SAME-SAMPLE COUPLING BOUNDS ADAM'S STEPS AND DAMPS TRAINING-ORDER NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam computes its first and second moments from the same minibatch. We show that this detail governs how Adam responds to the order of its training data. Linearizing a step, the adaptive contribution to order dispersion splits into a non-negative preconditioner term V_p and a numerator–denominator covariance C. We prove that under same-sample coupling C is negative whenever momentum agrees in sign with the current gradient, and that it outweighs V_p when β_2 is close to one; an independent second-moment sample makes C vanish, so decoupled memory can only amplify order dispersion. Coupling also gives Adam a per-coordinate step bound, a known consequence of Cauchy–Schwarz that we show holds if and only if the second-moment input dominates the numerator input. We test both consequences by changing the pairing and nothing else. In a small adapter, manipulating the alignment between numerator and second-moment fluctuations flips the sign of the adaptive correction in all 18 units. At mature ResNet-18 checkpoints, replacing each second-moment sample by an independent draw with the same task and marginal law increases order dispersion on three task pairs, at every checkpoint for parameter and prediction variance. On pretrained Pythia-160M the same replacement drives a small set of coordinates 10^2–10^5 times past the step bound at the first update and the model diverges within a few updates, even when fifteen of sixteen sequences in the second-moment minibatch are shared.

1 INTRODUCTION

Two fine-tuning runs can start from the same checkpoint, process exactly the same minibatches, and still finish at different parameters because the minibatches arrived in a different order. This order dispersion sets a floor on how reproducible a run is and on how small a difference between two data orderings or curricula can be detected. For memoryless gradient descent it is the familiar gradient commutator: applying domains A then B rather than B then A gives

$$\theta_{AB} - \theta_{BA} = \eta^2 \left(H_B g_A - H_A g_B\right) + O(\eta^3), \tag{1}$$

with $g_D = \nabla L_D$ and $H_D = \nabla^2 L_D$ (Dherin, 2023; Rukhovich et al., 2024; Sweeney, 2026a).

Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) change this picture because the update is not a function of the parameters alone. The first moment m remembers directions and the second moment v remembers coordinate scales. Fixed-clock memory can make shuffle order a first-order effect (Sweeney, 2026b), and finite-horizon analyses assign signed influence to perturbations of AdamW state (Liu & Li, 2026). What is missing is an account of the *sign*: when does adaptive preconditioning amplify order noise, and when does it suppress it?

Our answer turns on a detail that is usually treated as an implementation choice. In every Adam step, the numerator and the denominator are updated from the same minibatch gradient, so a minibatch that pushes a coordinate unusually hard also enlarges that coordinate's denominator in the same step (fig. 1). We call this *same-sample coupling*, and we show that it, rather than the marginal behaviour of either moment, decides whether the second moment damps or amplifies order dispersion and whether Adam's steps stay bounded.

Contributions.

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.

- **Coupling sets the sign.** The adaptive contribution to local order dispersion is $\Delta V = V_p + C$ with $V_p \geq 0$. We prove that coupling makes C negative wherever momentum agrees in sign with the current gradient, and large enough to outweigh V_p when β_2 is close to one, while an independent second-moment sample makes $C = 0$ (theorem 1).
- **Coupling bounds the step, and only exact coupling.** We characterize Adam's known per-coordinate step bound: it holds if and only if the second-moment input dominates the numerator input coordinatewise (theorem 2).
- **Interventions on the pairing alone.** Manipulating numerator–denominator alignment flips the sign of the adaptive correction in every unit of a small adapter; breaking the pairing increases order dispersion at mature ResNet-18 checkpoints on three task pairs, and destabilizes pretrained Pythia-160M even when all but one of sixteen sequences are shared (section 5).

Two further results are reported in support: a clock-phase law for when optimizer memory enters the order contrast at leading order (section 3, section A.2), and the batch-size target of the second moment together with what transporting it buys at a batch-size switch (section 5.4).

2 Related work

Order effects and optimizer memory. Backward-error and Lie-bracket analyses relate $A \to B$ versus $B \to A$ to noncommuting gradient fields (Dherin, 2023; Rukhovich et al., 2024; Sweeney, 2026a). Sweeney (2026b) shows that fixed-clock optimizer memory makes shuffle order a first-order source of fine-tuning noise; our clock-phase law describes how the order interpolates as memory rate and optimizer age are scaled. Liu & Li (2026) linearize AdamW as a finite-horizon input–output system; their second-moment write-in contains the same-sample term $2(1-\beta_2)g \odot \delta g$. We derive what that term implies for the sign of order dispersion and test it by intervening on the pairing.

The numerator–denominator relation. Kingma & Ba (2015, Sec. 2.1) note that Adam's effective step is bounded, and Cauchy–Schwarz bounds of the form we use appear as lemmas in convergence analyses under $\beta_1^2 < \beta_2$ (Zou et al., 2019; Défossez et al., 2022). AdaShift (Zhou et al., 2019) and ADOPT (Taniguchi et al., 2024) deliberately remove the current gradient from the second moment to repair convergence counterexamples (Reddi et al., 2018), and ADOPT adds clipping to keep steps finite. Adafactor's update clipping (Shazeer & Stern, 2018) and StableAdamW (Wortsman et al., 2023) address loss spikes that arise when gradients outrun a slowly updated second moment. Heavy-tailed and class-imbalanced gradient noise in language models (Zhang et al., 2020; Kunstner et al., 2024) explains why such events are common there. Our contribution is to connect this stability property to the sign of order effects, to show that it requires exact rather than approximate coupling, and to measure both by intervention.

Batch size and optimizer state. SDE analyses give square-root learning-rate scaling for adaptive methods (Malladi et al., 2022); batch-size-invariant second-moment estimators (Wang & Aitchison, 2024) and token-scaled moment half-lives (Marek et al., 2025) address the same dependence from the estimator side, and optimizer resets are a common remedy at objective changes (Asadi et al., 2023).

3 Setup: order contrasts and when memory matters

A *word* $w = (d_1, \ldots, d_K)$ is a sequence of minibatch updates. We compare words that contain the same minibatches in different orders, so that differences between their endpoints are due to chronology alone. The *order dispersion* of a fixed bank of minibatches is the variance of the endpoint over orderings of that bank; averaging over independently drawn banks gives the conditional order variance reported throughout.

Optimizer memory enters this contrast at leading order only in some regimes. If $1 - \beta \asymp \eta^p$ and the optimizer's age (its step count) is $s \asymp \eta^{-q}$, a gradient's bias-corrected write-in weight scales as $\eta^{\min(p,q)}$, so memory contributes to the order contrast at $\Theta(\eta^{1+\min(p,q)})$ (theorem 3, section A.2). Fixed decay rates ($p \to 0$) make order a first-order effect, as in Sweeney (2026b); the linear scaling $p = q = 1$ of continuous analyses restores the η^2 order of eq. (1). Numerical sweeps recover every

predicted exponent to within 10^{-3}. The rest of the paper concerns the regime in which memory matters and asks what sets the sign of its contribution.

4 SAME-SAMPLE COUPLING

4.1 THE SIGN OF THE ADAPTIVE CONTRIBUTION

At a given step, center the word-dependent bias-corrected numerator and preconditioner across words: $\widehat{m}_w = \bar{m} + \delta m_w$ and $p_w = \bar{p} + \delta p_w$ with $p_w = (\sqrt{\widehat{v}_w} + \epsilon)^{-1}$. The word-dependent part of the update is, to first order, $\delta u_w = \underbrace{\bar{p} \odot \delta m_w}_{A_w} + \underbrace{\bar{m} \odot \delta p_w}_{B_w}$. Comparing its dispersion with that of a preconditioner held at its mean gives the identity

$$\Delta V \equiv \mathbb{E}_w \|\delta u_w\|^2 - \mathbb{E}_w \|A_w\|^2 = \underbrace{\mathbb{E}_w \|B_w\|^2}_{V_p \geq 0} + \underbrace{2\,\mathbb{E}_w \langle A_w, B_w \rangle}_{C}. \tag{2}$$

Letting the preconditioner vary with the data order costs $V_p \geq 0$; it can only help if the preconditioner moves against the numerator, $C < -V_p$. The identity is elementary. What matters is that same-sample coupling fixes the sign of C.

Theorem 1 (Coupling sets the sign)**.** *Consider one step from a shared optimizer state, t steps old, in which words differ only through the current gradient $g_w = \bar{g} + \delta_w$, $\mathbb{E}_w \delta_w = 0$, $\sigma^2 = \mathbb{E}_w \delta_w^{\odot 2}$, and the gradient s_w fed to the second moment. Let $a = (1-\beta_1)/(1-\beta_1^t)$, $b = (1-\beta_2)/(1-\beta_2^t)$ and $\bar{v} = \mathbb{E}_w \widehat{v}_w$. To first order in the fluctuations, for each coordinate i,*

$$C_i = -\,ab\,\frac{\bar{m}_i \operatorname{Cov}_w(g_{w,i}, s_{w,i}^2)}{\sqrt{\bar{v}_i}\,(\sqrt{\bar{v}_i} + \epsilon)^3}, \qquad V_{p,i} = b^2\,\frac{\bar{m}_i^2 \operatorname{Var}_w(s_{w,i}^2)}{4\bar{v}_i(\sqrt{\bar{v}_i} + \epsilon)^4}.$$

*Consequently: **(i)** under coupling ($s_w = g_w$), $\operatorname{Cov}_w(g_{w,i}, g_{w,i}^2) = 2\bar{g}_i\sigma_i^2$ and coordinate i damps ($C_i < -V_{p,i}$) if and only if $0 < \bar{m}_i\bar{g}_i < 2\frac{a}{b}\sqrt{\bar{v}_i}\,(\sqrt{\bar{v}_i} + \epsilon)$, and amplifies if $\bar{m}_i\bar{g}_i < 0$; **(ii)** if s_w is independent of g_w across words, $C = 0$ and $\Delta V = V_p \geq 0$.*

Proof. With shared history, $\delta m_w = a\,\delta_w$ and $\delta\widehat{v}_w = b\,(s_w^{\odot 2} - \mathbb{E}_w s_w^{\odot 2})$, and to first order $\delta p_w = p'(\bar{v})\,\delta\widehat{v}_w$ with $p'(v) = -\frac{1}{2}v^{-1/2}(\sqrt{v}+\epsilon)^{-2}$. Hence $A_{w,i} = a\bar{p}_i\delta_{w,i}$ and $B_{w,i} = b\,\bar{m}_i p_i'\,\delta(s_{w,i}^2)$, and taking expectations gives C_i and $V_{p,i}$. Under coupling, $\operatorname{Cov}_w(g, g^2) = 2\bar{g}\sigma^2$ and $\operatorname{Var}_w(g^2) = 4\bar{g}^2\sigma^2$ to first order, so $C_i + V_{p,i} = b\,\bar{m}_i\bar{g}_i\sigma_i^2\big(b\bar{m}_i\bar{g}_i - 2a\sqrt{\bar{v}_i}(\sqrt{\bar{v}_i}+\epsilon)\big)\big/\big(\bar{v}_i(\sqrt{\bar{v}_i}+\epsilon)^4\big)$, whose sign gives (i). Independence gives zero covariance, hence (ii). □

The condition in (i) is mild. Because $a/b = (1-\beta_1)(1-\beta_2^t)/((1-\beta_2)(1-\beta_1^t))$ is large when β_2 is close to one (about 100 for $(\beta_1, \beta_2) = (0.9, 0.999)$ at maturity), damping holds on every coordinate where momentum and the current mean gradient agree in sign unless $|\bar{m}_i\bar{g}_i|$ is many times $\bar{v}_i$. The general form makes the mechanism explicit: the sign of C is set by the covariance, across words, between the gradient that drives the numerator and the gradient that enters the second moment. Same-sample coupling makes that covariance as large as it can be; an independent sample makes it zero; a second moment fed from the wrong task can make it negative. Two predictions follow. Varying this covariance should move the adaptive correction through zero, and replacing the second-moment sample by an independent draw with the same marginal law should increase order dispersion. It is a statement about one step from a shared state; whether the sign persists over a rollout is an empirical question (section A.6).

4.2 THE COUPLED STEP BOUND

Coupling also has a non-perturbative consequence. Write $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2)s_t^{\odot 2}$ from $m_0 = v_0 = 0$, where g_t feeds the numerator and s_t the second moment; standard Adam has $s_t = g_t$.

Proposition 2 (Coupled step bound)**.** *Let $\beta_1^2 < \beta_2$, $r = \beta_1^2/\beta_2$ and $\kappa_t = (1-\beta_1)\sqrt{(1-r^t)/(1-r)}\,\sqrt{1-\beta_2^t}\,/\,\big((1-\beta_1^t)\sqrt{1-\beta_2}\big)$. **(i)** If $s_t^{\odot 2} \geq c\,g_t^{\odot 2}$ coordinatewise for all t and some $c > 0$, then $|\widehat{m}_t|/(\sqrt{\widehat{v}_t} + \epsilon) \leq \kappa_t/\sqrt{c}$ for every coordinate, t and $\epsilon \geq 0$; live Adam is the*

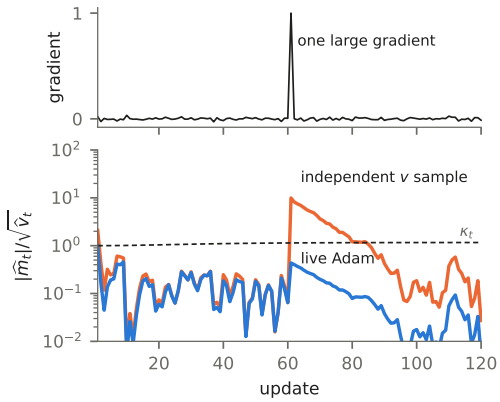

Figure 1: Why coupling bounds the step. A rarely active coordinate receives one large gradient (top). Under live Adam the same gradient enters the denominator, and the adaptive step stays below κ_t (dashed). With the second moment fed an independent draw from the same distribution, the denominator never sees the spike and the step exceeds the bound for tens of updates.

case $c = 1$, and the bound is attained. ***(ii)*** *If no such c exists, the ratio is unbounded over sequences as $\epsilon \to 0$.*

Proof. (i) Unroll $m_t = (1-\beta_1)\sum_{k<t}\beta_1^k g_{t-k}$, write $\beta_1^k = r^{k/2}\beta_2^{k/2}$ and apply Cauchy–Schwarz: $|m_t| \le (1-\beta_1)\big(\sum_{k<t} r^k\big)^{1/2}\big(\sum_{k<t}\beta_2^k g_{t-k}^2\big)^{1/2} \le (1-\beta_1)\sqrt{(1-r^t)/(1-r)}\,\sqrt{v_t/(c(1-\beta_2))}$; divide by the bias corrections. Equality holds for $c = 1$, $\epsilon = 0$ and $g_{t-k} \propto (\beta_1/\beta_2)^k$. (ii) After a history with $g = s = 0$, take a step with $g_t \neq 0$ and $s_t^2 < \delta g_t^2$; the ratio then exceeds $(1-\beta_1)/\big((1-\beta_1^t)(\sqrt{b\delta} + \epsilon/|g_t|)\big)$ with $b = (1-\beta_2)/(1-\beta_2^t)$, which grows without bound as $\delta, \epsilon \to 0$. □

For $(\beta_1, \beta_2) = (0.9, 0.95)$, $\kappa_\infty = 1.165$; for $(0.9, 0.999)$, $\kappa_\infty = 7.27$. Part (i) with $c = 1$ is known (Kingma & Ba, 2015; Zou et al., 2019; Défossez et al., 2022); the characterization in (ii) is what the experiments exercise. An independent draw from a heavy-tailed or token-sparse distribution violates domination with positive probability, and so does a minibatch that shares all but one sequence with the numerator's, since that sequence may carry most of a coordinate's gradient. The event that breaks the bound, a large numerator gradient meeting a denominator that has not seen it (fig. 1), is the extreme form of the positive covariance that theorem 1 shows coupling removes.

4.3 The batch size sets the second moment's target

For independent per-example gradients with mean μ and covariance Σ, the minibatch mean satisfies $q_B \equiv \mathbb{E}[\bar{g}_B^{\odot 2}] = \mu^{\odot 2} + \operatorname{diag}\Sigma/B$. The second moment therefore measures minibatch noise, and a change of batch size from B to B' moves its target. The state matching the new batch size is $v^\star = \mu^{\odot 2} + \frac{B}{B'}(v_{\text{old}} - \mu^{\odot 2})$; where noise dominates this is $v^\star \approx \frac{B}{B'} v_{\text{old}}$, a uniform rescaling of every adaptive step by $\sqrt{B'/B}$, which is the square-root rule of Malladi et al. (2022) applied at the switch.

5 Experiments

The experiments intervene on the pairing in three settings of increasing realism, then test the batch-size corollary. Independent checkpoints or warm-phase seeds are the statistical units; orders, banks and horizons are repeated measurements (section D).

5.1 CONTROLLED ALIGNMENT IN A SMALL ADAPTER

We freeze an ImageNet ResNet-18 feature extractor and train a 1,040-parameter adapter on six CIFAR-100 superclass pairs with three checkpoint seeds each (18 units), evaluating all 20 balanced length-six words with three updates from each task. The first-order quantity $V_p + C$ and the exact adaptive correction (the same dispersion difference computed without linearization) agree in sign for 96.2–99.7% of words at $B = 8$, and the correction takes both signs within at least 15 of the 18 units, so this is not the accuracy of a constant guess. This confirms that the linearization of theorem 1 is accurate over six updates; the causal test is the next one.

Keeping the action words fixed, we then vary the covariance in theorem 1 directly through the alignment ρ of the minibatches fed to the second moment: at $\rho = 1$ each update's second moment receives an independent minibatch from its own task, at $\rho = 0$ one from the other task, and intermediate levels average all balanced schedules (exactly 1, 9, 9 and 1 at $\rho = 0, 1/3, 2/3, 1$). The exact integrated correction is positive (amplification) at $\rho = 0$ in all 18 units and negative (damping) at $\rho = 1$ in all 18 units, for both β_2, and ordinary AdamW, whose second moment receives the numerator's own minibatch, damps more strongly still (table 1). The ordering across the three conditions is the ordering of the covariance, and a second moment fed from a randomly chosen task, independent of the numerator, gives a correction of essentially zero (median 0.0000 at both β_2), as part (ii) predicts when V_p is second order. The first-order and exact corrections change sign at the same alignment, to within 0.002 when interpolating between $\rho = 1/3$ and $2/3$ and within 0.01–0.02 for a line through all four levels.

The same assay compares live AdamW with a preconditioner frozen at its checkpoint value, the reference in eq. (2). At $B = 1$, live AdamW has the smaller order dispersion in all 18 units (median log ratio -2.7 at $\beta_2 = 0.999$ and -3.3 at 0.99), and at $B = 8$ in 15 and 16 of 18. Feeding the second moment an independent same-task minibatch removes most of the advantage: the log ratio of live to independent dispersion is -2.0 and -2.5 at $B = 1$, again in all 18 units, and the two constructions coincide exactly at full batch, where there is no sampling to couple.

5.2 BREAKING THE PAIRING AT MATURE CONVOLUTIONAL CHECKPOINTS

We repeat the independent-sample intervention at five ResNet-18 CIFAR-100 checkpoints trained for 20,000 AdamW updates. For a task pair, the *matched-independent* arm keeps task identity, fit population, the numerator minibatch stream, four minibatch banks and eight uniform order words, and replaces only the minibatch used for each second-moment update by an independent draw from the same task, attached to the same occurrence in every order. We report live/matched-independent ratios of the conditional order variance of the parameters, of predicted probabilities and of loss, and of the total parameter variance, at horizons of 6, 20 and 100 updates.

Breaking the pairing increases order dispersion on all three pairs (table 2). For parameter, prediction and total variance the ratio is below one at every horizon in every checkpoint of every pair (4/4, 5/5, 5/5; one-sided sign test $p = 1/16$ or $1/32$ for each pair and readout). Loss variance, the noisiest readout, is below one in 38 of 42 checkpoint–horizon cells. The effect grows with the horizon: at $H = 100$ live coupling reduces conditional parameter variance by 18–22% on every pair. Two caveats bound the interpretation. The checkpoints are shared across pairs, so the effective sample is five networks. And the independent arm also takes larger steps (theorem 2); we did not run a step-matched control, so part of the increase may reflect step size rather than the covariance itself.

5.3 BREAKING THE PAIRING IN A PRETRAINED LANGUAGE MODEL

We apply the intervention to pretrained Pythia-160M (Biderman et al., 2023) with the task pair GitHub code / Wikipedia. A 300-step AdamW warm phase on a mixture of both domains ($B = 16$ sequences of 256 tokens, learning rate 10^{-5}, $\beta_1 = 0.9$, $\beta_2 \in \{0.95, 0.999\}$, fp32) builds a mature optimizer state from the pretrained weights; four warm-phase seeds per β_2 are the units. The assay uses the ResNet design (four banks, eight orders) at $H \in \{6, 20\}$. Input-embedding rows receive token-sparse gradients, so an independent minibatch would leave the rows of absent tokens without a denominator; we keep both embedding tables live in both arms and intervene on the 124M weights through which every token's gradient flows, a conservative choice. Feeding the numerator's own minibatch to the second moment (an identity control) reproduces the live trajectory to numerical

Table 1: Median integrated adaptive correction (18 units) as the second moment's input moves from the other task's minibatch ($\rho = 0$) to an independent minibatch of the same task ($\rho = 1$) to the numerator's own minibatch (ordinary AdamW). Positive values mean amplification; the sign is positive at $\rho = 0$ and negative at $\rho = 1$ in every unit.

β_2	$\rho = 0$	$1/3$	$2/3$	1	same sample	zero crossing: exact / first order
.99	$+.145$	$+.043$	$-.030$	$-.078$	$-.431$	.532 / .530
.999	$+.071$	$+.022$	$-.018$	$-.050$	$-.268$	.519 / .518

Table 2: Live/matched-independent variance ratios at mature ResNet-18 checkpoints; below one means that breaking the pairing increases dispersion. Checkpoints are the units (four for the first pair, whose run on the fifth checkpoint did not complete, and five for the others); brackets are exact checkpoint-resampling intervals and are descriptive.

pair	H	parameters	predictions	loss	total
carnivores/vehicles	6	.910 [.899,.921]	.967 [.959,.974]	.958 [.926,.983]	.918 [.905,.932]
	20	.914 [.899,.929]	.967 [.952,.983]	.800 [.674,.927]	.920 [.903,.937]
	100	.777 [.696,.852]	.951 [.930,.975]	.636 [.438,.757]	.780 [.701,.853]
fish/furniture	6	.898 [.862,.926]	.959 [.947,.971]	1.007 [.972,1.056]	.900 [.855,.929]
	20	.901 [.875,.919]	.964 [.951,.975]	.864 [.718,.968]	.911 [.888,.929]
	100	.819 [.764,.875]	.934 [.906,.962]	.904 [.635,1.263]	.824 [.771,.877]
flowers/insects	6	.889 [.811,.938]	.968 [.962,.976]	.957 [.940,.975]	.904 [.836,.945]
	20	.898 [.862,.927]	.960 [.951,.973]	.936 [.895,.987]	.905 [.868,.936]
	100	.813 [.742,.884]	.923 [.891,.956]	.559 [.432,.703]	.815 [.745,.885]

precision. The learning rate, the dense-weight restriction and the step-bound endpoints were fixed after a pre-assay diagnostic (section D).

We monitor the largest adaptive step over the intervened coordinates at every update (table 3, fig. 2). As theorem 2 guarantees, live AdamW never exceeds κ_t (at most $0.95\,\kappa_t$ over 6,656 updates), which checks the implementation. The decoupled arm exceeds it from the first update: at update 1, between 0.06% and 0.4% of dense coordinates are above κ_t at $\beta_2 = 0.95$, the largest by 1.4–8.3×10^4, and between 0.01% and 0.09% at $\beta_2 = 0.999$, the largest by 1.6×10^2–2.1×10^3. These few coordinates are enough: held-out loss rises from 2.6 to 4.1–5.3×10^3 nats/token after six updates at $\beta_2 = 0.95$ and to 60–374 at $\beta_2 = 0.999$, after which the violation spreads to 2–7% of coordinates. The slower second moment carries more history and so loses less of its normalization when one sample is missing, as the coupling picture predicts.

The bound requires exact coupling. When the second-moment minibatch shares a fraction f of its sequences with the numerator's, the violation shrinks as f grows but persists at every update at $f = 15/16$, where one sequence in sixteen differs (median 5–$7 \times 10^3\,\kappa_t$ at $\beta_2 = 0.95$ and 0.9–$4 \times 10^2\,\kappa_t$ at 0.999); it disappears only at $f = 1$. An AdaShift-style second moment that receives the previous update's gradient exceeds the bound at 19 of 20 updates and raises loss to 31–1,579 nats/token. Because the decoupled arm diverges, order-dispersion ratios in this setting are uninformative; a stable decoupled arm would be needed to test theorem 1 at transformer scale.

5.4 The batch-size target and what transport buys

At the Pythia warm state we estimate q_B coordinatewise from 128 stratified minibatches at each $B \in \{4, 8, 16, 32, 64\}$, fit $q_B = a + c/B$ on $B \in \{4, 16, 64\}$ and predict $B \in \{8, 32\}$. The held-out relative error is 21% and 22%, against 28% and 168% for the best constant, and the fitted intercept matches the directly measured $\mu^{\odot 2}$ (norm ratio 0.89); since two halves of each estimate differ by 21–29%, this error is at the level of the estimator's own noise. Adam's own v lies nearest q_8 (relative distance 0.47, against 0.60 for q_{16} and 2.7 for q_{64}) rather than the training batch size of 16, plausibly because with $\beta_2 = 0.999$ it averages the whole warm phase.

At a live batch-size switch the practical question is whether transporting v improves the model. In three 70M-parameter decoders trained from scratch at $B = 64$ and switched to $B = 256$ (section C),

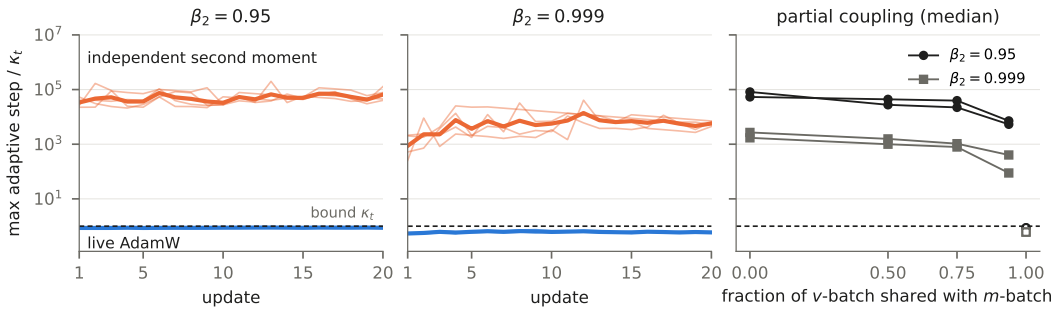

Figure 2: Largest per-coordinate adaptive step relative to the coupled bound κ_t (dashed), pretrained Pythia-160M. Left and centre: along one 20-update word; thin lines are warm-phase seeds, thick lines geometric means. Live AdamW stays below the bound; an independent second-moment sample with the same task and marginal law exceeds it by two to five orders of magnitude from the first update. Right: median over 20 updates when the second-moment minibatch shares a fraction of its sequences with the numerator minibatch (two seeds per β_2); the open markers at 1 are exact sharing.

Table 3: Coupled step bound on pretrained Pythia-160M: per-update maximum adaptive step over the 124M dense weights relative to κ_t, over all words; ranges are over four warm-phase seeds. Loss is held-out cross-entropy (2.5–2.6 at the start of the assay).

β_2	arm	updates above κ_t	median max$/\kappa_t$	largest	loss, $H = 6$	loss, $H = 20$
0.95	live	0 / 3328	0.85–0.88	0.95	2.6	2.6
	matched-independent	3328 / 3328	3.0–5.0×10^4	1.1×10^6	4.1–5.3×10^3	639–745
0.999	live	0 / 3328	0.58–0.62	0.85	2.5–2.6	2.5–2.6
	matched-independent	3328 / 3328	1.2–4.4×10^3	2.1×10^5	60–374	74–144

transport at learning rate 1.5×10^{-4} lowered held-out loss relative to inherited state at the same rate in all three (mean 0.050 nats/token). But the transported state is 0.251 times the inherited one in norm, almost exactly the noise-dominated value $1/4$, so transport doubles every step and at its first update coincides with keeping the state at twice the rate. Against that step-matched control the sources give $+0.004$, -0.032 and one loss spike of the control. The practical content of transport in this regime is the square-root step correction.

6 Discussion

The most instructive comparison in the paper is between the two coupling interventions. On a ResNet-18, replacing the second-moment sample with an independent draw makes training noticeably more order-sensitive, by roughly a fifth at 100 updates, but nothing breaks. On a pretrained language model the same replacement is catastrophic from the first update. Much differs between the two settings, including architecture, learning rate and training history, but the step bound points to one explanation that we find most plausible, though we have not isolated it: the gradient distribution. Adam can afford to divide by a running average because a coordinate that suddenly receives a large gradient also receives it in the denominator. Convolutional gradients rarely test this. Language-model gradients, heavy-tailed and tied to which tokens happen to appear in a batch, test it on a few coordinates at almost every step, and a few are enough.

The partial-coupling experiment sharpens the point. One might expect a second moment computed from a mostly overlapping minibatch to be nearly as good as the real one. It is not: with fifteen of sixteen sequences shared, the step bound still fails at every update. What protects Adam is not a good estimate of $\mathbb{E}[g^2]$ but the presence of the current gradient in the current denominator. This is a useful check for anyone who changes where the second moment comes from, by delaying it as AdaShift and ADOPT do, by estimating it from different samples than the update, or by rescaling or replacing it at an intervention. Such choices can be fine, and methods that make them typically add clipping, as ADOPT, Adafactor and StableAdamW do. Our results explain why that clipping is

needed and suggest where to look when it fires: at coordinates whose current gradient far exceeds their second-moment history.

The batch-size experiments are a cautionary footnote to the same story. The second moment does track minibatch noise, closely enough that a batch-size change leaves it miscalibrated by roughly the batch-size ratio, and correcting this at the switch helps relative to leaving it alone. But in the regimes we tested it is simply the square-root step correction, and the tempting reading that transporting optimizer state is a new ingredient for batch-size schedules is not supported.

Limitations. theorem 1 is a one-step, first-order statement, tested directly only in a small adapter; at transformer scale the decoupled arm diverges, so the order-dispersion prediction is untested there. The ResNet effect is not separated from the larger steps that decoupling produces, which a step-matched control would address. The convolutional study covers one dataset, three task pairs and five checkpoints, and the language-model study one pretrained model, one task pair and horizons of at most 20 updates. All interventions hold the optimizer hyperparameters fixed, including $\epsilon = 10^{-8}$; a larger ϵ or update clipping would soften the language-model result, and how damping scales in large-scale pretraining is open.

7 Conclusion

Adam computes its numerator and its denominator from the same minibatch. This is usually treated as an implementation detail, but it decides whether the second moment damps or amplifies the effect of data order, and it is what keeps Adam's steps bounded when gradients are heavy-tailed. Take it away while holding everything else fixed, and a convolutional network becomes more order-sensitive while a pretrained language model diverges within a few updates, even when the second moment misses only one sequence in sixteen.

AI use statement

Generative AI tools were used to assist with conceptual and theoretical development, mathematical derivations and their critique, experiment design, research-code development and debugging, interpretation of intermediate results, literature search and verification, and manuscript drafting and editing. The authors reviewed and revised all AI-assisted material, verified the cited literature against the original sources, and checked the mathematical and empirical claims against the corresponding derivations, code, and experimental outputs. The authors take responsibility for the final content of the work.

REFERENCES

Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep rl: An empirical study. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.17833.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

Benoit Dherin. Implicit biases in multitask and continual learning from a backward error analysis perspective, 2023. URL https://arxiv.org/abs/2311.00235.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Frederik Kunstner, Alan Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems*, 2024.

Kang Liu and Suyan Li. Finite-horizon input-output dynamics of minibatch perturbations in AdamW, 2026. URL https://arxiv.org/abs/2608.19762.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.10287.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful, 2025. URL https://arxiv.org/abs/2507.07101.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Alexey Rukhovich, Alexander Podolskiy, and Irina Piontkovskaya. Commute your domains: Trajectory optimality criterion for multi-domain learning, 2024. URL https://openreview.net/forum?id=h2TdQ99eR4. NeurIPS Mathematics of Modern Machine Learning Workshop.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, 2018.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order, 2026a. URL https://arxiv.org/abs/2606.24993.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise, 2026b. URL https://arxiv.org/abs/2606.29554.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. In *Advances in Neural Information Processing Systems*, 2024.

Xi Wang and Laurence Aitchison. Batch size invariant adam, 2024. URL https://arxiv.org/abs/2402.18824.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and low-precision training for large-scale vision-language models. In *Advances in Neural Information Processing Systems*, 2023.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, 2020.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. AdaShift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

A Derivations

A.1 Memoryless two-step bracket

Take $\theta_A = \theta - \eta g_A(\theta)$. Then $g_B(\theta_A) = g_B(\theta) - \eta H_B g_A + O(\eta^2)$, and the same for BA; subtracting gives eq. (1).

A.2 Clock phases

Consider bias-corrected normalized momentum, $m_t = \beta m_{t-1} + (1-\beta) g_{d_t}(\theta_{t-1})$, $\widehat{m}_t = m_t/(1-\beta^{s+t})$, $\theta_t = \theta_{t-1} - \eta \widehat{m}_t$. A gradient injected at position r contributes to the endpoint K through the weight

$$W_{r,K}^{(\eta)} = (1-\beta) \sum_{t=r}^{K} \frac{\beta^{t-r}}{1-\beta^{s+t}}. \tag{3}$$

Theorem 3 (Clock-phase asymptotics)**.** *Fix K and r; in plain terms, a gradient's write-in weight scales as $\max(1-\beta, 1/s)$. If $1-\beta(\eta) = c\eta^p + o(\eta^p)$ and $s(\eta) = T\eta^{-q} + o(\eta^{-q})$ with $c, T, p, q > 0$, then $W_{r,K}^{(\eta)} = \kappa_{p,q}(r,K)\, \eta^{\min(p,q)}(1+o(1))$ with*

$$\kappa_{p,q}(r,K) = (K-r+1) \times \begin{cases} T^{-1}, & q < p, \\ c/(1-e^{-cT}), & q = p, \\ c, & p < q. \end{cases}$$

Hence position-dependent replay contributes to the parameter endpoint at $\Theta(\eta^{1+\min(p,q)})$.

Proof. Since $\log\beta = -c\eta^p + o(\eta^p)$, $\beta^{s+t} = \exp[-cT\eta^{p-q} + o(\eta^{p-q})](1+o(1))$. Thus $1-\beta^{s+t}$ is asymptotic to $cT\eta^{p-q}$ when $q < p$, tends to $1 - e^{-cT}$ when $q = p$, and tends to 1 when $p < q$; also $\beta^{t-r} = 1 + o(1)$ for fixed K. Substitute into eq. (3). □

The Θ statement requires the leading positional coefficient $\sum_r \kappa_r [g_{d_r}(\theta_0) - g_{d'_r}(\theta_0)]$, which is proportional to $\sum_r (K-r+1)(g_{d_r} - g_{d'_r})$, to be nonzero; when it is, the replay-induced contrast is $\Theta(\eta^{1+\min(p,q)})$; parameter-induced changes in the gradients enter at higher order for C^2 fields and a shared warm state whose momentum is consistent with its age, so that $\widehat{m} = O(1)$. In AdamW, with $1-\beta_i \asymp \eta^{p_i}$, the two memory channels contribute at $\eta^{1+\min(p_1,q)}$ and $\eta^{1+\min(p_2,q)}$; after grouping terms of equal order, the smallest offset with a nonzero coefficient sets the observed exponent; we state this extension without proof and check it numerically (section A.3). Heuristically, fixed decay rates are the limit $p \to 0$ (outside the theorem's hypotheses), where the exponent tends to one and memory makes order a first-order effect, as in Sweeney (2026b); linear scaling $p = q = 1$ restores the η^2 order of the commutator. The law says how the exponent interpolates between these regimes as memory rate and optimizer age are scaled with the learning rate.

A.3 Active AdamW channels

Write $1-\beta_i \asymp \eta^{p_i}$ and $s \asymp \eta^{-q}$, and set $r_m = \min(p_1, q)$, $r_v = \min(p_2, q)$. A local AdamW word contrast organizes as $\Delta_\theta(\eta) = \eta^{1+r_m} A_m + \eta^{1+r_v} A_v +$ higher-order terms. Group all terms sharing an exponent before testing cancellation, and let $r_\star$ be the smallest offset whose grouped coefficient is nonzero; then $\|\Delta_\theta(\eta)\| = \Theta(\eta^{1+r_\star})$. Generically $r_\star = \min(p_1, p_2, q)$. When the two channels have different powers, their measured amplitudes locate the crossover between them.

A.4 Second-moment geometry beyond mean gradients

Let $z = (\tau, \theta, m, v)$, $R_\tau(m, v) = \frac{m/(1-e^{-a\tau})}{\sqrt{v/(1-e^{-b\tau})}+\epsilon}$ componentwise, g_D the mean gradient of domain D and $q_D = \mathbb{E}[g^{\odot 2}_{\text{sample}} \mid D]$. In the linear scaling regime $1-\beta_1 = a\eta$, $1-\beta_2 = b\eta$, $s = \tau_0/\eta$, AdamW follows the mean-field vector field $X_D(z) = (1,\ -R_\tau(m,v) - \lambda\theta,\ a(g_D - m),\ b(q_D - v))$ with decoupled weight decay λ. Write $X_D = F + G_D$ with $G_D = (0, 0, ag_D, bq_D)$. Because G_D has no θ component while g_D, q_D depend only on θ, $DG_B\, G_A = 0$ and $[G_A, G_B] = 0$. Hence, with $[X, Y] = DY\, X - DX\, Y$ and $\Phi_D = [F, G_D]$, $[X_A, X_B] = \Phi_B - \Phi_A$, and the first-order cycle curl vanishes. With $F^\theta = -R_\tau - \lambda\theta$, $P_m = \partial R_\tau/\partial m$ and $P_v = \partial R_\tau/\partial v$, $\pi_\theta[F, G_D] = aP_m g_D + bP_v q_D$, so

$$\pi_\theta[X_A, X_B] = aP_m(g_B - g_A) + bP_v(q_B - q_A). \tag{4}$$

Two domains with identical mean gradients but different raw second moments therefore still produce an $O(\eta^2)$ order effect, through a direction outside any mean-gradient analysis. This mean-field calculation contains no minibatch sampling and hence no coupling; it shows only that the second moment carries domain information, and at a real network that component is small (section B).

A.5 Projected signed feedback

For a fixed covector h, write $a_w = h^\top(\bar{p} \odot \delta m_w)$ and $b_w = h^\top(\bar{m} \odot \delta p_w)$. Then $\mathrm{Var}_w(a_w + b_w) - \mathrm{Var}_w(a_w) = \mathbb{E}_w b_w^2 + 2\mathbb{E}_w[a_w b_w]$, the scalar version of eq. (2).

A.6 From local feedback to endpoint dispersion

Ignoring weight decay, the endpoint perturbation sums preconditioned updates $u_t(w)$, so $\mathrm{Var}_w\left(\sum_t \eta_t u_t(w)\right) = \sum_{t,r} \eta_t \eta_r \,\mathrm{tr}\,\mathrm{Cov}_w(u_t, u_r)$. Theorem 1 controls the immediate diagonal part; cross-time covariance and the gradients met after trajectories separate decide whether the local sign persists, which is why the coupling interventions measure endpoint dispersion directly.

B Additional results

Clock phases. A 1,025-parameter tanh regressor alternates two full-batch synthetic domains while η, $1-\beta$ and the initial clock are scaled as in theorem 3, from ten initializations and from zero and warm momentum states. This is a deterministic float64 check of the asymptotics, so we report fitted slopes without intervals: every slope lies within 10^{-3} of its prediction (table 4), warm momentum leaves the exponent unchanged, and when the first- and second-moment channels have different powers their measured amplitudes predict the crossover between them to within a factor of 1.10.

Table 4: Deep-range clock-law slopes (identical for zero and warm momentum to four decimals).

cell	predicted	fitted
GD	2.0	2.0000
$p = .5, q = 1$	1.5	1.4994
$p = 1, q = .5$	1.5	1.4992
$p = 1, q = 1$	2.0	2.0000
$p = 2, q = 1$	2.0	2.0000

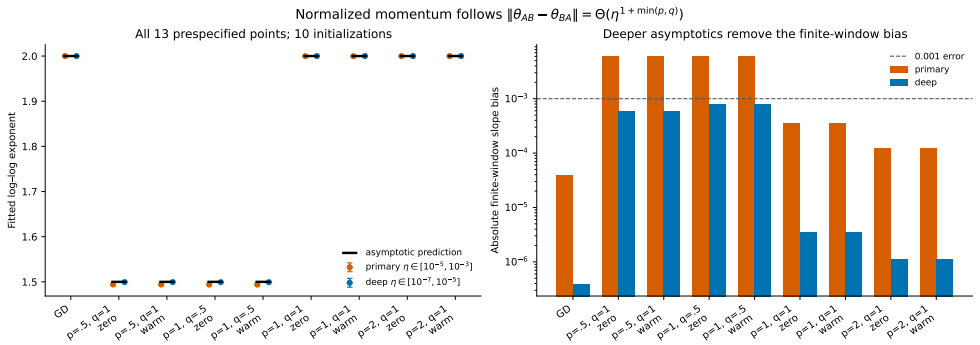

Figure 3: Clock-phase scaling experiment.

Second-moment-exclusive geometry. We build domains with sample gradients $g \pm r$ and $g \pm s$: equal mean gradients, different raw second moments. Across 12 seeds a nested-JVP calculation and the closed form eq. (4) agree to relative error below 9×10^{-15}, and at least 82.4% of the bracket norm lies outside the common mean-gradient span. At an ImageNet-pretrained ResNet-18 (final head, 640 CIFAR-10 images), adding the raw-q directions raises numerical rank from 9 to 18 at relative tolerance 10^{-4} (one dimension at 10^{-3}); the component outside the mean-gradient span is only 0.263–0.339% of each bracket column, so this direction exists at a real network but is small.

C LIVE BATCH-SIZE SWITCH AT 70M

Each source is a Pythia-style 70M decoder (six layers, width 512) trained from random initialization for 500 updates at $B = 64$ (AdamW, $\beta_2 = 0.999$, weight decay 0.1, peak rate 3×10^{-4}). At the switch to $B = 256$ three branches restore the same checkpoint and consume the same 200 ordered batches: inherited state at 3×10^{-4} (keep), inherited state at 1.5×10^{-4} (keep, half rate), and bias-corrected transport $\widehat{v} \leftarrow \frac{1}{4}\widehat{v} + \frac{3}{4}\widehat{m}^{\odot 2}$ at 1.5×10^{-4}.

Table 5: Held-out loss (nats/token) after 200 updates at $B = 256$. The transported state's norm is 0.2510–0.2514 times the inherited norm in every source.

source	at switch	keep, 3×10^{-4}	keep, 1.5×10^{-4}	transport, 1.5×10^{-4}
1	5.111	5.047	5.087	5.050
2	5.171	4.998	5.067	4.966
3	5.064	6.665	5.032	5.019

Effective steps. With $\epsilon \ll \sqrt{v}$, a state scaled by 0.251 multiplies every adaptive step by 1.996, so transport at 1.5×10^{-4} matches keep at 3×10^{-4} at the first update, up to the $\frac{3}{4}\widehat{m}^{\odot 2}$ term and half the decoupled weight decay. Afterwards the inherited state relaxes toward the new target: with $v_{\text{old}} \approx 4q_{256}$ in noise-dominated coordinates, $v_{200} \approx (0.999^{200} \cdot 4 + 1 - 0.999^{200})\, q_{256} = 3.46\, q_{256}$, so the inherited state's step grows by only 7.6% over the continuation. The half-rate keep arm therefore runs at roughly half the step size of the other two arms throughout, which accounts for its higher loss. Relative to the step-matched keep arm, transport is $+0.004$ and -0.032 nats/token in sources 1 and 2; in source 3 the keep arm's loss spiked between updates 100 and 200. An earlier exploratory pilot that applied transport without halving the rate, on a 160M model overtrained on a small corpus, raised loss, consistent with a doubled step.

D EXPERIMENTAL DESIGN AND STATISTICAL UNITS

table 6 lists the full configuration of every experiment; code, frozen designs, and compact result records accompany the submission.

Clock phases. Ten initializations are the units; learning-rate grid points within one initialization trace one curve. All grid points are kept.

Sign rule. Six task pairs are top-level clusters with three nested checkpoint/split seeds. The conditional contrast holds a minibatch bank fixed and changes only the order. In the independent-q arm each task occurrence draws a fresh second-moment minibatch from the same population; at full batch the two arms coincide exactly, which the code verifies.

Mature coupling. Independently trained step-20,000 ResNet-18 checkpoints (GroupNorm, AdamW $\eta = 10^{-3}$, $(\beta_1, \beta_2) = (.9, .999)$, $\epsilon = 10^{-8}$, weight decay 10^{-4}, $B = 64$) are the units. Each pair uses four independent banks, eight uniform orders, and horizons 6, 20, 100. The three pairs are those fixed in the experimental configuration; none was added or dropped after outcomes were seen. Deterministic kernels are used throughout.

Language-model coupling. Warm-phase seeds (independent warm data orders and bank draws from the shared pretrained model) are the units. The design, including the dense-weight restriction, the learning rate, and the step-bound endpoints, was fixed after a pre-assay diagnostic that showed (a) bf16 autocast raises Pythia's loss by 0.4 nats and makes attention backward passes nondeterministic, so all runs use fp32 with TF32 matmuls, and (b) learning rates of 10^{-4} and 3×10^{-5} degrade the pretrained model during the warm phase. A repeated live rollout measures the numerical floor.

Live 70M switch. Three random-initialization sources (six layers, width 512) trained for 500 updates at $B = 64$ on 347 GitHub documents are the units; each continuation consumes the same 200 ordered $B = 256$ batches and is evaluated on 5,263 packed sequences from 59 held-out documents.

Table 6: Reproducibility summary. Values are copied from the frozen experiment configurations and result manifests; the listed units are the independent statistical units.

experiment	configuration and statistical units
clock phase	1,025-parameter tanh regressor (input 14, hidden 64, scalar output), two fixed 128-example domains, float64 CPU, $c = T = 1$; units: ten initialization seeds 12001–12010; 13 learning-rate points in each primary and deep grid; 20,000 bootstrap resamples of initializations.
sign rule	1,040-parameter adapter (64 projected features, hidden size 16) on ImageNet-1K ResNet-18 V1 features; six CIFAR-100 superclass pairs × checkpoint seeds 5–7; 20 balanced length-six words; primary $B = 8$ with 576 repetitions, plus $B = 1$ (576 repetitions) and full batch; $\beta_2 \in \{.99, .999\}$, $\eta = 3 \times 10^{-4}$, $\epsilon = 10^{-8}$, weight decay 10^{-3}.
ResNet coupling	ResNet-18 (GroupNorm, 3×3 stem, no max-pool) trained 20,000 AdamW updates at $B = 64$ on CIFAR-100 ($\eta = 10^{-3}$, $(\beta_1, \beta_2) = (.9, .999)$, $\epsilon = 10^{-8}$, weight decay 10^{-4}); units: checkpoint seeds 2026091201, 1202, 1205, 1206 (large carnivores/vehicles 1) and additionally 1204 (fish/household furniture, flowers/insects); per pair: four independent banks, eight uniform orders, $H \in \{6, 20, 100\}$, $B = 64$; 1,000 evaluation examples; deterministic kernels; arms live and occurrence-matched independent second moment.
Pythia coupling	Pythia-160M, revision step143000 (commit b56d9be), 162M parameters of which 124M dense (intervened); fp32 with TF32 matmuls, no gradient clipping; GitHub/Wikipedia from a hashed document manifest (SHA-256 39722358…), non-oracle documents split by document-hash parity into warm (GitHub 220 documents / 48,691 windows; Wikipedia 236 / 5,215) and assay pools (221 / 72,460; 220 / 5,118) of 256-token windows; evaluation on 16 windows from each domain's held-out oracle documents (59 and 44). Warm phase 300 updates at $B = 16$ (8+8), learning rate 10^{-5} with 30-step linear warmup, $\beta_1 = 0.9$, $\epsilon = 10^{-8}$, weight decay 0.01; assay at learning rate 10^{-5}; units: warm-phase seeds 1–4 for each $\beta_2 \in \{0.95, 0.999\}$; four banks, eight uniform orders, $H \in \{6, 20\}$; one repeated live word per horizon measures the numerical floor.
Pythia q_B law	Same model and warm recipe ($\beta_2 = 0.999$, seed 1); 128 stratified minibatches ($B/2$ per domain) at each $B \in \{4, 8, 16, 32, 64\}$ from the assay pools; fit $B \in \{4, 16, 64\}$, held out $B \in \{8, 32\}$; split halves of 64 minibatches.
live 70M	Pythia-style 70M (six layers, width 512, eight heads, MLP width 2048, vocabulary 50,304, context 1,024); units: three random-initialization seeds 20260924–26; 500 source updates at $B = 64$ on 347 GitHub documents (AdamW, peak 3×10^{-4}, $\beta_2 = 0.999$, weight decay 0.1); three continuations of 200 updates at $B = 256$ on identical ordered batches; evaluation on 5,263 packed sequences from 59 held-out documents.