chi)$, so $\eta \propto K(\chi)^{-1/2}$; with $K \propto \chi^\gamma$ and $\dot\chi = -\eta$ this integrates to $\chi \propto (T-t)^{2/(2+\gamma)}$ and $\eta \propto (T-t)^{-\gamma/(2+\gamma)}$: $\gamma = \frac{1}{a} - 2$ gives $(T-t)^{2a-1}$ (the decay of Bordelon & Mori, 2026), $\gamma = \frac{1}{a} - 1$ gives $(T-t)^{(a-1)/(a+1)}$. The crossover $\chi_\times$ is where $\iota_{\rm bw}S_1(\chi) = S_2(\chi)$; at $M = 64$ the reduced system gives $\chi_\times = 857$ and 152 for $p_{\rm bw} = 1$ and 0 against a cooldown start at $\chi_s \approx 150$–280, which is why only $p_{\rm bw} = 0$ changes the profile at that width. *Numerically optimal schedules* (Figure 5(c)): at $M = 64$, $B = 8$, $\sigma^2 = 1$ the reduced system's optimal last-iterate schedule starts decaying at $t/T = 0.88$ without rounding and 0.80 at $p_{\rm bw} = 0$, and its penalty is 26% where the constant-rate penalty was 7%. *Peak rate.* Minimising $(\eta T)^{-(b-1)/a} + c\,q_{\rm bw}\sigma^2 c(\phi,\zeta)\,\eta/B$ in η gives $\eta^*_{\rm peak} \propto (B/(q_{\rm bw}\sigma^2 c(\phi,\zeta)))^{a/(a+b-1)}T^{-(b-1)/(a+b-1)}$ below the stability cap; in the reduced

system at $B = 1$, $\sigma^2 = 1$ the optimal peak rate is $\eta^*/\eta_{\max} = 0.7, 0.5, 0.35$ at $p_{\rm bw} = 2, 1, 0$, while at $B = 32$, $\sigma^2 = 0.1$ the minimiser lies above the stability cap for every $p_{\rm bw} \geq 0$ (no shift). *Exponents.* With $\eta_{\max}$ free and fixed bits the loss is $(\eta_{\max} T)^{-(b-1)/a} + C q_{\rm bw} \eta_{\max}^{1+1/a} T^{1/a} \operatorname{tr} \Sigma_d \sigma^2/(BM)$; the minimiser $\eta_{\max} \propto T^{-b/(a+b)}$ gives $T^{-(b-1)/(a+b)}$, and the rounding term stays below the bias iff $q_{\rm bw} \lesssim M \tau_T^{-b/a}/(\eta_{\max} \operatorname{tr} \Sigma_d)$, the slope $\frac{b}{2a}$ of Corollary F.2. *Clock rule.* It is the Lagrangian allocation (15) with the weights (113), $S_1 \asymp \chi^{1/a-1}$, $S_2 \asymp \chi^{1/a-2}$: $p_{\rm bw}(t) = c_{\rm bw} + \log_2 \eta_t + \frac{a-1}{2a} \log_2(1/\chi_t)$ and $p_{\rm fw}(t) = c_{\rm fw} + \log_2 \eta_t + \frac{1}{2} \log_2 \mathsf{U}_t + \frac{2a-1}{2a} \log_2(1/\chi_t)$. Under a linear cooldown $\chi \propto (T-t)^2$ and $\eta \propto \chi^{1/2}$, so $p_{\rm bw}$ falls by $\frac{1}{2a}$ bit per halving of χ while $p_{\rm fw}$ rises; under the polynomial decay $\chi \propto (T-t)^{2a}$ and $\eta \propto \chi^{(2a-1)/(2a)}$, which makes $p_{\rm fw}$ flat and $p_{\rm bw}$ fall by $\frac{1}{2}$ bit per halving of χ.

F.7 Algorithm 1: a closed-form precision schedule

Algorithm 1 turns Corollary F.1, Proposition F.6 and Theorem F.8 into a procedure, which the main text names in one sentence (§4.3): solve the exact-arithmetic dynamics once, form the response weights, choose the mean bits of each location from the tolerance or the compute, and distribute them over time by (15); whether it transfers to networks is not a consequence of the theory.

Algorithm 1 (closed-form precision schedule; the constructive form of Corollary F.1, Proposition F.6 and Theorem F.8)

Input: a, b, σ^2, M, B, T, the step-size schedule (η_s), the tolerance ϑ (or the budget f), the format menu.

1. Solve the exact-arithmetic deterministic-equivalent dynamics (10) once for $\mathsf{U}_s, \mathsf{R}_s, \chi_s$.
2. Form the response weights (113) for each location and the chosen output rule.
3. Mean bits per location: $\bar{p}_r = \frac{1}{2} \log_2\big(c_r \sum_s W_r(s)/(\vartheta\, \mathcal{E}_0(T))\big)$, $c_r = 0.09 \operatorname{tr} \Sigma_d/(2BM)$, plus the joint constraint of (110) for the average; under a compute budget use (116) instead.
4. Allocate over time by (15); round to the format menu with the mean re-fitted; cast the partners to the operands' widths.
5. Calibrate c_r once at small M against the exact recursion (9).

Output: the schedule $(p_{{\rm fw},s}, p_{{\rm bw},s})_{s<T}$. Step 4 is optimal in the first-order-response model; the joint optimization of (η_s, p_s) by automatic differentiation through (10) is Algorithm B (§F.8). Applying the algorithm to a network requires (a, b, σ^2) for the layer at hand (from its feature spectrum and the slope of its exact-arithmetic loss curve) and the calibration of step 5 on a small-width run of that network; whether the resulting schedule transfers is the subject of the neural-network tests (a two-layer MLP on CIFAR-10 and a small transformer), not of the theory above.

F.8 Joint optimisation of learning-rate and precision schedules

The reduced system (10) is affine in the coefficients $(\iota, \bar{\alpha}, \gamma)$, which are linear in $q = 4^{-p}$ up to the profile of Lemma B.1, so its outputs are differentiable in $(\log \eta_t, p_{{\rm fw},t}, p_{{\rm bw},t})$ and reverse-mode automatic differentiation through the recursion costs one forward pass. Minimising $\mathcal{E}_{\rm DE}^{\rm out}(T) + \lambda \sum_t [w(p_{{\rm fw},t}) + w(p_{{\rm bw},t})]$ jointly over the two schedules, parametrised on a grid dense in $\log(T-t)$, gives in one run the joint optimum of Sections 4.3–4.4; this is what Bordelon & Mori (2026) do numerically for the learning rate alone, and it is how panel (c) of Figure 5 was produced (160 knots, 120 Adam steps, 6 seconds per gradient at $M = 64$ on a CPU). It is not a theorem—the objective is the reduced system and the relaxation is continuous—but it is a computable reference against which Algorithm 1 can be scored, and rounding its solution to a format menu and re-evaluating it with the exact recursion gives a certified value for the actual algorithm.

G EXPERIMENTAL DETAILS

Common model and programs. Unless stated otherwise: $a = 2$, $b = \frac{3}{2}$, $\sigma^2 = 0.1$, $d = 4M$, one feature map per width, $B = 32$, $\eta = 0.5/g_1$, $\tau_T = M^a$; precisions are written $(p_{\mathrm{fw}}, p_{\mathrm{bw}})$. $\lambda_j = j^{-a}$, $\mathbb{E}[\lambda_j \theta_j^2] = j^{-b}$; the feature map has i.i.d. $N(0, 1/M)$ entries with a fixed seed per width (seed 1 unless stated). Recursion runs use the target-averaged problem (Corollary B.4); Monte Carlo runs use one realised target and compare with the recursion on the same problem. The exact recursion (9) is solved in float64 with the averaged output by the formula of Corollary B.8; the numpy and PyTorch implementations agree to 10^{-15} on CPU. The reduced system (10) is solved with $\iota, \bar{\alpha}, \gamma_{\mathrm{fwbw}}$ of (26) computed from the same G. The bit-exact trainer resolves the rounding interval of every float64 operand exactly (powers of two) and draws one uniform random variable per entry; the only randomness is the data and the rounding decisions. At $\eta = 0.5/g_1$ the compute-optimal horizon is $T = M^a/\eta = 2g_1 M^a$ steps. The exact-recursion runs at $M \geq 256$ and those with three feature maps (seeds 1–3), the rounded-SGD runs at $M = 8$, 32, 128 and 512, the compute-law runs and the network runs used one NVIDIA A100 GPU each (PyTorch, float64 for the recursion, the reduced system and the trainer); every run folder records its command, code version and batch-job identifier.

Panels of the main-text figures. Every panel of Figures 1, 3, 2, 4, 5 and 7 is drawn from a run described in this appendix, with the common settings above unless stated ($a = 2$, $b = \frac{3}{2}$, $\sigma^2 = 0.1$, $d = 4M$, $B = 32$, $\eta = 0.5/g_1$, $\tau_T = M^a$; three feature maps at $M \leq 512$ and one at $M = 1024$). Figure 1: (a) the penalty shares of the last iterate at one bit, medians over the feature maps (Appendix G); (b) the A/B ratios of the last iterate at $M \in \{64, 128\}$, precision switched at $T/2$ (Appendix G); (c) the bits needed p^ϑ at $\vartheta \in \{10\%, 1\%, 0.1\%\}$ from the $(1, 1)$ runs at $M = 16$–512 (Appendix G). Figure 3: (a) the $M = 512$ Monte Carlo comparison (Appendix G); (b) the maximum over $\tau_t \in [M^a/10, M^a]$ of the relative error of the predicted penalty, $M = 16$–1024 (Appendix G); (c) the reduced system and the exact recursion at $M = 1024$, $p = 1$ (Appendix G). Figure 4: (a) the ratio r at $p = 1$ from the recursion, the reduced system and the $M = 128$, 512 Monte Carlo runs; (b) the interaction term of the average at $(a, b) \in \{(2, \frac{3}{2}), (2, \frac{5}{4}), (\frac{5}{2}, \frac{3}{2})\}$, normalized at $M = 32$, with the fitted exponents 0.48, 0.70, 0.93 in the legend (Appendix G); (c) the two ratios of Proposition 5(ii) at $p = 1$ and $p = 2$ with one fitted constant each (Appendix G). Figure 2: (a) the forward-to-backward ratio of the last iterate at $p = 1, 2$, $M = 32$–1024, with the fitted constant 1.43 (Appendix G); (b) the A/B runs at $M \in \{64, 128\}$, precision switched at $T/2$, cooldown $\phi = 1$ (Appendix G); (c) the compute-law runs at $M = 16$–512 (Appendix G). Figure 5: (a) the time-allocation runs at $M = 64$ (Appendix G); (b) the cooldown runs at $p_{\mathrm{bw}} = 2$, $p_{\mathrm{fw}} = \infty$, $\zeta = 0.02$, $M = 32$–512 (Appendix G); (c) the optimal-schedule run at $M = 64$, $B = 8$, $\sigma^2 = 1$, $p_{\mathrm{bw}} \in \{\infty, 1, 0\}$ (Appendix G). Figure 7: the transformer and MLP runs of Appendix G, with the 10% and 50% cooldowns and, in (c), the checkpoint continuations.

Rounding-error moments. $m_p(g)$, $\alpha_p(g)$ and the Hermite coefficients $a_n(p, g)$ are computed interval by interval over all binades carrying Gaussian mass (16-point Gauss–Legendre per interval; relative error $\leq 10^{-14}$ against a 40-digit multiprecision evaluation). The cross-covariances $c_{ab}(G_{kk}, G_{mm}, \rho_{km})$ use the Mehler series to degree 128 (degree 40 in the CPU runs, which agree with the GPU runs to every digit reported here) for $|\rho| \leq 0.9$ and an exact conditional quadrature above; the $k = m$ term is the exact same-variable moment. For input rounding the cross coefficients $c(\lambda_j, \hat{G}_{kk}, \varkappa_{jk})$ use the same rule.

Exact recursion against rounded SGD. Small widths: $M = 6$–8, $d = 4M$, $B \in \{3, 4\}$, $\eta = 0.4/g_1$–$1/g_1$, $T = 100$–300, configurations $(1, 1), (0, 2), (2, 0), (\infty, 1)$, 2×10^4–4×10^4 runs; largest $|\mathrm{recursion} - \mathrm{MC}|/\mathrm{s.e.}$ 2.5. With 2×10^5 runs at $M = 8$ and 32 ($B = 4$, $\eta = 0.4/g_1$, $T = 300$, configurations $(1, 1), (2, 2), (1, \infty), (\infty, 1)$) the largest $|z|$ is 3.1 over 192 comparisons; the comparisons are correlated, since all configurations share one stream of data. At the scale of the paper ($B = 32$, $\eta = 0.5/g_1$, $\tau_T = M^a$, realised target of feature map 1, 40 log-spaced checkpoints; the trainer samples the features directly, $z \sim N(0, G)$ and $y = z^\top v_*^S + N(0, \sigma_{\mathrm{eff}}^2)$, which is the exact joint law): $M = 128$ with 16 384 runs (31 715 steps) of $(1, 1), (1, \infty), (\infty, 1), (\infty, 2)$, exact arithmetic, and the cooldown schedule below with $\phi = 0.2$ and 2, largest $|z| = 3.0$ over 692 comparisons; $M = 512$ with 2048 runs (5.0×10^5 steps, Figure 3(a)) of the same five constant-rate configurations, largest $|z| = 2.5$ over 390. At $M = 512$ the Monte Carlo resolves the penalty of $(1, 1)$: $(5.47 \pm 0.18) \times 10^{-3}$ against 5.50×10^{-3} for the last iterate and $(2.12 \pm 0.07) \times 10^{-5}$ against

2.09×10^{-5} for the average (penalty from the rounded and the exact-arithmetic runs, standard errors added in quadrature). Exact enumeration of all 4096 rounding outcomes of one batch step ($B = 3$, $M = 2$) against the batch map of Lemma B.6: 1.3×10^{-14}; enumeration of all 16 384 outcomes of one independent-copy step ($M = 3$, $d = 4$, $p_x = p_{\mathrm{fw}} = p_{\mathrm{bw}} = 1$) against Lemma C.2: 5×10^{-15}. A further run at $M = 6$, $d = 24$, $B = 4$, $\eta = 0.4/g_1$, $T = 120$, 4000 runs, configurations $(1, 1)$ and $(\infty, 1)$, gives the largest $|\mathrm{recursion} - \mathrm{MC}|/\mathrm{s.e.}$ 1.2 for $(1, 1)$ and 2.3 for $(\infty, 1)$.

Deterministic-equivalent dynamics against the exact recursion. $B = 32$, $\eta = 0.5/g_1$, trajectories recorded to $2M^a/\eta$ (120 log-spaced checkpoints), configurations $(2, 2)$, $(1, 1)$, $(1, \infty)$, $(\infty, 1)$ and exact arithmetic, $M = 16, 32, 64, 128, 256, 512$ (feature maps 1–3) and $M = 1024$ (feature map 1 only, so the last point of Figure 3(b) rests on one map). Reported: the maximum over the saved checkpoints with $\tau_t \in [M^a/10, M^a]$ (37 checkpoints at $M = 16$, 18 at $M = 1024$) of $|\Delta_{\mathrm{DE}} - \Delta|/\Delta$ for last iterate and average. Feature map 1, $M = 16, \ldots, 1024$. Last iterate: $(2, 2)$: $13.10, 7.75, 2.46, 2.36, 0.62, 0.67, 0.39\%$; $(1, \infty)$: $14.87, 7.53, 2.49, 2.48, 0.62, 0.73, 0.45\%$; $(\infty, 1)$: $6.85, 9.24, 2.73, 1.14, 0.73, 0.21, 0.09\%$; $(1, 1)$: $13.00, 7.88, 2.59, 2.34, 0.76, 0.50, 0.45\%$. Average: $(2, 2)$: $9.65, 5.74, 2.58, 1.47, 0.50, 0.50, 0.41\%$; $(1, \infty)$: $24.01, 12.65, 5.28, 1.83, 0.26, 0.55, 0.37\%$; $(\infty, 1)$: $11.93, 8.56, 3.35, 1.50, 0.73, 0.56, 0.41\%$; $(1, 1)$: $10.01, 5.75, 2.73, 1.77, 0.41, 0.49, 0.39\%$. Over the three feature maps, the median over the eight (configuration, output) pairs is $12.5, 7.6, 3.2, 1.8, 0.62, 0.50, 0.40\%$ and the largest value $28.9, 12.7, 6.2, 3.5, 1.07, 0.73, 0.45\%$. The decay is close to M^{-1} up to $M = 256$; beyond $M = 256$ the local slopes of the forward-rounded configurations lie between -0.04 and -0.4, and the error is of the order of its spread over feature maps (0.19–0.73% at $M = 512$). At $M \le 256$ the dominant finite-M correction is the correlation between $(\bar{W}_t)_{kk}$ and $m_p(G_{kk})$ in the forward rounding variance. The backward-only configuration, whose law is Theorem E.12, keeps decaying: for the last iterate at $\tau_T = M^a$ its signed error on feature map 1 is $-0.53, -0.35, -0.20, -0.10, -0.04\%$ at $M = 64, \ldots, 1024$. On the risk scale, $|\mathcal{E}_{\mathrm{DE}} - \mathcal{E}|$ is at most 0.13% of the exact-arithmetic risk; the excess risk itself agrees within 1% at $p = 2$.

Joint term and cooldown. Joint term: $p = 1$ and $p = 2$, $\tau_T = M^a$, $B = 32$, $\eta = 0.5/g_1$; penalties $\Delta(p, p), \Delta(p, \infty), \Delta(\infty, p)$ from the exact recursion and the reduced system at T; ratio $r = \Delta(p, p)/[\Delta(p, \infty) + \Delta(\infty, p)]$ and interaction $I = \Delta(p, p) - \Delta(p, \infty) - \Delta(\infty, p)$ for both outputs; feature maps 1–3 at $M \le 512$ and 1–2 at $M = 1024$ (means reported; the last point of Figure 4(a) rests on two maps). At $p = 1$ and $M = 32, 64, 128, 256, 512, 1024$ the ratio is $1.15, 1.25, 1.44, 1.76, 2.37, 3.39$ for the average (reduced system within 0.01; 1.05 at $M = 16$ on feature map 1) and $1.017, 1.017, 1.018, 1.020, 1.023, 1.029$ for the last iterate. The interaction of the average grows like $M^{0.48}$ (local slopes 0.44–0.54; reduced system 0.46), against M^{a-b} from the interaction row $q_{\mathrm{fw}} q_{\mathrm{bw}} k_T^{a-b+2}/M^2$ of Theorem 4. At $(a, b) = (2, \frac{5}{4})$ ($M = 32$–256) and $(\frac{5}{2}, \frac{3}{2})$ ($M = 32$–128) the fitted exponents are 0.70 and 0.93 against 0.75 and 1 (0.45 for $(2, \frac{3}{2})$ over $M = 32$–256), and the forward term of the average follows M^{1-b} (-0.23 and -0.50). From $p = 1$ to $p = 2$ ($M = 128$–512) the single-copy penalties fall by 3.92 and the interaction of the average by 15.4, against 4 and 16. For the last iterate $r - 1$ is 4.2 to 4.8 times smaller at $p = 2$, a term of order q_{bw}, and its interaction grows with local slopes rising from 0.41 to 0.86, as the term $\varrho^2 q_{\mathrm{fw}} q_{\mathrm{bw}} r_t \tau_t / M$ discussed after Theorem 4 predicts. From rounded SGD alone (above), $r = 1.42 \pm 0.10$ at $M = 128$ and 2.40 ± 0.15 at $M = 512$ for the average, against 1.39 and 2.38 from the recursion on the same realised problem. At one bit the whole penalty of the average stays below 0.1% of its exact-arithmetic risk up to $M = 1024$ (Appendix G), so the interaction decides the penalty of the average but not yet its loss. Cooldown: $p_{\mathrm{bw}} = 2$, $p_{\mathrm{fw}} = \infty$, $\eta_{\mathrm{peak}} = 0.5/g_1$ for $T_1 = M^a/\eta_{\mathrm{peak}}$ steps, then a linear decay to $\eta_{\mathrm{peak}}/50$ over T_c steps with $\sum_{s \ge T_1} \eta_s = \phi\, \eta_{\mathrm{peak}} T_1$, $\phi \in \{0.05, 0.1, 0.2, 0.5, 1, 2\}$; retained fraction $= [\mathcal{E}_{p_{\mathrm{bw}}}(T) - \mathcal{E}_{\mathrm{clean}}(T)]/[\mathcal{E}_{p_{\mathrm{bw}}}(T_1) - \mathcal{E}_{\mathrm{clean}}(T_1)]$ of the last iterate. At $M = 512$ the retained fraction is $0.738, 0.662, 0.572, 0.443, 0.347, 0.260$ for $\phi = 0.05, \ldots, 2$ (feature map 1), against the leading curve $0.751, 0.668, 0.568, 0.422, 0.317, 0.228$ of Proposition E.10(iii) with $\zeta = 0.02$: the recursion lies 0.013 below the curve for the shortest cooldown and 0.032 above it for the longest. The mean over feature maps 1–3 is 0.69–0.23 at $M = 32$, 0.72–0.25 at $M = 64$ and 0.73–0.25 at $M = 128$, and the retained fraction changes by less than 0.007 between $M = 128$ and 512. The reduced system gives $0.741, 0.664, 0.574, 0.444, 0.348, 0.261$ at $M = 512$ and is flat in M within 0.008 from $M = 32$ to 512. Over the same cooldowns the exact-arithmetic error $\frac{1}{2}\|v_T - v_*^S\|_G^2$ of the last iterate falls to 0.47 ($\phi = 0.05$) down to 0.043 ($\phi = 2$) of its value at T_1 ($M = 512$): a

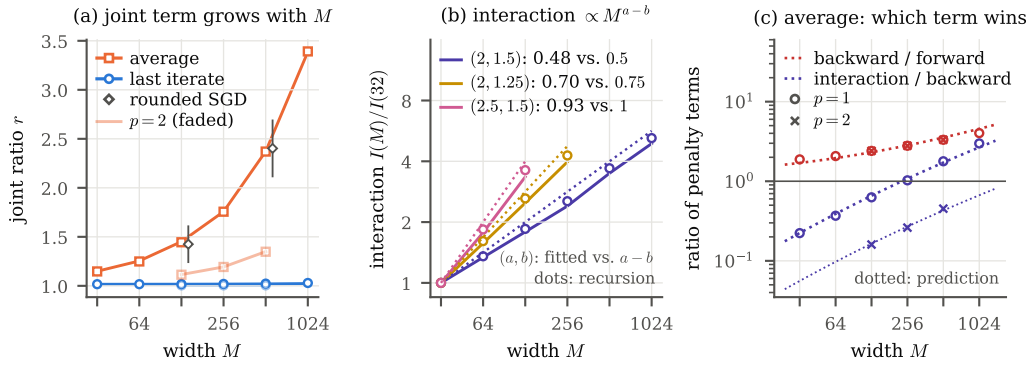


Figure 4: Joint term and interaction, continuing Figure 3. (a) Ratio r of the penalty with both copies rounded to the sum of the single-copy penalties at $\tau_T = M^a$, at $p = 1$ ($p = 2$ faded): exact recursion (markers), DE dynamics (lines) and, for the average, rounded SGD (diamonds, ± 2 s.e.). (b) Interaction term of the average, normalized at $M = 32$, for three pairs (a, b): recursion (dots), dynamics (lines) and M^{a-b} (dotted); the legend gives the fitted exponent. (c) Average: backward over forward term and interaction over backward term at $p = 1$ and $p = 2$, against the predictions of Proposition 5(ii) (dotted).

cooldown removes SGD noise much faster than rounding noise. Relative to the full exact-arithmetic risk, including A_S, the cooled backward penalty at $p_{\text{bw}} = 2$ is 0.005–0.013% at $M = 512$. Rounded SGD at $M = 128$ matches the recursion on these schedules at every recorded time, but the retained fraction itself is below its resolution (the penalty at T_1 is about 9×10^{-6}, the standard error of a difference of two Monte Carlo means about 2×10^{-5}).

Timing and time allocation. $M \in \{64, 128\}$, $B = 32$, $\eta = 0.5/g_1$, stable phase $T = M^a/\eta$, cooldown of the same optimisation time ($\phi = 1$) linear to $\eta/50$. A/B: precision switched at $T/2$ and the second value held through the cooldown; forward test at $p_{\text{bw}} = 4$, backward test at $p_{\text{fw}} = \infty$; low/high $= 1/3$, constant comparator 2. Time allocation: the schedules (15) at $\bar{p} = 2$ from the response weights (113), built from the exact-arithmetic trajectory of the exact recursion and the spectrum, rounded to integers in $[0, 12]$ with the mean re-fitted to 2.00 ± 0.01; constant-rate phase only; comparator: constant 2 on the same location. A/B ratios (high-early over low-early): forward copy $15, 1.2, 15$ ($M = 64$) and $15, 1.2, 15$ ($M = 128$) for last, average, cooled; backward copy $5.5, 0.33, 8.2$ ($M = 64$) and $5.6, 0.34, 8.7$ ($M = 128$). Time-allocation ratios at $M = 64$: measured $0.002, 0.474, 0.734, 0.819$ against predicted $0.006, 0.469, 0.646, 0.838$ for (forward, last), (backward, last), (backward, average), (forward, average), the last three within 0.09 of the prediction; the integer schedules span $p \in [0, 11]$, $[1, 6]$, $[0, 4]$, $[0, 3]$. The $M = 128$ time-allocation units are reported in the run folder. In summary, 99% of the forward weight of a last iterate sits in the last 10% of the steps, so the schedule at mean two bits ranges over eleven bits and removes 99.8% of the forward penalty at equal cost; the backward weight is spread, and the gain is $2\times$ for the last iterate (bits late) and $1.4\times$ for the average (bits early), so the two output rules want opposite backward schedules ($1 \to 3$ bits at mid-run beats $3 \to 1$ by $5.5\times$ for the last iterate and loses by $3\times$ for the average). Elapsed-time ramps ($p \propto \log_2 \tau_t$) gain nothing: the right clock is the remaining time.

Time exponents. $p = 1$, $B = 32$, $\eta = 0.5/g_1$, $\tau_T = M^a$, 300 log-spaced checkpoints; configurations exact arithmetic, $(1, \infty)$, $(\infty, 1)$, $(1, 1)$; interaction penalty $= \Delta(1, 1) - \Delta(1, \infty) - \Delta(\infty, 1)$; least-squares slopes of log penalty against $\log \tau_t$. Exact recursion at $M = 64, 128$ (window $\tau_t \in [M^a/100, M^a/3]$) and reduced system with the coefficients of the realised feature map at $M = 512, 1024$ and $\sigma^2 \in \{0.1, 1\}$ (window $\tau_t \in [10^{-3}, 10^{-2}]M^a$, i.e. $k_t \in [0.03M, 0.1M]$). Slopes (forward-last, interaction-average, forward-average, backward-last): exact $M = 64$: $0.70, 1.02, 0.13, 0.13$; exact $M = 128$: $0.69, 1.03, 0.15, 0.16$; reduced $M = 128$: $0.71, 1.07, 0.17, 0.17$; reduced $M = 512$: $0.75, 1.16, 0.20, 0.26$; $M = 1024$: $0.74, 1.19, 0.22, 0.28$; at $\sigma^2 = 1$: $0.74, 1.11, 0.20, 0.40$ ($M = 512$) and $0.74, 1.15, 0.22, 0.42$ ($M = 1024$); exact recursion at $M = 1024$ on the same feature map (the trajectories of the comparison above): $0.73, 1.15, 0.21, 0.27$, within 0.04 of the reduced system. At $M \leq 128$ the learned-mode count $\sum_j (1 - r_j^{2t})$ is already saturating towards M inside the window and the residual is still falling, which

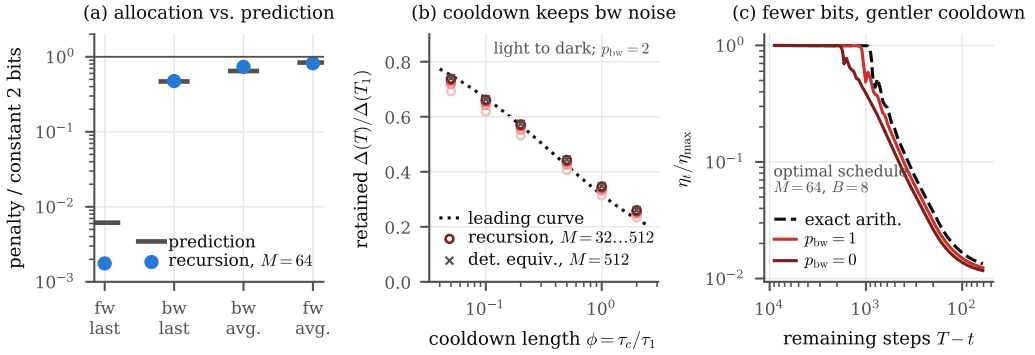


Figure 5: Allocation, cooldown and learning rate, continuing Figure 2. (a) The allocation (15) at $\bar{p} = 2$: measured penalty relative to a constant two bits, against the prediction, for the four copy–output pairs at $M = 64$. (b) Backward penalty retained after a linear cooldown of relative length ϕ: recursion at $M = 32$–512 (light to dark), DE dynamics at $M = 512$ (crosses), leading curve of Proposition E.10 (dotted). (c) Optimal learning-rate schedule at $M = 64$, $B = 8$: fewer backward bits, earlier and gentler decay.

is why the exact-recursion slopes sit below the asymptotic ones while agreeing with the finite-width first-order response. The backward slope is $\frac{2-b}{a}$ while k_t^{1-b} dominates the residual and $\frac{1}{a}$ once σ^2 does; at $\sigma^2 = 0.1$ the crossover ($k_t \approx \sigma^{-2/(b-1)} = 100$) lies at the edge of the window even at $M = 2048$. Exact recursion at $M = 256, 512, 1024$ on the same feature maps as the reduced system (trajectories of the comparison above; 23, 21 and 19 checkpoints in the window $k_t \in [0.03M, 0.1M]$): $0.71, 0.96, 0.10, 0.22$; $0.74, 1.10, 0.18, 0.25$; $0.73, 1.15, 0.21, 0.27$. At $M = 1024$ it differs from the reduced system by at most 0.04, and in the window $\tau_t \in [M^a/100, M^a/3]$ the same trajectories reproduce the exact-recursion slopes above at $M = 64$ and 128. Penalty fractions at $\tau_T = M^a$ (last iterate, exact-arithmetic risk including A_S): Appendix G. On $\tau_t \in [M^a/100, M^a/3]$ at $M = 128$ the exact-recursion and reduced-system slopes ($0.69/0.71$, $1.03/1.07$, $0.15/0.17$, $0.16/0.17$) equal those of the finite-width first-order response (moved from §5).

Where to round: penalty shares and the regime map. Exact recursion at $\tau_T = M^a$, $p = 1$ (and $p = 2$), feature maps as in the comparison and the joint term above. Shares of the exact-arithmetic risk, including A_S, at one bit and $M = 16, \ldots, 1024$ (median over feature maps): last iterate, forward $0.49, 1.06, 1.99, 3.88, 7.08, 14.6, 30.5\%$ and backward $0.040, 0.033, 0.040, 0.046, 0.054, 0.064, 0.077\%$; average with both copies rounded 0.016 to 0.086%. The forward share grows like $M^{0.98}$, against M^{a-1}. Proposition 5(i): the ratio of the forward to the backward penalty of the last iterate is $29, 49, 84, 134, 230, 377$ at $M = 32, \ldots, 1024$, and $1.43\, M^{a-b}/(\sigma^2 + M^{1-b})$ within 9% at every width; at $p = 2$ it is $84.1, 134.4, 229.7$ at $M = 128, 256, 512$, equal to the values at $p = 1$ to three digits. Proposition 5(ii): for the average, backward over forward is $1.9, 2.1, 2.4, 2.8, 3.3, 4.0$, within 20% of $(\sigma^2 + M^{1-b})/M^{1-b}$, and interaction over backward is $0.22, 0.37, 0.63, 1.03, 1.78, 2.99$ at $p = 1$ and 3.9 times smaller at $p = 2$. The drivers of §3: the squared parameter norm at $\tau_T = M^a$ grows like $M^{1.50}$, against M^{a-b+1}, and the backward penalty divided by ϱR_T, with R_T the residual of the exact-arithmetic run, stays within $\pm 10\%$ (last iterate 0.0086–0.0106, average 0.0023–0.0029) while R_T falls from 0.53 to 0.16 between $M = 16$ and 1024. The measured column of Table 1 is $p^\vartheta = 1 + \log_4(\text{share at one bit}/\vartheta)$ at $M = 1024$ and $\vartheta = 0.1$ ($2 + \log_4$ of the share for the joint row, whose penalty scales with $q_{\text{fw}} q_{\text{bw}}$); one more bit divides the single-copy penalties by 3.92 to 4.00 and the interaction by 15.4. Its cooled entry uses the cooldown run above at $M = 512$, $\phi = 1$ and $p_{\text{bw}} = 2$. The "theory" column of Table 1 uses $m_p(g) \approx 4^{-p} g/(8 \ln 4)$ in place of $4^{-p} g$, which lowers every entry by $\frac{1}{2}\log_2(8 \ln 4) = 1.7$ bits and the joint entry by 3.5; with 4^{-p} in place of m_p the joint entry would ask for 3.7 bits that the exact recursion does not need. Figures 1(a), 2(a) and 4(c) show the shares and the two ratios of Proposition 5 against the width.

Compute law. Exact recursion at $M = 16, 24, 32, 48, 64, 96, 128, 192, 256, 384, 512$ (feature map 1), configurations exact, (p, p) and (∞, p) for $p = 1, \ldots, 4$, trajectories to $\tau_t = 2M^a$ (200 log-spaced checkpoints) read at $\tau_T = M^a$ by log-log interpolation. Cost $f = BTM[w(p_{\text{fw}}) + w(p_{\text{bw}})]$

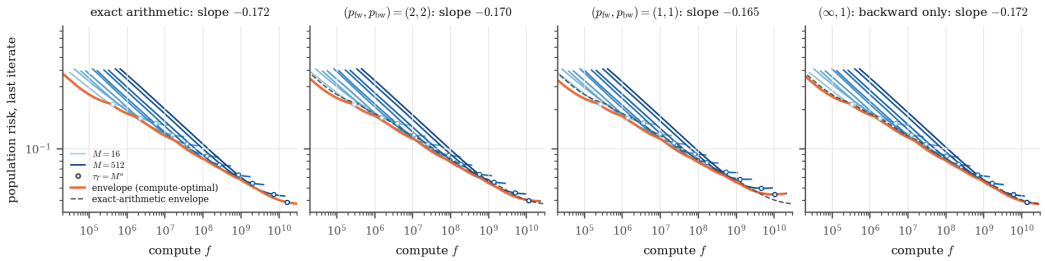


Figure 6: Compute-optimal frontier of the last iterate, exact recursion at $M = 16$–512 (light to dark), $\tau_t \le 2M^a$: the risk of every width against the compute f, the point $\tau_T = M^a$ on each trajectory (circles) and the envelope over widths (orange), for exact arithmetic, $(p_{\mathrm{fw}}, p_{\mathrm{bw}}) = (2,2)$, $(1,1)$ and backward rounding alone at one bit. The exact-arithmetic envelope (dashed) is repeated in every panel, and each title gives the fitted exponent of the envelope over $f \in [10^6, 3 \times 10^9]$.

with $w(p) = (p+9)/16$ ($p_0 = 9$, $p_{\mathrm{ref}} = 7$) and $w = 1$ for an exact copy. In exact arithmetic the risk, including A_S, falls like $f^{-0.168}$ for the last iterate and $f^{-0.170}$ for the average, against $-(b-1)/(a+1) = -0.167$. For the last iterate at (p,p) the penalty share grows like $f^{0.32}$, against $f^{(a-1)/(a+1)} = f^{1/3}$, from 0.46% at $M = 16$ to 14.9% at $M = 512$ at one bit, and one more bit divides the penalty by 3.97 to 4.00 (3.92 to 3.99 with backward rounding alone). The bits needed, $p^\vartheta = 1 + \log_4(\text{share at one bit}/\vartheta)$, therefore grow by 0.162 bit per doubling of compute for every ϑ (0.4 to 2.9 bits at $\vartheta = 1\%$), against $\frac{a-1}{2(a+1)} = \frac{1}{6}$ in Theorem F.8(i); here $\mathsf{n} \le 0.04$. At one bit the local exponent of the risk of the last iterate over the five widest runs is -0.137 instead of -0.162 (-0.155 at two bits). Not tested at these widths: the penalty of the average is at most 0.05% of its risk at one bit, and its local exponents do not depend on the precision to three digits, so the floor and the exponents of Theorem F.8(iii) lie beyond $M = 512$; with backward rounding alone the bits grow by 0.03 to 0.04 per doubling, the value of $\frac{1}{2}\log_2(\mathsf{n} + \varrho)$ in its crossover ($\sigma^2 M^{b-1} \le 2.3$ here), not the asymptotic $\frac{1}{12}$. Figures 2(c) and 1(c) show the penalty shares and the bits needed against compute, and Figure 6 the risk of every width against compute with the envelopes: over $f \in [10^6, 3 \times 10^9]$ the envelope falls like $f^{-0.172}$ in exact arithmetic and at $(\infty, 1)$, like $f^{-0.170}$ at $(2,2)$ and $f^{-0.165}$ at $(1,1)$, so backward rounding leaves the compute-optimal law unchanged and the bend of Theorem F.8(iii) is still small at these widths.

Optimal learning-rate schedule. Deterministic-equivalent dynamics at $M = 64$, $B = 8$, label noise $\sigma^2 = 1$, $T = 8868$ steps and peak rate $\eta_{\max} = 0.46$, with the forward copy exact and the backward copy at $p_{\mathrm{bw}} \in \{\infty, 1, 0\}$; the step-size sequence that minimizes the last-iterate excess risk was found numerically (Figure 5(c)). The optimal risk is 0.0072, 0.0077 and 0.0091 for $p_{\mathrm{bw}} = \infty, 1, 0$, against 0.045–0.049 at the constant peak rate, and the rate first falls below 99% of its peak at $t/T = 0.89$, 0.87 and 0.79: with fewer backward bits the decay starts earlier and is gentler.

Input rounding. $M = 8$, $d = 32$, $B = 4$, $\eta = 0.5/g_1$, $T = 60$, seed 3, 10^5 runs of the bit-exact trainer with the d-dimensional input rounded at p_x bits under shared or independent copies and the post-feature operands at $(p_{\mathrm{fw}}, p_{\mathrm{bw}}) \in \{(\infty,\infty), (1,1), (0,0)\}$; the Gaussian-smoothing solver of Proposition C.7 / Theorem C.8 on the same realised problem. Exact cases ($p_{\mathrm{fw}} = p_{\mathrm{bw}} = \infty$): $\max|z| = 1.85$ (independent, $p_x = 1$), 1.74 ($p_x = 2$), 1.44, 1.56 (shared). Approximate cases, relative error of the post-feature penalty at T (solver minus Monte Carlo, $\pm$ one s.e.): independent $(1,1)$: $-5.2 \pm 2.5\%$, $-2.9 \pm 2.4\%$, $-2.9 \pm 2.4\%$ at $p_x = 1, 2, 4$; independent $(0,0)$, $p_x = 2$: $-1.4 \pm 1.1\%$; shared $(1,1)$: $-8.3 \pm 2.7\%$, $-2.4 \pm 2.5\%$, $+1.6 \pm 2.5\%$; shared $(0,0)$, $p_x = 2$: $-2.7 \pm 1.2\%$; the excess risk itself within 0.6% in every case. Coefficient-level probe (one feature map, $M = 16$, $d = 64$, $p_{\mathrm{fw}} = p_{\mathrm{bw}} = 2$, 4×10^6 Gaussian draws): the marginal $\mathbb{E}\nu_{\mathrm{fw}}(\hat{z}_k)$ is exact to noise for $p_x \ge 2$; the weighted terms err by 0.5–1.4% at $p_x = 1$, 0.15–0.3% at $p_x = 2$, $\le 0.04\%$ at $p_x = 4$; the independent-copy same-coordinate pair coefficient errs by $+2.0\%$, -0.7%, -1.3% at $p_x = 1, 2, 4$ and does not improve with p_x. The shared-copy floor is $\asymp 16^{-p_x}$.

Networks. Every linear layer is replaced by a layer whose forward product uses $Q_{p_{\mathrm{fw}}}(A)W^\top$ and whose weight-gradient product uses $Q_{p_{\mathrm{bw}}}(A)^\top D$, with independent rounding decisions for the two

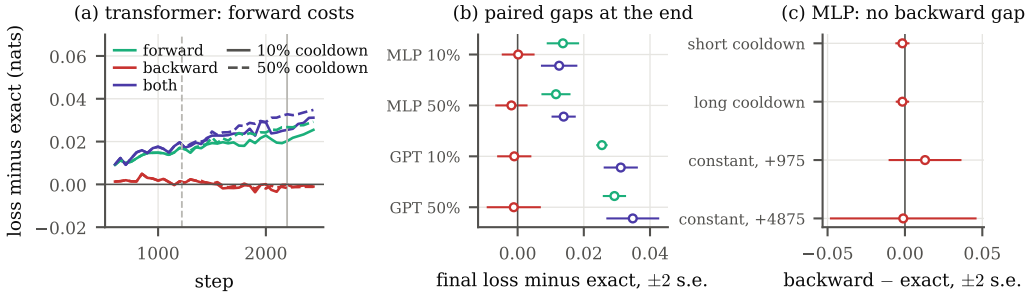


Figure 7: Networks with SR at $p = 2$ bits in every linear layer. (a) Transformer: validation loss minus exact arithmetic along training for the forward copy, the backward copy and both, with a 10% (solid) and a 50% (dashed) cooldown starting at the vertical lines. (b) Final validation loss minus exact arithmetic, seed-paired, ± 2 s.e., MLP (16 seeds) and transformer (3 seeds). (c) MLP, backward copy against exact arithmetic, continuations restored from one checkpoint per seed (± 2 s.e.).

copies, the grid of §2 and $p = 2$; the gradient with respect to the input, the weights, the optimizer state and all other arithmetic are unrounded (fp32, and bf16 autocast for the transformer, which represents the rounded values exactly). Conditions: exact arithmetic, forward copy only, backward copy only, both; the runs of one seed share their initialization and data order. MLP: 3072–1024–10 with ReLU on CIFAR-10, SGD with momentum 0.9 and weight decay 5×10^{-4}, batch 128, 25 epochs (9750 steps), peak rate 0.01, 200 warmup steps, linear cooldown to zero over the last 10% or 50% of the steps, 16 seeds; validation on the test set. Transformer: 6 layers, width 384, 6 heads, context 256, GPT-2 vocabulary, 30M parameters, AdamW (peak rate 6×10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1), batches of 64×256 tokens, 40M FineWeb-Edu tokens (2441 steps), 100 warmup steps, the same cooldowns, 3 seeds; the tied output layer is not rounded; validation on 40 fixed batches. Final validation loss minus exact arithmetic, seed-paired mean $\pm$ standard error, for the 10% and 50% cooldowns. MLP: forward $+0.0137 \pm 0.0024$, $+0.0116 \pm 0.0022$; backward $+0.0002 \pm 0.0025$, -0.0019 ± 0.0024; both $+0.0125 \pm 0.0028$, $+0.0139 \pm 0.0018$; the validation accuracy is within 0.2 points of exact arithmetic in every case. Transformer: forward $+0.0255 \pm 0.0009$, $+0.0292 \pm 0.0018$; backward -0.0010 ± 0.0026, -0.0012 ± 0.0041; both $+0.0312 \pm 0.0026$, $+0.0348 \pm 0.0040$. At peak rate 0.02 the MLP gives the same picture (forward $+0.006$ and $+0.015$; backward within 1.4 standard errors of zero). Checkpoint controls (MLP, exact and backward-only, 16 seeds): each condition trains one prefix of 7800 steps (200 of warmup) at the peak rate, and its model, momentum, random state and data stream are restored for a linear cooldown to zero over 975 or 4875 steps and for 4875 steps at the peak rate. Backward minus exact validation loss: -0.0016 ± 0.0023 (short cooldown), -0.0016 ± 0.0022 (long cooldown), $+0.013 \pm 0.012$ and -0.001 ± 0.024 (constant rate, after 975 and 4875 steps). The MLP overfits in this regime (training loss 0.08 after the long cooldown, validation accuracy 52–56%), and the transformer is trained briefly (validation loss 4.75–4.83). The validation loss is computed with the rounding of the forward copy active, so the forward and both gaps include the rounding at inference; the checkpoint controls have an exact forward pass. In exact arithmetic the transformer's validation loss falls by 0.15 nats over the 10% cooldown and by 0.47 nats over the 50% cooldown (means over the three seeds); if this is the SGD noise that the cooldown removes and backward rounding at two bits adds about $0.09 \cdot 4^{-2}$ of it (Proposition 5(i)), the backward gap to expect after the cooldown is about 0.001 and 0.003 nats, below the resolution of the measurement. Figure 7 is the full-size copy of Figure 7.

Runtime. The exact recursion costs four dense $M \times M$ products per step; $M = 256$ takes about 5 minutes for $\tau_T = M^a$ on a laptop CPU; on one A100 GPU in float64, $M = 1024$ (4.3×10^6 steps, to $\tau_t = 2M^a$) takes 1.2 hours per precision configuration, together with the reduced system. Rounded SGD takes 0.1–0.9 hours per configuration at $M = 128$ (16 384 runs) and 0.6–2.9 hours at $M = 512$ (2048 runs). One MLP run takes 12 seconds and one transformer run 8 minutes.

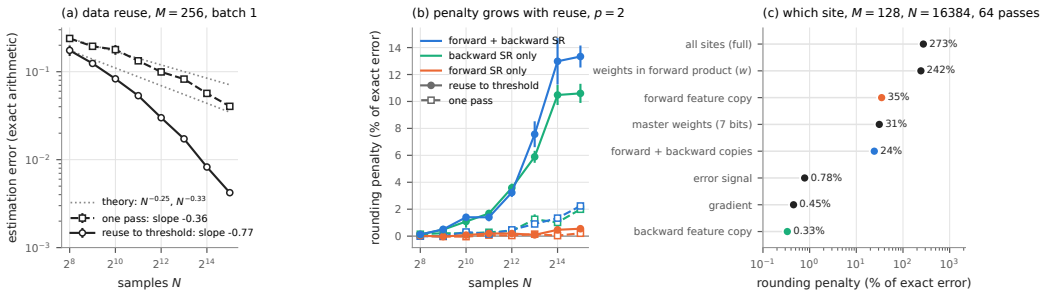


Figure 8: **Data reuse and stochastic rounding (E10). (a)** Estimation error of SGD in exact arithmetic on a finite dataset of N samples, in the setting of Lin et al. (2025) ($a = 2$, $b = 1.5$, $M = 256$, batch size 1, geometric step-size decay, label noise $\sigma^2 = 1$). One pass over the data (squares, dashed) decays as $N^{-0.36}$; reusing the data for $N^{a/b}/\eta$ steps (12–60 passes; circles, solid) decays as $N^{-0.77}$. Dotted lines: the predicted rates $N^{-(b-1)/a} = N^{-1/4}$ and $N^{-(b-1)/b} = N^{-1/3}$. **(b)** Same runs with the features rounded to $p = 2$ bits in the forward copy, the backward copy, or both, reported as a percentage of the exact run's error on the same data and sample order. After one pass every penalty is at most 2.2%. With reuse the backward penalty grows with N to about 10%, while the forward penalty stays below 1%. **(c)** Penalty of each rounding location used alone, at batch size 32, $M = 128$, $N = 16384$, after 64 passes (constant step size, $\sigma^2 = 0.1$; 2 bits, master weights 7 bits). Rounding the weights in the forward product accounts for almost all of the penalty of rounding every location (242% of 273%); the error signal, the gradient and the backward feature copy each cost below 1%. Mean $\pm$ s.e. over 4 seeds in (a, b) and 2 seeds in (c); each penalty is averaged over 16 (a, b) or 64 (c) paired rounding draws.

H Predictions beyond the model: other rounding sites and data reuse

Other rounding sites. We predict that rounding the gradient acts, late in training, like rounding the backward copy, and that rounding the stored weights, when the update is smaller than the rounding interval, adds a penalty that a cooldown neither removes nor amplifies and that halves rather than quarters per bit. The reasoning is a first-order heuristic. Theorem 1 closes because the two operands are rounded before they meet the state. Every other location rounds a state-dependent quantity and the second moments do not close (Appendix B); the profile of Lemma B.1 still gives a first-order heuristic, $\mathbb{E}[\nu_p(X) \mid \text{state}] \approx 0.09 \cdot 4^{-p}\, \mathbb{E}[X^2 \mid \text{state}]$ for an argument spread over many rounding intervals.

site	injected covariance per step (feature coordinates)	scales with	driver	geometry
forward operand	$\frac{\eta^2}{B} \iota_{\text{fw}} \mathsf{U}_t (G + M_b)$	η^2/B	parameter norm	aligned (exact)
backward operand	$\frac{\eta^2}{B} \mathsf{R}_t M_b$	η^2/B	residual	isotropic (exact)
gradient $\hat{g}_t$	$\eta^2\, 0.09 q_g [\operatorname{diag}(GC_tG) + \frac{1}{B} \operatorname{diag}(\mathsf{R}_t G + \dots)]$	η^2 (no $1/B$ early)	$\|G\delta_t\|^2$, then residual$/B$	isotropic
weights, small update	$\eta\, 2^{-p_u} \mathbb{E}\lvert v_k \rvert \lvert \hat{g}_k \rvert$	η (penalty η-free)	$\|v\|\,\|g\|$	isotropic
weights, large update	$0.09 \cdot 4^{-p_u} \mathbb{E} v_k^2$	η^0 (penalty $\propto 1/\eta$)	parameter norm	isotropic
residual	$(\mathsf{R}_t G + 2GC_tG) \times 0.09 q_r$	η^2/B	residual	aligned

Gradient rounding is, late in training, exactly a backward operand with $q_{\text{bw}} \to q_g$ ($\eta^2\, 0.09 q_g \mathsf{R}_t G_{kk}/B = (\eta^2/B) \mathsf{R}_t\, 0.09 q_g\, \operatorname{tr} \Sigma_d/M$ per coordinate) and early in training a batch-independent term $\propto \operatorname{tr}(GC_tG) = \sum_j g_j^2 e_j \asymp k_t^{1-a-b} + \varrho r_t$; it is removed by a cooldown (η^2). Rounding the stored weights, as in SR training of quantized weights (Li et al., 2017; Zhao et al., 2025), is different in kind: when the update is smaller than the rounding interval ($\eta \lvert \hat{g}_k \rvert \ll 2^{-p_u} \lvert v_k \rvert$, the regime of bf16 master weights under Adam, where relative updates are 10^{-3}–10^{-4} against 2^{-7}) the stored value sits on the grid and the error variance is $\approx \eta \lvert \hat{g}_k \rvert \cdot 2^{-p_u} \lvert v_k \rvert$, linear in η, so the stationary per-mode variance $\propto 2^{-p_u} \lvert v \rvert \lvert \hat{g} \rvert / g_j$ and the penalty $\asymp 2^{-p_u} k_t \sqrt{\mathsf{U}_t \mathsf{R}_t \operatorname{tr} \Sigma_d / B}/M$ are η-free: a cooldown neither removes nor amplifies them, and they halve rather than quarter per bit. In the large-update regime the variance is $0.09 \cdot 4^{-p_u} v_k^2$ and the stationary penalty $\asymp 4^{-p_u} 0.09\, \mathsf{U}_t k_t / (\eta M)$ grows as the rate decays, so the storage precision must rise by $\frac{1}{2} \log_2(\eta_{\text{peak}}/\eta_{\text{fin}})$ bits through a

cooldown. The test is the bit-exact trainer (which needs three switches) against a surrogate recursion with the table's terms inserted, reporting its error as Appendix G does.

Data reuse. We predict that the bits are set by the number of steps, not of samples, so that multi-pass training needs more bits per sample: $p_{\mathrm{bw}} \approx \frac{1}{2}\log_2 N$ and $p_{\mathrm{fw}} \approx \frac{a}{2b}\log_2 N$ at maximal reuse and fixed width. The reasoning: with L steps over N samples, Lin et al. (2025) show that multi-pass SGD matches one-pass SGD with $\tau = \eta L$ as long as $L \lesssim N^{a/b}/\eta$, so data can be reused for $N^{a/b-1}$ passes at no loss and the compute-optimal allocation $L \sim C^{a/(a+1)}$, $M \sim C^{1/(a+1)}$ needs only $N \sim C^{b/(a+1)}$ samples. Since the rounding decisions are fresh at every access, the rounding rows of Theorem 4 apply with $\tau_t \to \eta t$ below the threshold: the bits are set by the step count L, not by N, which gives the per-sample counts above. Above the threshold the training residual falls and the backward noise with it, while the forward noise does not. The test is the bit-exact trainer at $M = 64$, $N \in \{M, 4M, 16M, 64M\}$, up to $4N^{a/b}$ steps, comparing the penalty at step L with the one-pass recursion at the same L.