When and How Does Muon Precondition? The Role of Structured Gradient Noise

Anonymous authors
Paper under double-blind review

Abstract

Muon has become a popular optimizer for large language model training, yet its theoretical foundations remain poorly understood. In this work, we study how gradient noise geometry shapes Muon's preconditioning benefits. We first show, theoretically and empirically, that under the classical isotropic noise assumption, Muon offers no acceleration beyond a rescaling of the learning rate. This motivates us to investigate structured noise arising from mini-batch sampling in associative memory with heterogeneous task frequencies. We establish an exact balancing threshold for two tasks and extend the analysis to multiple tasks, revealing how task coverage and gradient reliability jointly shape learning. Our analysis shows that Muon can soften the dependence on task frequency from linear to square-root, helping balance the learning of frequent and rare tasks. We further identify rotation-induced negative transfer in SignSGD that Muon avoids, and explain how momentum reduces sampling noise while tracking changing gradients. Controlled numerical experiments validate these mechanisms, and a preliminary LLM pretraining study illustrates the practical promise of fast–slow momentum.

1 Introduction

Muon has emerged as a prominent alternative to Adam for training the matrix parameters of large language models (LLMs) (Jordan et al., 2024; Liu et al., 2025). Its defining operation replaces a gradient or momentum matrix by an approximate polar factor, flattening its nonzero singular values before the parameter update. This matrix-level normalization offers a different way to allocate updates across directions from Adam's coordinatewise rescaling. Recent frontier-model reports, including DeepSeek–V4.1–Flash (DeepSeek-AI, 2026) and Kimi–K3 (Kimi Team et al., 2026), adopt Muon variants for matrix optimization, with head-wise orthogonalization further adapting the update to attention structure. These developments make understanding when and why Muon improves learning an increasingly practical theoretical question.

Existing theory has established convergence guarantees for practical orthogonalization procedures (Kim & Oh, 2026) and identified preconditioning benefits in structured optimization problems (Ma et al., 2026). Full-batch associative-memory analyses further show how spectral normalization can balance learning across tasks with heterogeneous frequencies (Wang et al., 2026; Li et al., 2026a). In LLM pretraining, however, Muon acts on stochastic mini-batch gradients, and nonlinear normalization need not preserve the benefits predicted from population gradients. In particular, writing msgn for the partial polar map, generally $\mathbb{E}[\mathrm{msgn}(G_t) \mid W_t] \neq \mathrm{msgn}(\mathbb{E}[G_t \mid W_t])$. Recent work has begun to characterize stochastic spectral dynamics and geometry-dependent noise scales (Paquette et al., 2026; Naganuma et al., 2026); the interaction between task occurrence, gradient reliability, and temporal averaging motivates the question we study here:

> **Question.** *When does Muon retain a directional learning advantage under stochastic gradient noise, and how do noise structure, batch size, and momentum determine this advantage?*

Isotropic noise limits preconditioning. Standard stochastic optimization analyses often control gradient noise through a scalar variance bound (Ghadimi & Lan, 2013), leaving its directional structure unspecified. We use isotropic Gaussian noise as a diagnostic baseline. In the fixed-dimensional low-SNR limit, we prove that Muon's conditional mean update is, to first order, a scalar multiple of

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.

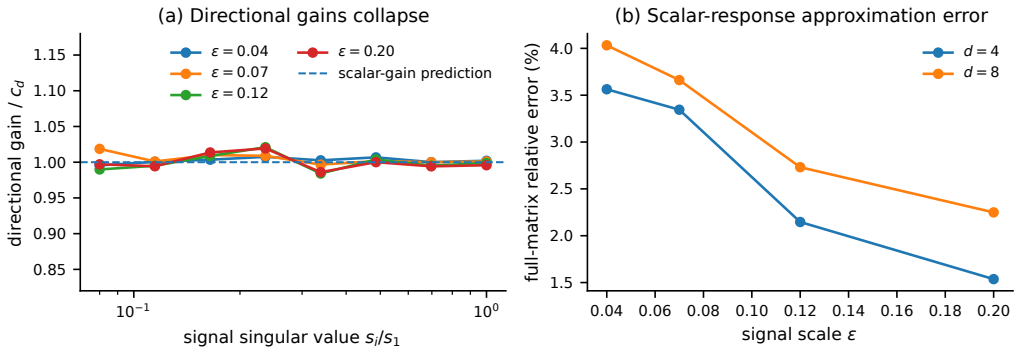


Figure 1: **Isotropic noise removes first-order directional preconditioning at low SNR.** Left: for an anisotropic $d = 8$ signal, the response gain along each singular direction, normalized by c_d, collapses near one across a wide range of signal singular values. Right: the full conditional-mean matrix remains within a few percent of the scalar-response prediction over the tested low-SNR amplitudes.

the population gradient, as confirmed numerically in Figure 1. SGD with a matched learning rate therefore reproduces this first-order mean drift: applying the partial polar map to each stochastic mini-batch gradient does not, by itself, produce directional preconditioning in the first-order mean response. Explaining such gains in this regime requires examining noise structure.

Sampling controls balancing. We study associative memory with heterogeneous task frequencies p_j (Wang et al., 2026; Li et al., 2026a). Mini-batch sampling can leave task directions entirely unobserved; without momentum, Muon's nullspace preservation then prevents updates along those directions. For two tasks, we prove an exact balancing threshold B_{bal}: below it, margins grow at unequal linear rates; above it, their gap has uniformly bounded exponential moments. For rare tasks, $B_{\mathrm{bal}} \sim (\log 2)/p_{\min}$. In a higher-rank hierarchy, local perturbation estimates identify the expected task count Bp_j as a controlling scale. Experiments exhibit corresponding frequency-dependent crossovers, with frequent groups approaching balanced growth while rare groups remain occurrence-limited.

Label noise limits reliability. Label fluctuations persist at the population optimum, so frequent observation alone does not ensure a reliable gradient. Exact gradient statistics and a scalar projected counting model distinguish regimes governed by task counts Bp_j and corruption counts $Bp_j\rho_\alpha$. Near the noisy optimum, an isotropic covariance closure with a locally homogeneous frequency approximation predicts a Muon contraction coefficient proportional to $\sqrt{Bp}$, compared with SGD's linear dependence on the local frequency scale p.

Orthogonal equivariance differs from coordinatewise sign updates. We use SignSGD as a proxy for Adam's coordinatewise adaptivity (Balles & Hennig, 2018; Kunstner et al., 2023). Muon is orthogonally equivariant whereas SignSGD is basis dependent. In a two-task rotated sparse-support construction, a batch containing only the frequent task leaves the rare-task Muon margin unchanged but can decrease the corresponding SignSGD margin under certain conditions.

Momentum trades averaging for tracking. Momentum reduces sampling fluctuations, with frozen-state variance-equivalent batch size $B_{\mathrm{eff}} = B(1+\beta)/(1-\beta)$, but longer memory introduces tracking lag. Under local tracking assumptions, the sufficient window $\eta \ll 1-\beta \ll Bp_j$ motivates longer memory for rare, slowly varying directions and shorter memory where rapid tracking is needed. This provides a sampling-based perspective on curvature-conditioned multiscale momentum (Zhu et al., 2026): the local Hessian block $p_j\Gamma_j$ links rare tasks to low-curvature directions when label statistics are comparable. It also connects to occurrence-dependent forgetting in Activation-Keyed Momentum (Hong & Qu, 2026), supporting selective long-memory averaging as a means to improve estimates in slowly varying directions.

Our contributions are threefold:

- **A diagnostic for stochastic directional gains.** We prove a low-SNR expansion under isotropic Gaussian noise, showing that Muon's first-order conditional mean response can be matched by learning-rate-rescaled SGD despite samplewise spectral normalization.
- **A theory of structured sampling.** We establish an exact two-task balancing threshold and derive local higher-rank sampling estimates. With stochastic labels, we separate exact gradient statistics, scalar projected counting laws, and a square-root frequency prediction under isotropic covariance and locally homogeneous frequency approximations. A momentum-free Muon–SignSGD comparison isolates rotation-induced negative transfer under sparse sampling.
- **A statistical basis for heterogeneous momentum.** We characterize frozen-state variance reduction and local tracking error under temporal averaging, explaining when longer memory can help rare directions and providing a sampling-based interpretation of fast–slow momentum designs. Controlled experiments test these mechanisms, and a preliminary pretraining comparison illustrates their practical relevance.

2 Related Work

Muon and preconditioning. Muon applies approximate polar normalization to matrix-valued momentum (Jordan et al., 2024; Liu et al., 2025). Existing work establishes convergence with Newton–Schulz iterations (Kim & Oh, 2026), preconditioning benefits in structured matrix problems (Ma et al., 2026), and generalization gains from balanced learning on imbalanced data (Vasudeva et al., 2026). Magakyan et al. (2026) further identify scaling laws for momentum spectra and their implications for orthogonalization accuracy. We complement these perspectives by studying how stochastic gradient structure governs Muon's conditional mean response.

Stochastic spectral dynamics. Paquette et al. (2026) characterize batch-dependent preconditioning regimes for Muon and SignSGD in high-dimensional matrix least squares. Naganuma et al. (2026) derive non-Euclidean gradient-noise scales for adaptive batch selection, while Eschenhagen et al. (2026) examine how stochasticity and temporal adaptation distinguish Shampoo from spectral descent. Our analysis focuses on random task support and persistent label fluctuations: we identify an exact finite-batch balancing threshold and separate task occurrence from gradient reliability. The isotropic Gaussian result provides a complementary fixed-dimensional diagnostic for the absence of first-order directional preconditioning.

Associative memory learning. Wang et al. (2026) connect Muon's advantage over Adam to balanced learning of rare associations. Li et al. (2026a) analyze hierarchical task frequencies, label noise, and scaling laws through implicit matrix preconditioning. Beyond orthogonal embeddings, Kim et al. (2026) derive recovery and capacity scaling with Gaussian embeddings, including critical batch-size behavior. Building on the orthogonal associative-memory framework, we isolate finite-batch sampling effects through exact two-task dynamics and a higher-rank frequency-dependent regime map. Our SignSGD comparison further identifies rotation-induced negative transfer when tasks are absent from a batch.

Heterogeneous momentum. Muon momentum has been studied as spectral filtering (Li et al., 2026b), while activation-keyed and curvature-conditioned methods adapt memory across directions (Hong & Qu, 2026; Zhu et al., 2026). We complement these designs with a sampling-based account of variance reduction and tracking lag.

3 Preliminaries

3.1 Notations.

For $G \in \mathbb{R}^{d\times d}$ with SVD $G = U\Sigma V^\top$, define the partial polar factor $\mathrm{msgn}(G) := U\,\mathrm{diag}(\mathbf{1}\{\sigma_i > 0\})V^\top$, which replaces each nonzero singular value by one and reduces to $UV^\top$ when G is full rank. The map is positively scale invariant, $\mathrm{msgn}(aG) = \mathrm{msgn}(G)$ for $a > 0$, and orthogonally equivariant, $\mathrm{msgn}(QGR^\top) = Q\,\mathrm{msgn}(G)R^\top$ for orthogonal Q, R. It also preserves the right nullspace: $\ker(\mathrm{msgn}(G)) = \ker(G)$.

Given the stochastic gradient G_t, we study the idealized Muon dynamics

$$M_t = \beta M_{t-1} + (1-\beta)G_t, \qquad W_{t+1} = W_t - \eta\,\mathrm{msgn}(M_t),$$

where $\eta > 0$ is the learning rate and $\beta \in [0, 1)$ is the momentum coefficient; $\beta = 0$ recovers Muon without momentum. Our theoretical analysis uses exact partial polar normalization in task coordinates; rectangular representations are introduced where needed. Unless stated otherwise, mini-batches are independent across iterations and weight decay is omitted. The pretraining experiment uses the practical construction cited in Section 6.

3.2 Warmup: Isotropic Noise Limits Preconditioning

Consider momentum-free Muon. Condition on the current iterate W and write $G = \bar{G} + \sigma\Xi$, where $\bar{G} = \nabla\mathcal{L}(W)$ and $\Xi \in \mathbb{R}^{d\times d}$ has i.i.d. standard Gaussian entries. Although each polar update has a flat nonzero spectrum, its mean need not retain a preconditioning benefit.

Proposition 3.1 (Isotropic Gaussian mean response)**.** *Fix d and let $c_d = d^{-2}\mathbb{E}\|\Xi\|_* > 0$. As $\|\bar{G}\|_F/\sigma \to 0$,*

$$\mathbb{E}[\mathrm{msgn}(\bar{G} + \sigma\Xi)] = \frac{c_d}{\sigma}\bar{G} + O_d\left(\frac{\|\bar{G}\|_F^3}{\sigma^3}\right),$$

with the remainder measured in Frobenius norm.

At the same parameter state, Muon with learning rate η_{M} has first-order conditional mean update $-\eta_{\mathrm{M}}(c_d/\sigma)\bar{G}$, matching SGD with $\eta_{\mathrm{S}} = \eta_{\mathrm{M}}c_d/\sigma$. Thus its effective first-order preconditioner is scalar: all singular directions receive the same multiplicative gain, leaving their relative scaling unchanged. This equivalence concerns conditional means, not update covariances or stochastic trajectories. Figure 1 confirms the common directional gain for an anisotropic signal. Appendix A gives the proof via rotational symmetry. In the fixed-dimensional low-SNR regime, directional preconditioning in the first-order mean response therefore requires structure absent from this isotropic benchmark, motivating our analysis of structured sampling.

4 Structured Sampling in Associative Memory

4.1 Associative memory and stochastic gradients

We study a linear associative memory (Wang et al., 2026; Li et al., 2026a) that maps query embeddings to answer logits through a trainable matrix W. There are $K \geq 2$ associations $(E_j, \widetilde{E}_j)$, sampled with probabilities $p_j > 0$, where $\sum_j p_j = 1$. Write $P = \mathrm{diag}(p_1, \ldots, p_K)$. Let $E = [E_1, \ldots, E_K]$ and $\widetilde{E} = [\widetilde{E}_1, \ldots, \widetilde{E}_K]$, with $E^\top E = \widetilde{E}^\top\widetilde{E} = I_K$. For query E_j, answer i has logit $\widetilde{E}_i^\top W E_j = \overline{W}_{ij}$, where $\overline{W} = \widetilde{E}^\top W E$. The prediction probabilities and deterministic-label population loss are

$$S_{ij} = \frac{\exp(\overline{W}_{ij})}{\sum_{k=1}^K \exp(\overline{W}_{kj})}, \qquad \mathcal{L}(W) = -\sum_{j=1}^K p_j \log S_{jj}.$$

For a batch of size B, let $q_j = N_j/B$ be the empirical task frequency and $Q = \mathrm{diag}(q)$, where $Bq \sim \mathrm{Multinomial}(B, p)$. The negative stochastic gradient is $\widetilde{E}(I - S)QE^\top$ (Appendix B). Orthogonal equivariance of the partial polar map gives the exact task-coordinate Muon update without momentum:

$$\overline{W}_{t+1} = \overline{W}_t + \eta\ \mathrm{msgn}\big((I - S_t)Q_t\big). \tag{1}$$

If task j is absent, $q_j = 0$, its gradient column in task coordinates is zero. Nullspace preservation then makes the corresponding polar-update column zero. This absence guarantee applies to the momentum-free update; historical momentum can remain nonzero when a task is absent from the current batch.

4.2 TWO TASKS: AN EXACT FINITE-BATCH BALANCING TRANSITION

With $K = 2$, the stochastic gradient is rank one, which makes the finite-batch sampling effect exactly solvable. Let $p_1 \geq p_2 =: p_{\min} > 0$ and let Δ_1, Δ_2 be the two target-vs.-nontarget margins.

Theorem 4.1 (Two-task finite-batch balancing transition)**.** *Assume $\Delta_1(0) = \Delta_2(0) = 0$ and fix $\eta > 0$. Let $B_{\mathrm{bal}} = \log 2/[-\log(1 - p_{\min})]$. If $B < B_{\mathrm{bal}}$, then $(\Delta_1, \Delta_2)/(\eta t) \to \sqrt{2}\,(p_1^B, 1 - p_1^B)$ almost surely. If $B > B_{\mathrm{bal}}$, then $\sup_t \mathbb{E}e^{a|\Delta_1 - \Delta_2|} < \infty$ for some $a > 0$ and $(\Delta_1 - \Delta_2)/t \to 0$ almost surely.*

The threshold is controlled by the probability p_1^B of a batch containing only the frequent task. As its margin pulls far ahead, a pure task-1 batch advances only Δ_1, while every batch containing task 2 asymptotically favors the rare task. The boundary $p_1^B = 1/2$ separates outward from restoring tail drift. For $p_{\min} \ll 1$, $B_{\mathrm{bal}} \sim (\log 2)/p_{\min}$: batches of order the inverse tail frequency suffice for equal growth to leading linear order. The critical case $B = B_{\mathrm{bal}}$ is not covered by the theorem.

4.3 HIGHER RANK: FREQUENCY-DEPENDENT DYNAMICAL REGIMES

We next ask how the two-task results extend when many associations occur at different population frequencies. To represent multiple frequency scales while retaining a tractable matrix structure, we group the $K = MC$ associations into M blocks of size C. Block μ has total probability $\widetilde{p}_\mu$, distributed uniformly among its C tasks, so that

$$p_j = \frac{\widetilde{p}_\mu}{C}, \qquad j \in \mu, \qquad \sum_{\mu=1}^{M} \widetilde{p}_\mu = 1.$$

This block model can be viewed as a coarse-grained representation of a heterogeneous frequency distribution: different blocks correspond to different frequency levels, while tasks within the same block are statistically exchangeable. We take $\overline{W}_0 = 0$. Full-batch dynamics preserve this block symmetry, whereas mini-batch fluctuations generally break it. Let $U_\mu \in \mathbb{R}^{K \times (C-1)}$ have orthonormal columns spanning the within-block contrast subspace.

Proposition 4.2 (Higher-rank sampling and local polar response)**.** *For the momentum-free dynamics: (i) along the full-batch trajectory from $\overline{W}_0 = 0$, every within-block contrast receives the unit polar response, $U_\mu^\top \operatorname{msgn}((I - S)P)U_\nu = \delta_{\mu\nu} I_{C-1}$; (ii) task j is observed with probability $1 - (1 - p_j)^B \sim Bp_j$ as $Bp_j \to 0$ and receives no update when absent; (iii) at a fixed block-symmetric reference state, the contrast-to-contrast derivative with respect to within-block sampling fluctuations vanishes. The first-order contrast-to-average coupling for block μ is $O_{\mathbb{P}}((Bp_j)^{-1/2})$, $j \in \mu$, under the well-sampled assumptions of Appendix D.5.*

These results identify a frequency-dependent sampling scale: rare tasks face an exact occurrence constraint, while well-sampled blocks have small local coupling at a symmetric reference state. They do not establish uniform long-time balancing of the stochastic hierarchy. Appendix D.7 separately proves finite-horizon convergence to the full-batch flow under large-batch and small-step limits.

Figure 2(b) shows corresponding trajectory-level crossovers: measured late margin-growth rates approximately align against Bp_j and approach the full-batch response as expected task counts increase. Frequent and rare groups can therefore exhibit different empirical regimes at the same batch size.

Appendices C and D give the exact two-task proof and the higher-rank perturbation analysis underlying the approach to balanced dynamics.

4.4 LABEL NOISE: THREE STOCHASTIC REGIMES

We next make the observed answer stochastic, with $0 < \alpha < 1$. Conditional on task j, symmetric label noise leaves the target label unchanged with probability $c_\alpha = 1 - \alpha + \alpha/K$ and returns a nontarget label with aggregate probability $\rho_\alpha = 1 - c_\alpha$. The crucial difference from task sampling is that label fluctuations remain nonzero at the finite population optimum.

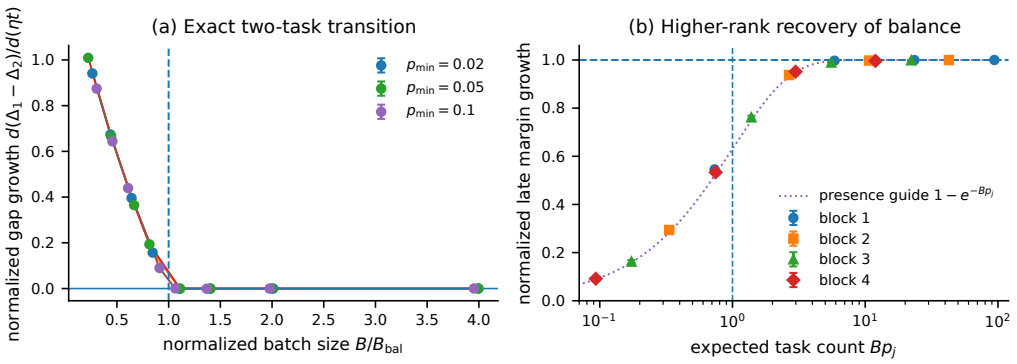


Figure 2: **Mini-batch sampling controls the transition from frequency-limited to balanced growth.** Left: the two-task gap-growth transition occurs at the exact threshold $B/B_{\text{bal}} = 1$. Right: in the higher-rank hierarchy, late margin-growth rates from different frequency blocks collapse when plotted against the expected task count Bp_j and approach the full-batch unit response as that count increases. Error bars summarize 64 trajectories; the dotted curve is the task-presence guide $1 - e^{-Bp_j}$.

Proposition 4.3 (Task frequency and gradient reliability decouple under label noise)**.** *Let d_j be the conditional label distribution, $\delta_j = Se_j - d_j$, and $\Gamma_j = \operatorname{diag}(d_j) - d_j d_j^\top$. For the stochastic gradient column $\widehat{g}_j$, define $\mathrm{SNR}_j^2 = \|\mathbb{E}\widehat{g}_j\|_2^2 / \operatorname{tr}\operatorname{Cov}(\widehat{g}_j)$. With deterministic labels and nonzero residual, $\mathrm{SNR}_j^2 = Bp_j/(1-p_j)$. With nondegenerate label noise, $\mathrm{SNR}_j^2 \sim Bp_j\|\delta_j\|_2^2 / \operatorname{tr}\Gamma_j \to 0$ as $\delta_j \to 0$.*

With deterministic labels, the stochastic fluctuation of a task shrinks together with its residual, so the number of task observations is enough to characterize directional reliability. Label noise breaks this one-scale picture: near the population optimum the mean residual vanishes, while the random observed label continues to fluctuate. A task can therefore be sampled many times ($Bp_j \gg 1$) and still have a low-SNR gradient.

Label noise introduces a second sampling scale, $Bp_j\rho_\alpha$, the expected number of corrupted task-j observations. When ρ_α is small, it can be much smaller than the task count Bp_j, leaving a broad intermediate range in which the task appears regularly but corruptions are still rare.

Proposition 4.4 (Projected regimes and local closure under label noise)**.** *Assume $\rho_\alpha < 1/2$. In the scalar longitudinal counting model: (i) if $Bp_j \ll 1$, the mean margin drift is $\eta Bp_j(1 - 2\rho_\alpha) + O(\eta(Bp_j)^2)$; (ii) if $1 \ll Bp_j \ll \rho_\alpha^{-1}$, the large-margin normalized drift is close to the unit deterministic-label response; (iii) if $Bp_j\rho_\alpha \gg 1$, that large-margin drift is negative. Separately, near the noisy population optimum, an isotropic covariance closure with a locally homogeneous frequency approximation predicts a Muon contraction coefficient proportional to $\sqrt{Bp}$, compared with p for SGD. These scalar and closure predictions are not exact formulas for the complete categorical matrix dynamics.*

The counting model separates occurrence-limited, corruption-sparse, and corruption-resolved behavior. A broad corruption-sparse window requires small ρ_α. Resolving corruptions reverses the projected large-margin drift, motivating a separate local analysis around the finite population optimum. The local closure predicts weaker frequency dependence for Muon than for SGD; Figure 3 tests this prediction in the complete categorical model, beyond the closure assumptions.

Appendix E derives the complete categorical covariance and local Gaussian limit, then the projected counting dynamics and covariance closure used for the explicit regime formulas.

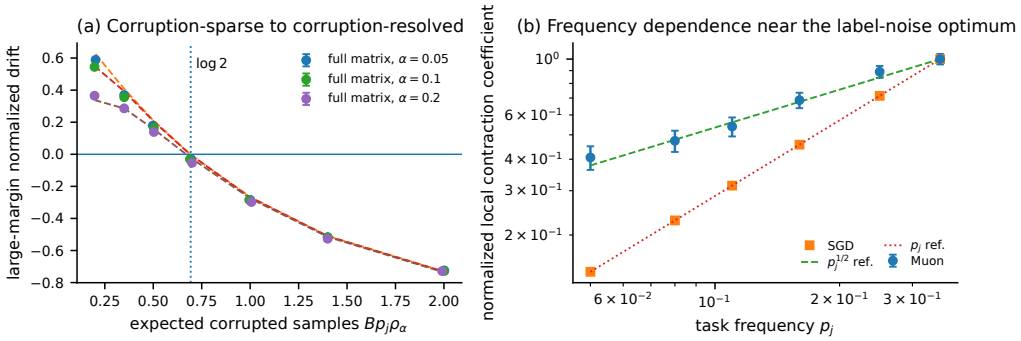


Figure 3: **Corrupted observations introduce a second sampling scale.** Left: full-matrix Muon changes from positive to negative large-margin drift as the expected number of corrupted task-j observations, $Bp_j\rho_\alpha$, crosses an order-one scale; dashed curves show the exact projected counting formula, whose crossover approaches $\log 2$ in the Poisson limit. Right: near the population optimum under label noise, the full-categorical Muon response scales empirically as $p_j^{0.494}$, close to the square-root prediction under the auxiliary covariance and frequency approximations, whereas the SGD coefficient scales linearly in p_j.

4.5 Sparse support: Muon versus SignSGD

Without momentum, sparse mini-batch sampling also distinguishes polar normalization from coordinatewise sign normalization. Muon is orthogonally equivariant and preserves the missing-task nullspace; SignSGD is coordinate dependent, so a basis rotation can turn an update for an observed task into an update of an unobserved task.

Proposition 4.5 (Muon avoids rotation-induced updates to a missing task)**.** *Consider momentum-free updates in the two-task model with query embedding* $E = \left[\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right]$, $0 < \theta < \pi/2$, *and let* $\widetilde{c} = \|\widetilde{E}(1,-1)^\top\|_1$. *Then a batch with no task-2 samples gives* $\Delta_2^{\mathrm{Muon}}(t+1) - \Delta_2^{\mathrm{Muon}}(t) = 0$, *whereas SignSGD gives* $\Delta_2^{\mathrm{Sign}}(t+1) - \Delta_2^{\mathrm{Sign}}(t) = \eta\widetilde{c}(\cos\theta - \sin\theta)$. *Hence* $\theta > \pi/4$ *makes every frequent-only batch decrease the rare-task margin, and* $p_1^B > (1+\cot\theta)/2$ *is sufficient for linear decay of the rare-task margin almost surely.*

The contrast is especially clear at $B = 1$. Muon updates only the sampled task: the rare margin still grows, but on its slow occurrence clock. After rotation, the same stream of frequent-task observations produces a negative component in the rare SignSGD margin. Figure 4 shows the long-time consequence for $p_2 = 0.05$ and $\theta = 1.15$ radians.

Appendix F gives the exact rotated recursion and the long-time negative-transfer proof.

5 Momentum: Averaging and Tracking

Momentum changes the sampling regime seen by the polar map by averaging stochastic gradients across optimization steps. At a frozen state in the deterministic-label model, the stationary EMA $M_t = (1-\beta)\sum_{r\geq 0}\beta^r G_{t-r}$ can be expressed using exponentially averaged task frequencies, reducing their sampling variance without changing the physical mini-batch size.

Proposition 5.1 (Averaging scale and tracking under momentum)**.** *For mini-batches of size* B, *the stationary EMA frequency of task* j *has variance-equivalent batch size* $B_{\mathrm{eff}}(\beta) = B(1+\beta)/(1-\beta)$; *relative frequency fluctuations vanish as* $B_{\mathrm{eff}}p_j \to \infty$. *If* $\|S_t - S_{t-r}\|_F \leq L_S\eta r$, *the frozen-state buffer error is bounded by* $L_S\eta\beta/(1-\beta)$. *Thus* $\eta/(1-\beta) \to 0$ *is sufficient for this error to vanish.*

At a frozen state, replacing B by B_{eff} reproduces task-frequency variances, but not task-absence probabilities or the full batch distribution. For β close to one, $B_{\mathrm{eff}} \approx 2B/(1-\beta)$, so $1-\beta \ll Bp_j$

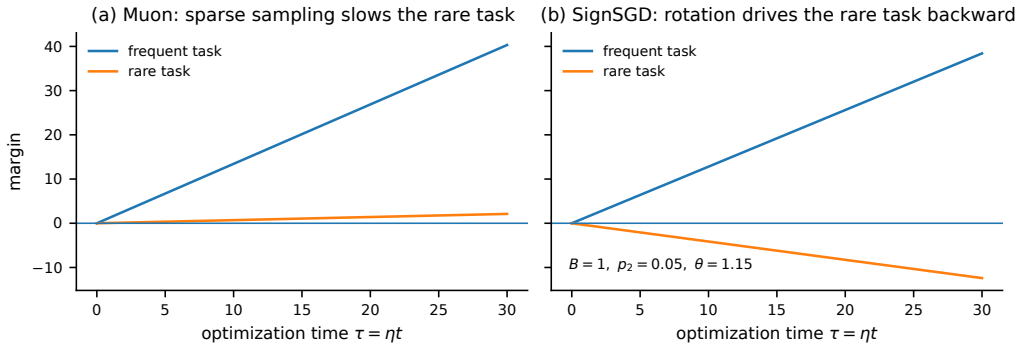


Figure 4: **Sparse support yields slow positive tail learning for Muon but negative transfer for SignSGD.** Both panels use the same $B = 1$, $p_1 = 0.95$, $p_2 = 0.05$ mini-batch streams over 64 runs. Left: Muon preserves support, so the rare-task margin remains nonnegative and grows on its occurrence clock. Right: at $\theta = 1.15 > \pi/4$, rotated SignSGD drives the rare-task margin downward while the frequent task continues to improve. The positive answer-embedding scalar $\widetilde{c}$ is absorbed into the SignSGD learning rate. Curves show medians and ribbons show interquartile ranges.

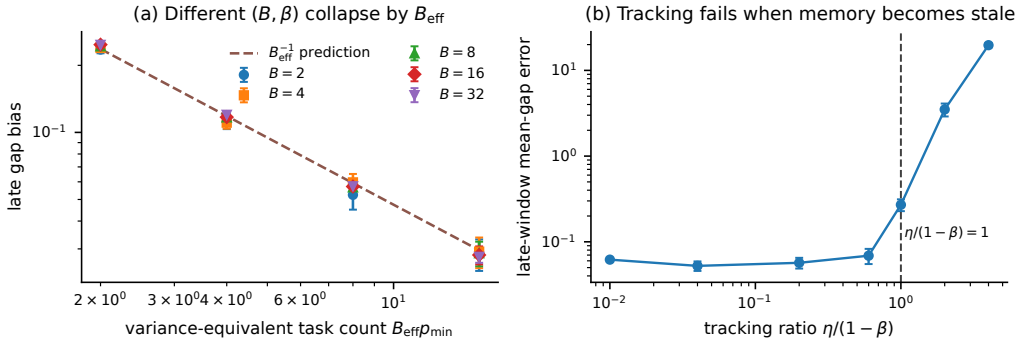


Figure 5: **Momentum trades variance reduction against tracking.** Left: different physical batch sizes and momentum coefficients nearly collapse when the two-task gap bias is plotted against the variance-equivalent task count $B_{\text{eff}} p_{\min}$. Right: with B and β fixed and a common horizon $\tau_{\max} = 50$, the late-window mean-gap error stays small while $\eta/(1-\beta)$ is small and grows once the EMA horizon becomes comparable to the state-evolution time scale.

suppresses relative frequency fluctuations. EMA updates are also temporally correlated: reduced instantaneous variance need not reduce long-time diffusion (Appendix C.5).

Under the local tracking assumptions, combining averaging with a small frozen-state buffer error gives the sufficient window

$$\eta \ll 1 - \beta \ll B p_j.$$

Transferring a buffer-error bound to a polar-direction bound additionally requires local control of the nonzero singular values or of relative perturbations. The window is therefore a local design guide, not a global convergence guarantee. It motivates longer memory for rare, slowly varying directions and shorter memory where rapid tracking is needed. Under stochastic labels, averaging also reduces persistent label fluctuations.

Appendix D.6 derives the EMA variance and tracking decomposition, and Appendix G gives a complementary scalar bias–variance calculation.

6 Fast–Slow Momentum in Language-Model Pretraining

From theory to practice. Our analysis motivates longer memory for rare, slowly varying directions and shorter memory where rapid tracking is needed. To examine this principle in language-model pretraining, we use the curvature-conditioned fast–slow momentum construction of Zhu et al. (2026). It retains the fast Muon update and adds a slow-momentum contribution in estimated flat directions, identified through the lower singular subspaces of the fast momentum. This provides a practical counterpart to the averaging–tracking trade-off in our theory, without requiring explicit task-frequency estimates.

Fast–slow Muon (schematic)

Decay rates $0 \le \beta_{\mathrm{f}} < \beta_{\mathrm{s}} < 1$; mixing weight $\lambda > 0$; step size η_t. Initialize $M^{\mathrm{f}} = M^{\mathrm{s}} = 0$.

for each step t **do**
  Compute gradient G_t
  $M^{\mathrm{f}} \leftarrow \beta_{\mathrm{f}} M^{\mathrm{f}} + G_t$
  $M^{\mathrm{s}} \leftarrow \beta_{\mathrm{s}} M^{\mathrm{s}} + G_t$
  $H_t \leftarrow \beta_{\mathrm{f}} M^{\mathrm{f}} + G_t$
  Estimate flat projector $\mathcal{Q}_t$ from H_t
  $D_t^{\mathrm{f}} \leftarrow \mathrm{msgn}(H_t)$
  $D_t^{\mathrm{s}} \leftarrow \mathcal{Q}_t(\mathrm{RowScale}(M^{\mathrm{s}}))$
  $W_{t+1} \leftarrow W_t - \eta_t(D_t^{\mathrm{f}} + \lambda D_t^{\mathrm{s}})$
end for

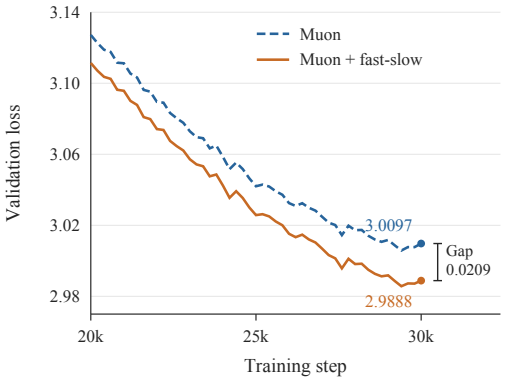


Figure 6: **Fast–slow momentum: core update and pretraining results.** Left: schematic momentum mixing following Zhu et al. (2026); RowScale normalizes rows, and parameter and update scaling follow the cited implementation. Right: validation loss from steps 20k to 30k for a 0.12B dense language model, with a final reduction of 0.0209. Other training settings are fixed; curves are unsmoothed, with one run per method.

Pretraining results. We compare standard Muon with this fast–slow variant on a 0.12B-parameter dense language model trained for 30,000 steps, keeping the remaining training configuration fixed. Across all 138 shared checkpoints from step 2,600 to 30,000, the fast–slow variant has lower validation loss. Figure 6 shows the final 10,000 steps. At step 30,000, validation loss decreases from 3.0097 to 2.9888, a reduction of 0.0209. These single-run results illustrate the practical relevance of heterogeneous momentum; variability across seeds and the effects of individual components remain to be evaluated. Appendix H.8 documents the trajectory comparison.

7 Conclusion

We study when Muon's preconditioning benefits survive stochastic gradient noise. Under isotropic low-SNR noise, its first-order conditional mean update matches learning-rate-rescaled SGD, showing that samplewise spectral normalization alone does not guarantee directional preconditioning. Structured sampling yields richer behavior: we establish an exact two-task balancing threshold and analyze higher-rank sampling effects governed by Bp_j. Label noise further separates task coverage from gradient reliability; an isotropic covariance closure with a locally homogeneous frequency approximation predicts square-root frequency dependence near the noisy optimum. A separate momentum-free SignSGD comparison highlights Muon's protection against rotation-induced negative transfer under sparse sampling.

Our momentum analysis connects frozen-state variance reduction with local tracking error, motivating longer memory for rare, slowly varying directions. This provides a sampling-based interpretation of heterogeneous momentum designs, supported by a preliminary 0.12B pretraining experiment in which fast–slow momentum lowers validation loss relative to Muon. Extending the analysis to evolving, non-orthogonal representations and validating these benefits across training scales remain important directions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with technical aspects of some mathematical proofs and derivations, and to assist with translation of the manuscript. We did not use generative AI tools to formulate the research questions, conceptual framework, or theorem statements. Additionally, we used generative AI tools to generate and edit code for controlled synthetic numerical experiments and figure generation. We reviewed all AI-assisted work. All AI-assisted mathematical arguments were checked and verified by the authors; AI-generated numerical code was inspected and tested for correctness, and all translated or edited text and generated figures were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code-derived results, and other artifacts produced with the aid of generative AI.

REFERENCES

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. arXiv:2609.19969, 2026. URL https://arxiv.org/abs/2609.19969.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying Shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. arXiv:2602.09314, 2026. URL https://arxiv.org/abs/2602.09314.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. arXiv:1309.5549, 2013. URL https://arxiv.org/abs/1309.5549.

Euijin Hong and Guannan Qu. Activation-Keyed Momentum: An anisotropic momentum update via the delta rule. arXiv:2608.19491, 2026. URL https://arxiv.org/abs/2608.19491.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Online technical note, 2024. URL https://kellerjordan.github.io/posts/muon/.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4c0129f1f7a036af5ed2409f1f9f8131-Abstract-Conference.html. arXiv:2601.19156.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. arXiv:2603.26554, 2026. URL https://arxiv.org/abs/2603.26554.

Kimi Team et al. Kimi K3: Open frontier intelligence. arXiv:2607.24653, 2026. URL https://arxiv.org/abs/2607.24653.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2602.05725. arXiv:2602.05725.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in Muon via spectral filtering. arXiv:2606.03899, 2026b. URL https://arxiv.org/abs/2606.03899.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. arXiv:2502.16982, 2025. URL https://arxiv.org/abs/2502.16982.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. arXiv:2601.13474, 2026. URL https://arxiv.org/abs/2601.13474.

Gagik Magakyan, Pablo A. Parrilo, and Asuman Ozdaglar. Spectral scaling laws of Muon. arXiv:2606.04058, 2026. URL https://arxiv.org/abs/2606.04058.

Hiroki Naganuma, Shagun Gupta, Youssef Briki, Ioannis Mitliagkas, Irina Rish, Parameswaran Raman, and Hao-Jun Michael Shi. Adaptive batch sizes using non-euclidean gradient noise scales for stochastic sign and spectral descent. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.03001. arXiv:2602.03001.

Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon eclipses SignSGD. arXiv:2605.09552, 2026. URL https://arxiv.org/abs/2605.09552.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YzjS4jcfmS. arXiv:2510.22980.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3eec5006051d9544e717067de3220198-Abstract-Conference.html. arXiv:2509.26030.

Shuchen Zhu, Yuxin Fang, Mingze Wang, and Kun Yuan. Curvature-conditioned multiscale momentum with sphere constraints for LLM pretraining. arXiv:2608.28442, 2026. URL https://arxiv.org/abs/2608.28442.

Guide to the appendices. Appendix A proves the isotropic Gaussian response result. Appendix B records the associative-memory setup used throughout the structured analysis. Appendix C gives the exact two-task dynamics and their large-batch/momentum refinements. Appendix D supports the higher-rank regime map through sparse-task gating, full-batch contrast dynamics, well-sampled perturbation analysis, momentum averaging, and a final technical continuum refinement. Appendix E separates exact categorical label-noise results from the projected counting model and the additional covariance closure. Appendix F proves the Muon–SignSGD sparse-support contrast. Appendix G contains an optional scalar EMA bias–variance calculation, and Appendix H collects numerical protocols and supplementary diagnostics.

A Isotropic Gaussian Mean Response: Proof of Proposition 3.1

Throughout this appendix, we condition on the current iterate and suppress the time index. We write

$$G = \bar{G} + \sigma\Xi, \qquad \bar{G} = \nabla\mathcal{L}(W),$$

where $\Xi \in \mathbb{R}^{d\times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries and $\sigma > 0$.

A.1 Gaussian-smoothing proof

Define

$$F(H) := \mathbb{E}[\mathrm{msgn}(\Xi + H)], \qquad H \in \mathbb{R}^{d\times d},$$

and write

$$\langle A, B\rangle_F := \mathrm{tr}(A^\top B).$$

Smoothness by Gaussian convolution. Let

$$\phi(A) = (2\pi)^{-d^2/2} \exp\left(-\frac{1}{2}\|A\|_F^2\right)$$

be the standard Gaussian density on $\mathbb{R}^{d\times d}$. Changing variables yields

$$\begin{aligned} F(H) &= \int \operatorname{msgn}(A)\phi(A-H)\,\mathrm{d}A \\ &= \mathbb{E}\left[\operatorname{msgn}(\Xi)\exp\left(\langle H,\Xi\rangle_F - \frac{1}{2}\|H\|_F^2\right)\right]. \end{aligned}$$

For $\|H\|_F \le r$, derivatives of any fixed order of the exponential factor are bounded by a constant, depending only on the derivative order and r, times a polynomial in $\|\Xi\|_F$ multiplied by $e^{r\|\Xi\|_F}$. Since $\|\operatorname{msgn}(\Xi)\|_F \le \sqrt{d}$ and

$$\mathbb{E}\left[(1+\|\Xi\|_F^k)e^{r\|\Xi\|_F}\right] < \infty$$

for every fixed $k \ge 0$ and $r > 0$, differentiation under the expectation is justified locally to every order. The same domination also implies continuity of the resulting derivatives. Hence F is smooth, and

$$DF(0)[X] = \mathbb{E}[\operatorname{msgn}(\Xi)\langle X,\Xi\rangle_F] \tag{2}$$

for every $X \in \mathbb{R}^{d\times d}$. This argument differentiates the shifted Gaussian density, not the nonsmooth map msgn.

Orthogonal equivariance of the derivative. We use the following elementary lemma.

Lemma A.1. *Let $T : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ be linear and satisfy*

$$T(UXV^\top) = UT(X)V^\top$$

for every $X \in \mathbb{R}^{d\times d}$ and every $U, V \in O(d)$. Then $T(X) = cX$ for some scalar $c \in \mathbb{R}$ and all X.

Proof. Fix i, j, and let D_1, D_2 be arbitrary diagonal sign matrices satisfying $(D_1)_{ii} = (D_2)_{jj} = 1$. Then

$$D_1 E_{ij} D_2^\top = E_{ij}.$$

Hence equivariance gives

$$T(E_{ij}) = D_1 T(E_{ij}) D_2^\top.$$

Varying all remaining diagonal signs independently forces every entry of $T(E_{ij})$ except entry (i, j) to vanish. Thus $T(E_{ij}) = c_{ij}E_{ij}$. Independent row and column permutations map any E_{ij} to any other matrix unit, so equivariance forces all c_{ij} to equal a common scalar c. Linearity completes the proof. □

The identities

$$\operatorname{msgn}(UAV^\top) = U\operatorname{msgn}(A)V^\top, \qquad U^\top \Xi V \overset{\mathrm{d}}{=} \Xi$$

hold for all $U, V \in O(d)$. Therefore

$$\begin{aligned} F(UHV^\top) &= U\,\mathbb{E}\left[\operatorname{msgn}(U^\top \Xi V + H)\right] V^\top \\ &= UF(H)V^\top. \end{aligned}$$

Differentiating at zero gives

$$DF(0)[UXV^\top] = UDF(0)[X]V^\top.$$

By Lemma A.1, there is a scalar c_d such that

$$DF(0)[X] = c_d X.$$

Identification of the coefficient. Taking the trace of this linear operator in the orthonormal basis $\{E_{ij}\}_{i,j=1}^{d}$ and using equation 2, we obtain

$$\begin{aligned} d^2 c_d &= \sum_{i,j} \langle E_{ij}, DF(0)[E_{ij}] \rangle_F \\ &= \sum_{i,j} \mathbb{E}[\mathrm{msgn}(\Xi)_{ij} \Xi_{ij}] \\ &= \mathbb{E}\langle \mathrm{msgn}(\Xi), \Xi \rangle_F. \end{aligned}$$

For an SVD $\Xi = U\Sigma V^\top$,

$$\langle \mathrm{msgn}(\Xi), \Xi \rangle_F = \mathrm{tr}(\Sigma) = \|\Xi\|_*.$$

Consequently,

$$c_d = \frac{1}{d^2} \mathbb{E}\|\Xi\|_* > 0.$$

The expectation is finite because $\|\Xi\|_* \le \sqrt{d}\,\|\Xi\|_F$, and it is strictly positive because Ξ is nonzero almost surely. For $d = 1$, $c_1 = \mathbb{E}|\xi| = \sqrt{2/\pi}$, recovering the SignSGD coefficient.

Taylor theorem with remainder. Since $\mathrm{msgn}(-A) = -\mathrm{msgn}(A)$ and $\Xi \overset{\mathrm{d}}{=} -\Xi$, we have $F(-H) = -F(H)$. Consequently,

$$F(0) = 0, \qquad D^2F(0) = 0.$$

Because F is smooth and the dimension is fixed,

$$M_d := \sup_{\|Z\|_F \le 1} \|D^3 F(Z)\|_{\mathrm{op}} < \infty,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the norm of a trilinear map induced by the Frobenius norm on its arguments and output. Taylor's theorem with integral remainder gives

$$F(H) - DF(0)[H] = \frac{1}{2} \int_0^1 (1-s)^2 D^3 F(sH)[H, H, H]\, \mathrm{d}s.$$

Hence, for every $\|H\|_F \le 1$,

$$\|F(H) - c_d H\|_F \le \frac{M_d}{6} \|H\|_F^3.$$

Finally, since $\sigma > 0$ and the matrix sign is invariant under positive rescaling,

$$\mathrm{msgn}(\bar{G} + \sigma\Xi) = \mathrm{msgn}\left(\Xi + \frac{\bar{G}}{\sigma}\right).$$

Taking $H = \bar{G}/\sigma$, for $\|\bar{G}\|_F/\sigma \le 1$ we obtain

$$\left\| \mathbb{E}[\mathrm{msgn}(\bar{G} + \sigma\Xi)] - \frac{c_d}{\sigma} \bar{G} \right\|_F \le \frac{M_d}{6} \frac{\|\bar{G}\|_F^3}{\sigma^3}.$$

This proves the claimed $O_d(\|\bar{G}\|_F^3/\sigma^3)$ remainder as $\|\bar{G}\|_F/\sigma \to 0$.

B Associative-Memory Setup and Gradient Derivation

B.1 Gradient of the Associative Memory Model

Recall that

$$\overline{W} = \widetilde{E}^\top W E, \qquad S_{ij} = \frac{\exp(\overline{W}_{ij})}{\sum_{k=1}^{K} \exp(\overline{W}_{kj})}.$$

For task j, the cross-entropy loss is

$$\ell_j(W) = -\log S_{jj} = -\widetilde{E}_j^\top W E_j + \log\left(\sum_{k=1}^{K} e^{\widetilde{E}_k^\top W E_j} \right).$$

Differentiating gives

$$\nabla_W \ell_j(W) = -\widetilde{E}_j E_j^\top + \sum_{k=1}^{K} S_{kj} \widetilde{E}_k E_j^\top$$
$$= -\left(\widetilde{E}_j - \sum_{k=1}^{K} S_{kj} \widetilde{E}_k \right) E_j^\top .$$

Hence, with $P = \mathrm{diag}(p_1, \ldots, p_K)$,

$$\nabla \mathcal{L}(W) = -\sum_{j=1}^{K} \sum_{k=1}^{K} p_j (\delta_{kj} - S_{kj}) \widetilde{E}_k E_j^\top$$
$$= -\widetilde{E}(I - S) P E^\top .$$

Similarly, for a mini-batch with empirical frequencies $q(t) = (q_1(t), \ldots, q_K(t))$ and $Q_t = \mathrm{diag}(q(t))$,

$$\nabla \mathcal{L}(W_t; \mathcal{B}_t) = -\widetilde{E}(I - S_t) Q_t E^\top .$$

Since the partial polar factor is equivariant under left and right orthogonal transformations,

$$\mathrm{msgn}\left(\widetilde{E} A E^\top\right) = \widetilde{E}\, \mathrm{msgn}(A) E^\top ,$$

the task-coordinate Muon update is

$$\overline{W}_{t+1} = \overline{W}_t + \eta\, \mathrm{msgn}((I - S_t) Q_t) .$$

C Two-Task Finite-Batch Dynamics and Refinements

The two-task model is retained as an exactly solvable specialization of the finite-batch sampling mechanism. Because its stochastic gradient is rank one, it should not be interpreted as evidence for genuinely higher-rank spectral effects. Throughout this appendix, we assume $K = 2$ and $p_1 \geq p_2 = p_{\min} > 0$. Each batch consists of B independent samples from (p_1, p_2), independently across iterations. Subsections C.1–C.4 establish the exact finite-B statements used in Theorem 4.1; the final subsection records large-batch and momentum-averaging refinements rather than a second phase-transition claim. Thus

$$B q_1(t) \sim \mathrm{Binom}(B, p_1), \qquad q_2(t) = 1 - q_1(t).$$

C.1 Exact Reduction to the Margin Dynamics

Define

$$\Delta_1(t) = (\overline{W}_t)_{11} - (\overline{W}_t)_{21}, \qquad \Delta_2(t) = (\overline{W}_t)_{22} - (\overline{W}_t)_{12},$$

and

$$u_i(t) = \frac{1}{1 + e^{\Delta_i(t)}}.$$

For two classes,

$$1 - S_{11}(t) = S_{21}(t) = u_1(t), \qquad 1 - S_{22}(t) = S_{12}(t) = u_2(t).$$

Therefore

$$(I - S_t) Q_t = \begin{bmatrix} q_1(t)(1 - S_{11}(t)) & -q_2(t) S_{12}(t) \\ -q_1(t) S_{21}(t) & q_2(t)(1 - S_{22}(t)) \end{bmatrix}$$
$$= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} q_1(t) u_1(t) & -q_2(t) u_2(t) \end{bmatrix} .$$

This matrix has rank one whenever at least one $q_i(t)$ is nonzero, which always holds. For a nonzero rank-one matrix $ab^\top$,

$$\mathrm{msgn}(ab^\top) = \frac{a}{\|a\|_2} \frac{b^\top}{\|b\|_2}.$$

Since $\|(1,-1)^\top\|_2 = \sqrt{2}$, we obtain

$$\mathrm{msgn}((I - S_t)Q_t) = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \frac{[q_1(t)u_1(t) \quad -q_2(t)u_2(t)]}{\sqrt{q_1^2(t)u_1^2(t) + q_2^2(t)u_2^2(t)}}.$$

Taking the corresponding column differences gives

$$\Delta_i(t+1) - \Delta_i(t) = \sqrt{2}\eta \frac{q_i(t)u_i(t)}{\sqrt{q_1^2(t)u_1^2(t) + q_2^2(t)u_2^2(t)}}. \tag{3}$$

C.2 Full-batch Dynamics

Set $q_i(t) = p_i$ and take the continuous-time limit $\tau = \eta t$. Then

$$\dot{\Delta}_i = \sqrt{2}\, \frac{p_i u_i}{\sqrt{p_1^2 u_1^2 + p_2^2 u_2^2}}.$$

Dividing the two equations gives

$$\frac{d\Delta_1}{d\Delta_2} = \frac{p_1 u_1}{p_2 u_2} = \frac{p_1(1 + e^{\Delta_2})}{p_2(1 + e^{\Delta_1})}.$$

Hence

$$p_2(1 + e^{\Delta_1})\, d\Delta_1 = p_1(1 + e^{\Delta_2})\, d\Delta_2.$$

Using $\Delta_1(0) = \Delta_2(0) = 0$,

$$p_2(\Delta_1 + e^{\Delta_1} - 1) = p_1(\Delta_2 + e^{\Delta_2} - 1). \tag{4}$$

Moreover,

$$\sqrt{2} \le \frac{d}{d\tau}(\Delta_1 + \Delta_2) = \sqrt{2}\, \frac{p_1 u_1 + p_2 u_2}{\sqrt{p_1^2 u_1^2 + p_2^2 u_2^2}} \le 2.$$

Thus $\Delta_1 + \Delta_2 \to \infty$. Therefore, equation 4 implies that both margins diverge, and the exponential terms dominate:

$$p_2 e^{\Delta_1} \sim p_1 e^{\Delta_2}.$$

Therefore

$$\Delta_1 - \Delta_2 \longrightarrow \log \frac{p_1}{p_2}, \qquad p_1 u_1 \sim p_2 u_2.$$

Substituting back into the ODE yields

$$\dot{\Delta}_i \longrightarrow 1, \qquad \frac{\Delta_i(\tau)}{\tau} \longrightarrow 1.$$

C.3 Small Mini-batches: Occurrence-limited Imbalance

Let

$$A_t = \{q_1(t) = 1\}, \qquad \mathbb{P}(A_t) = p_1^B.$$

On A_t,

$$\Delta_1(t+1) - \Delta_1(t) = \sqrt{2}\eta, \qquad \Delta_2(t+1) - \Delta_2(t) = 0.$$

On A_t^c, both increments are nonnegative and their difference is at least $-\sqrt{2}\eta$. Hence

$$D(t+1) - D(t) \ge \sqrt{2}\eta\big(2\mathbf{1}_{A_t} - 1\big), \qquad D(t) := \Delta_1(t) - \Delta_2(t).$$

By the strong law of large numbers,

$$\liminf_{t\to\infty} \frac{D(t)}{\eta t} \ge \sqrt{2}(2p_1^B - 1) \qquad \text{a.s.}$$

Thus $D(t) \to +\infty$ when $p_1^B > 1/2$.

We can then identify the exact asymptotic rates. Since

$$\frac{u_1(t)}{u_2(t)} = \frac{1 + e^{\Delta_2(t)}}{1 + e^{\Delta_1(t)}} \longrightarrow 0,$$

every batch containing task-2 asymptotically allocates the normalized update to task-2, while a pure task-1 batch allocates it to task-1. Consequently,

$$\frac{\Delta_1(t)}{\eta t} \xrightarrow{\text{a.s.}} \sqrt{2}p_1^B, \qquad \frac{\Delta_2(t)}{\eta t} \xrightarrow{\text{a.s.}} \sqrt{2}(1-p_1^B).$$

The transition $p_1^B = 1/2$ is equivalent to

$$B_{\text{bal}} = \frac{\log 2}{-\log p_1} = \frac{\log 2}{-\log(1-p_{\min})}.$$

C.4 Large Mini-batches: Recovery of Balanced Growth

Assume

$$p_1^B < \frac{1}{2}.$$

We first establish an exact stability result for the discrete-time gap

$$D(t) := \Delta_1(t) - \Delta_2(t).$$

Restoring drift in the tails. From equation 3,

$$D(t+1) - D(t) = \sqrt{2}\eta \frac{q_1(t)u_1(t) - q_2(t)u_2(t)}{\sqrt{q_1^2(t)u_1^2(t) + q_2^2(t)u_2^2(t)}}. \tag{5}$$

In particular,

$$|D(t+1) - D(t)| \le \sqrt{2}\eta.$$

Since both margins start from zero and have nonnegative increments,

$$\Delta_i(t) \ge 0, \qquad \forall t.$$

If $D(t) \ge 0$, then

$$\begin{aligned}\frac{u_1(t)}{u_2(t)} &= \frac{1+e^{\Delta_2(t)}}{1+e^{\Delta_1(t)}} \\ &= \frac{1+e^{\Delta_2(t)}}{1+e^{\Delta_2(t)+D(t)}} \\ &\le 2e^{-D(t)}.\end{aligned}$$

Thus $u_1/u_2 \to 0$ uniformly over the common offset of the two margins as $D \to +\infty$.

If $q_2(t) = 0$, then $q_1(t) = 1$ and $D(t+1) - D(t) = \sqrt{2}\eta$.

If $q_2(t) > 0$, then $q_2(t) \ge 1/B$. Since the batch proportions take values in a finite set, equation 5 and the bound above imply, uniformly over all batches with $q_2(t) > 0$,

$$D(t+1) - D(t) \longrightarrow -\sqrt{2}\eta \qquad \text{as } D(t) \to +\infty.$$

Consequently,

$$\mathbb{E}[D(t+1) - D(t) \mid \mathcal{F}_t] \longrightarrow \sqrt{2}\eta\left(2p_1^B - 1\right) < 0 \qquad \text{as } D(t) \to +\infty. \tag{6}$$

Similarly, if $D(t) \le 0$, then

$$\frac{u_2(t)}{u_1(t)} \le 2e^{D(t)} = 2e^{-|D(t)|}.$$

Hence, as $D(t) \to -\infty$, every batch containing task-1 satisfies

$$D(t+1) - D(t) \longrightarrow \sqrt{2}\eta,$$

whereas a pure task-2 batch satisfies

$$D(t+1) - D(t) = -\sqrt{2}\eta.$$

Therefore,

$$\mathbb{E}[D(t+1) - D(t) \mid \mathcal{F}_t] \longrightarrow \sqrt{2}\eta\left(1 - 2p_2^B\right) > 0 \qquad \text{as } D(t) \to -\infty, \tag{7}$$

where we used $p_2^B \le p_1^B < 1/2$.

It follows that there exist constants $R, \delta > 0$ such that

$$\mathbb{E}[D(t+1) - D(t) \mid \mathcal{F}_t] \le -\delta \qquad \text{when } D(t) \ge R,$$

and

$$\mathbb{E}[D(t+1) - D(t) \mid \mathcal{F}_t] \ge \delta \qquad \text{when } D(t) \le -R.$$

Exponential moment bound. Let

$$X_t := D(t+1) - D(t), \qquad L := \sqrt{2}\eta,$$

so that $|X_t| \le L$. Increasing R if necessary, assume $R > L$.

When $D(t) \ge R$, the sign of $D(t)$ cannot change in one step, and hence

$$|D(t+1)| = D(t) + X_t.$$

For any $\theta > 0$,

$$e^{\theta X_t} \le 1 + \theta X_t + \frac{\theta^2 L^2}{2} e^{\theta L}.$$

Therefore, for sufficiently small $\theta > 0$,

$$\mathbb{E}\left[e^{\theta X_t} \mid \mathcal{F}_t\right] \le 1 - \frac{\theta\delta}{2} \qquad \text{when } D(t) \ge R.$$

The same argument applied to $-X_t$ gives

$$\mathbb{E}\left[e^{-\theta X_t} \mid \mathcal{F}_t\right] \le 1 - \frac{\theta\delta}{2} \qquad \text{when } D(t) \le -R.$$

Define

$$V(D) := e^{\theta|D|}.$$

There then exist constants $\rho < 1$ and $b_0 < \infty$ such that

$$\mathbb{E}[V(D(t+1)) \mid \mathcal{F}_t] \le \rho V(D(t)) + b_0 \mathbf{1}_{\{|D(t)| \le R\}}.$$

Iterating this inequality yields

$$\sup_{t \ge 0} \mathbb{E} e^{\theta|D(t)|} < \infty. \tag{8}$$

Hence, for every $\varepsilon > 0$,

$$\mathbb{P}(|D(t)| > \varepsilon t) \le b_\theta e^{-\theta\varepsilon t},$$

and therefore

$$\sum_{t=1}^{\infty} \mathbb{P}(|D(t)| > \varepsilon t) < \infty.$$

By the Borel–Cantelli lemma,

$$\frac{D(t)}{t} \longrightarrow 0 \qquad \text{a.s.} \tag{9}$$

This proves vanishing asymptotic gap growth whenever $p_1^B < 1/2$.

Late-time effective drift. We next refine the description of the gap in the well-sampled regime. The remainder of this subsection concerns a late-time, small-step approximation rather than an exact invariant law of the discrete process.

First note that both margins diverge almost surely. A pure task-i batch occurs independently with probability $p_i^B > 0$, and on every such batch,

$$\Delta_i(t+1) - \Delta_i(t) = \sqrt{2}\eta.$$

The strong law therefore implies

$$\liminf_{t\to\infty} \frac{\Delta_i(t)}{\eta t} \geq \sqrt{2} p_i^B > 0 \qquad \text{a.s.}$$

Consequently,

$$u_i(t) = e^{-\Delta_i(t)} \left(1 + o(1)\right).$$

Define the centered gap

$$\gamma(t) := D(t) - \log \frac{p_1}{p_2}.$$

Then

$$\frac{p_1 u_1(t)}{p_2 u_2(t)} = e^{-\gamma(t)} \left(1 + o(1)\right).$$

For

$$a = \frac{q_1}{p_1}, \qquad b = \frac{q_2}{p_2},$$

define

$$\varphi(\gamma; a, b) := \sqrt{2} \frac{a e^{-\gamma} - b}{\sqrt{a^2 e^{-2\gamma} + b^2}}. \tag{10}$$

Then, uniformly for γ in compact sets,

$$\frac{\gamma(t+1) - \gamma(t)}{\eta} = \varphi\left(\gamma(t); \frac{q_1(t)}{p_1}, \frac{q_2(t)}{p_2}\right) + o(1) \qquad \text{as } t \to \infty. \tag{11}$$

Define the corresponding late-time mean drift

$$\psi_B(\gamma) := \mathbb{E}\left[\varphi\left(\gamma; \frac{q_1}{p_1}, \frac{q_2}{p_2}\right)\right]. \tag{12}$$

For $a, b > 0$,

$$\partial_\gamma \varphi(\gamma; a, b) = -\sqrt{2} \frac{ab e^{-\gamma} \left(a e^{-\gamma} + b\right)}{\left(a^2 e^{-2\gamma} + b^2\right)^{3/2}} < 0.$$

If $b = 0$, then $\varphi(\gamma; a, 0) = \sqrt{2}$, whereas if $a = 0$, $\varphi(\gamma; 0, b) = -\sqrt{2}$. Thus φ is nonincreasing in γ for every batch and strictly decreasing for every mixed batch. Since mixed batches occur with positive probability, ψ_B is strictly decreasing.

Moreover,

$$\lim_{\gamma\to+\infty} \psi_B(\gamma) = \sqrt{2}\left(2p_1^B - 1\right) < 0,$$

and

$$\lim_{\gamma\to-\infty} \psi_B(\gamma) = \sqrt{2}\left(1 - 2p_2^B\right) > 0.$$

Hence ψ_B has a unique zero, denoted by

$$\gamma_B^*.$$

Suppose now that

$$B p_{\min} \to \infty.$$

Since

$$\mathrm{Var}\left(\frac{q_i}{p_i}\right) = \frac{1 - p_i}{B p_i},$$

we have

$$\frac{q_i}{p_i} \xrightarrow{\mathbb{P}} 1.$$

Since $|\varphi| \le \sqrt{2}$, convergence in probability together with uniform boundedness implies

$$\psi_B(\gamma) \longrightarrow \psi_\infty(\gamma) := \sqrt{2}\frac{e^{-\gamma}-1}{\sqrt{e^{-2\gamma}+1}}$$

for every fixed γ.

The limiting drift is strictly decreasing and has the unique zero $\gamma = 0$. Fix any $\delta > 0$. Since

$$\psi_\infty(-\delta) > 0, \qquad \psi_\infty(\delta) < 0,$$

the same inequalities hold for ψ_B when $Bp_{\min}$ is sufficiently large. By monotonicity,

$$|\gamma_B^*| < \delta.$$

Therefore,

$$\gamma_B^* \longrightarrow 0 \qquad \text{as } Bp_{\min} \to \infty. \tag{13}$$

For fixed $p_1, p_2 > 0$, this convergence can be sharpened. Define

$$h(q) := \varphi\left(0; \frac{q}{p_1}, \frac{1-q}{p_2}\right).$$

A direct calculation gives

$$h(p_1) = 0, \qquad h'(p_1) = \frac{1}{p_1 p_2}, \qquad h''(p_1) = \frac{p_1 - p_2}{p_1^2 p_2^2}.$$

Since

$$\mathbb{E}[q_1 - p_1] = 0, \qquad \mathrm{Var}(q_1) = \frac{p_1 p_2}{B},$$

a second-order expansion gives

$$\psi_B(0) = \frac{p_1 - p_2}{2Bp_1p_2} + o\left(\frac{1}{B}\right).$$

Furthermore,

$$\psi_B'(0) = -1 + o(1).$$

Expanding the zero of ψ_B around zero therefore yields

$$\gamma_B^* = \frac{p_1 - p_2}{2Bp_1p_2} + o\left(\frac{1}{B}\right). \tag{14}$$

Thus mini-batch fluctuations induce a small positive shift of the effective gap when $p_1 > p_2$, and this shift vanishes in the large-batch limit.

Local diffusion approximation. Let

$$\lambda_B := -\psi_B'(\gamma_B^*) > 0$$

and

$$v_B := \mathrm{Var}\left[\varphi\left(\gamma_B^*; \frac{q_1}{p_1}, \frac{q_2}{p_2}\right)\right].$$

Writing

$$x(t) := \gamma(t) - \gamma_B^*,$$

the local recursion suggested by equation 11 is

$$x(t+1) - x(t) \simeq -\eta\lambda_B x(t) + \eta\xi_t,$$

where

$$\mathbb{E}[\xi_t] \simeq 0, \qquad \mathrm{Var}(\xi_t) \simeq v_B.$$

In the well-sampled regime,

$$\lambda_B = 1 + o(1).$$

Moreover, using

$$h'(p_1) = \frac{1}{p_1 p_2}$$

and

$$\mathrm{Var}(q_1) = \frac{p_1 p_2}{B},$$

the delta method gives

$$v_B = \frac{1 + o(1)}{B p_1 p_2}. \tag{15}$$

On the rescaled time

$$\tau = \eta t,$$

the corresponding local diffusion approximation is

$$dx \simeq -\lambda_B x \, d\tau + \sqrt{\eta v_B} \, d\mathcal{W}_\tau. \tag{16}$$

The associated Ornstein–Uhlenbeck stationary variance is

$$\frac{\eta v_B}{2\lambda_B} = \frac{\eta}{2B p_1 p_2} \left(1 + o(1)\right).$$

Thus, within the late-time local diffusion approximation,

$$D = \log \frac{p_1}{p_2} + \gamma_B^* + O_{\mathbb{P}}\left(\sqrt{\frac{\eta}{B p_1 p_2}}\right). \tag{17}$$

Together with equation 13, this recovers the full-batch gap in the well-sampled, small-step limit.

C.5 Momentum: Averaging Across Steps and Tracking

We now consider

$$g_t = (q_1(t)u_1(t), q_2(t)u_2(t))^\top, \qquad m_t = \beta m_{t-1} + (1-\beta) g_t,$$

with

$$0 < \beta < 1, \qquad \epsilon := 1 - \beta.$$

The normalized margin update is

$$\Delta(t+1) - \Delta(t) = \sqrt{2}\eta \frac{m_t}{\|m_t\|_2}.$$

We first freeze the margins and characterize how momentum averages task frequencies across iterations. We then restore the slow evolution of the margins.

Frozen margins and the EMA task frequency. Fix $u_1, u_2 > 0$. After initialization transients, the stationary momentum admits the representation

$$m_t = (1-\beta) \sum_{k=0}^{\infty} \beta^k g_{t-k}.$$

Define the exponentially averaged task-1 frequency

$$Z_t := (1-\beta) \sum_{k=0}^{\infty} \beta^k q_1(t-k). \tag{18}$$

Since $q_2(t) = 1 - q_1(t)$ and

$$(1-\beta) \sum_{k=0}^{\infty} \beta^k = 1,$$

we obtain the exact frozen-margin representation

$$m_t = \begin{pmatrix} u_1 Z_t \\ u_2(1 - Z_t) \end{pmatrix}. \tag{19}$$

The first two moments of Z_t are

$$\mathbb{E}[Z_t] = p_1, \tag{20}$$

$$\begin{aligned} \mathrm{Var}(Z_t) &= (1-\beta)^2 \sum_{k=0}^{\infty} \beta^{2k} \frac{p_1 p_2}{B} \\ &= \frac{1-\beta}{1+\beta} \frac{p_1 p_2}{B}. \end{aligned} \tag{21}$$

It is therefore natural to define the variance-equivalent batch size

$$B_{\text{eff}} := B \frac{1+\beta}{1-\beta}, \tag{22}$$

for which

$$\mathrm{Var}(Z_t) = \frac{p_1 p_2}{B_{\text{eff}}}.$$

Since both tasks have positive probability and $\beta > 0$,

$$0 < Z_t < 1 \qquad \text{a.s.}$$

under the stationary representation. Thus EMA averaging assigns positive weight to both tasks almost surely even when individual tasks rarely appear in a mini-batch.

Momentum-induced effective gap. Define again

$$\gamma = D - \log \frac{p_1}{p_2}.$$

At late time,

$$\frac{p_1 u_1}{p_2 u_2} = e^{-\gamma} \left(1 + o(1)\right).$$

Using equation 19, the normalized frozen-buffer gap increment is therefore asymptotically

$$\eta \varphi\left(\gamma; \frac{Z_t}{p_1}, \frac{1 - Z_t}{p_2}\right),$$

where φ is defined in equation 10.

Define the frozen-momentum mean drift

$$\psi_{B,\beta}(\gamma) := \mathbb{E}\left[\varphi\left(\gamma; \frac{Z}{p_1}, \frac{1-Z}{p_2}\right)\right], \tag{23}$$

where Z has the stationary marginal distribution defined by equation 18.

Since $0 < Z < 1$ almost surely,

$$\partial_\gamma \varphi\left(\gamma; \frac{Z}{p_1}, \frac{1-Z}{p_2}\right) < 0 \qquad \text{a.s.}$$

Hence $\psi_{B,\beta}$ is strictly decreasing. Moreover,

$$\lim_{\gamma \to -\infty} \psi_{B,\beta}(\gamma) = \sqrt{2},$$

whereas

$$\lim_{\gamma \to +\infty} \psi_{B,\beta}(\gamma) = -\sqrt{2}.$$

Therefore there exists a unique $\gamma^*_{B,\beta}$ such that

$$\psi_{B,\beta}(\gamma^*_{B,\beta}) = 0.$$

Unlike the single-batch mean drift, the stationary frozen-momentum drift has no analogue of the threshold $p_1^B = 1/2$: for every $\beta > 0$, both tasks receive positive weight almost surely in the stationary EMA, producing restoring drift in both tails. This is the basic mechanism by which momentum suppresses task-frequency fluctuations across steps.

Recovery of the full-batch gap. From equation 21,

$$\mathrm{Var}\left(\frac{Z}{p_1}\right) = \frac{p_2}{B_{\mathrm{eff}} p_1},$$

and

$$\mathrm{Var}\left(\frac{1-Z}{p_2}\right) = \frac{p_1}{B_{\mathrm{eff}} p_2}.$$

Hence, if

$$B_{\mathrm{eff}} p_{\min} \longrightarrow \infty, \tag{24}$$

then

$$\frac{Z}{p_1} \xrightarrow{\mathbb{P}} 1, \qquad \frac{1-Z}{p_2} \xrightarrow{\mathbb{P}} 1.$$

Since $|\varphi| \le \sqrt{2}$,

$$\psi_{B,\beta}(\gamma) \longrightarrow \psi_\infty(\gamma) = \sqrt{2}\frac{e^{-\gamma}-1}{\sqrt{e^{-2\gamma}+1}}$$

for every fixed γ.

As before, monotonicity and uniqueness of the zero of ψ_∞ imply

$$\gamma^*_{B,\beta} \longrightarrow 0 \qquad \text{as } B_{\mathrm{eff}} p_{\min} \to \infty. \tag{25}$$

Since

$$B_{\mathrm{eff}} = B\frac{1+\beta}{1-\beta},$$

condition (24) is equivalent, up to the bounded factor $1+\beta$, to

$$\frac{1-\beta}{B p_{\min}} \longrightarrow 0.$$

For fixed $p_1, p_2 > 0$, the finite-sampling bias can be sharpened by a second-order expansion. Using the function h defined in the previous subsection,

$$\mathbb{E}[h(Z)] = \frac{1}{2}h''(p_1)\,\mathrm{Var}(Z) + o\big(B_{\mathrm{eff}}^{-1}\big)\,.$$

Therefore,

$$\psi_{B,\beta}(0) = \frac{p_1 - p_2}{2B_{\mathrm{eff}} p_1 p_2} + o\big(B_{\mathrm{eff}}^{-1}\big)\,.$$

Moreover,

$$\psi'_{B,\beta}(0) = -1 + o(1).$$

Expanding the unique zero of $\psi_{B,\beta}$ around zero gives

$$\gamma^*_{B,\beta} = \frac{p_1 - p_2}{2B_{\mathrm{eff}} p_1 p_2} + o\big(B_{\mathrm{eff}}^{-1}\big)\,. \tag{26}$$

Equivalently,

$$\gamma^*_{B,\beta} = \frac{1-\beta}{1+\beta}\frac{p_1 - p_2}{2B p_1 p_2} + o\big(B_{\mathrm{eff}}^{-1}\big)\,.$$

Thus momentum suppresses the finite-sampling displacement of the effective gap according to B_{eff}.

Tracking slowly evolving margins. We now return to the actual dynamics, in which $u_i(t)$ evolve over time. The EMA has a memory length of order

$$\epsilon^{-1} = (1-\beta)^{-1}$$

iterations. Over one iteration, the margins move by $O(\eta)$. In the late-time balanced regime, the relative variation of the mean gradient over one iteration is therefore also $O(\eta)$. Across one EMA window, this produces a relative tracking error of order

$$O\Big(\frac{\eta}{\epsilon}\Big)\,. \tag{27}$$

On the other hand, the frozen-buffer fluctuation scale follows from equation 21. Near balanced growth, where

$$p_1 u_1 \asymp p_2 u_2,$$

we obtain

$$\frac{\|m_t - \mathbb{E}[m_t]\|_2}{\|\mathbb{E}[m_t]\|_2} = O_{\mathbb{P}}\left(\sqrt{\frac{\epsilon}{Bp_{\min}}}\right). \tag{28}$$

Thus the two scale-separation requirements are

$$\frac{\eta}{\epsilon} \ll 1, \qquad \frac{\epsilon}{Bp_{\min}} \ll 1. \tag{29}$$

Equivalently,

$$\eta \ll 1 - \beta \ll Bp_{\min}.$$

The first condition ensures that the EMA tracks the slowly evolving gradient direction, while the second ensures that momentum averaging provides sufficiently small relative task-frequency fluctuations.

Long-time diffusion with momentum. Momentum reduces the instantaneous variance of the buffer but also introduces serial correlations. We now examine their combined effect on the leading long-time diffusion coefficient.

Freeze the margins and write

$$\bar{g} = (p_1 u_1, p_2 u_2)^\top, \qquad n_t := m_t - \bar{g}.$$

The fresh-gradient covariance is

$$\Sigma = \frac{p_1 p_2}{B} \begin{bmatrix} u_1^2 & -u_1 u_2 \\ -u_1 u_2 & u_2^2 \end{bmatrix}.$$

The stationary EMA covariance satisfies

$$\mathrm{Cov}(n_t) = \frac{1-\beta}{1+\beta}\Sigma.$$

More generally,

$$\mathrm{Cov}(n_t, n_{t+k}) = \beta^{|k|}\frac{1-\beta}{1+\beta}\Sigma.$$

Consequently,

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \mathrm{Cov}(n_0, n_k) &= \frac{1-\beta}{1+\beta}\left(1 + 2\sum_{k=1}^{\infty}\beta^k\right)\Sigma \\ &= \Sigma.\end{aligned} \tag{30}$$

Thus the reduction in instantaneous variance is exactly compensated, at the covariance level, by the increased correlation time. To translate this into fluctuations of the gap, define

$$\ell(m) := \sqrt{2}\frac{m_1 - m_2}{\|m\|_2}.$$

Then

$$D(t+1) - D(t) = \eta\, \ell(m_t).$$

Linearizing around $\bar{g}$,

$$\ell(m_t) - \ell(\bar{g}) \simeq \nabla\ell(\bar{g})^\top n_t.$$

Hence, by equation 30,

$$\sum_{k\in\mathbb{Z}} \mathrm{Cov}(\ell(m_0), \ell(m_k)) \simeq \nabla\ell(\bar{g})^\top \Sigma \nabla\ell(\bar{g}).$$

At the full-batch balanced point

$$p_1 u_1 = p_2 u_2,$$

a direct calculation gives

$$\nabla \ell(\bar{g})^\top \Sigma \nabla \ell(\bar{g}) = \frac{1}{B p_1 p_2}. \tag{31}$$

Since $\gamma^*_{B,\beta} \to 0$ in the well-sampled regime, the same coefficient holds to leading order around the momentum-induced equilibrium.

Therefore, on the slow time

$$\tau = \eta t,$$

the local gap dynamics are approximated by

$$d\gamma \simeq -\left(\gamma - \gamma^*_{B,\beta}\right) d\tau + \sqrt{\frac{\eta}{B p_1 p_2}}\, d\mathcal{W}_\tau, \tag{32}$$

up to corrections associated with finite effective sample size and tracking error. In particular, the local stationary fluctuation scale is

$$O_{\mathbb{P}}\left(\sqrt{\frac{\eta}{B p_1 p_2}}\right),$$

which matches the leading fluctuation scale of the momentum-free well-sampled regime.

Combining the momentum-induced equilibrium, the tracking error, and the local diffusion scale gives the late-time approximation

$$D = \log \frac{p_1}{p_2} + \gamma^*_{B,\beta} + O\left(\frac{\eta}{1-\beta}\right) + O_{\mathbb{P}}\left(\sqrt{\frac{\eta}{B p_1 p_2}}\right). \tag{33}$$

Finally, consider a scale-separated regime satisfying

$$\frac{B p_{\min}}{1-\beta} \longrightarrow \infty, \qquad \frac{\eta}{1-\beta} \longrightarrow 0.$$

The first condition implies

$$B_{\text{eff}} p_{\min} \longrightarrow \infty,$$

and hence, by equation 25,

$$\gamma^*_{B,\beta} \longrightarrow 0.$$

Moreover, since $p_1 \geq 1/2$ and $p_2 = p_{\min}$,

$$\frac{\eta}{B p_1 p_2} = O\left(\frac{\eta}{1-\beta}\frac{1-\beta}{B p_{\min}}\right) \longrightarrow 0.$$

Thus the late-time local approximation gives

$$\Delta_1 - \Delta_2 = \log \frac{p_1}{p_2} + o_{\mathbb{P}}(1), \qquad \frac{B p_{\min}}{1-\beta} \to \infty, \qquad \frac{\eta}{1-\beta} \to 0. \tag{34}$$

Momentum therefore acts as cross-step task aggregation. Under frozen margins, it replaces the noisy within-batch frequency by an exponentially averaged frequency with variance-equivalent batch size $B_{\text{eff}} = B(1+\beta)/(1-\beta)$. When the EMA also tracks the slowly evolving margins, this variance reduction drives the gap toward the full-batch value. At the same time, the reduction in instantaneous variance is compensated by increased correlation across steps, leaving the leading local long-time diffusion coefficient unchanged.

D Higher-Rank Sampling Dynamics and Momentum Averaging

Throughout this section, the $K = MC$ tasks are partitioned into M blocks of size C. Tasks in block μ have common population frequency

$$p_j = \frac{\tilde{p}_\mu}{C}, \qquad \sum_{\mu=1}^{M} \tilde{p}_\mu = 1.$$

The organization mirrors Proposition 4.2. Appendix D.1 gives the exact occurrence-gated clock in the frequency-limited regime, and Appendix D.3 gives the balanced full-batch contrast response. Appendices D.4–D.5 then establish how the well-sampled response approaches that full-batch behavior and quantify the residual $O_{\mathbb{P}}((Bp_j)^{-1/2})$ coupling; the cancellation removes the direct contrast perturbation, leaving contrast-to-average coupling as the leading first-order effect. Appendix D.6 treats momentum as averaging across steps with a separate tracking condition. A finite-horizon continuum theorem is retained in Appendix D.7 as a technical refinement.

D.1 Sparse regime: exact occurrence-gated clock and block-symmetrized reference

We begin with the part of the hierarchical dynamics that is exact for arbitrary batch size. The partial polar factor has the same right nullspace as its input: if $Gx = 0$, then $\mathrm{msgn}(G)x = 0$. Indeed, for an SVD $G = U\Sigma V^\top$, $Gx = 0$ implies $\Sigma V^\top x = 0$; the coordinates of $V^\top x$ corresponding to positive singular values therefore vanish, which also gives $\mathrm{msgn}(G)x = 0$.

For task j, if $q_j(t) = 0$ then

$$(I - S_t)Q_t e_j = 0,$$

so

$$(\overline{W}_{t+1} - \overline{W}_t)e_j = 0. \tag{35}$$

Since $\|\mathrm{msgn}(G)\|_{\mathrm{op}} \le 1$,

$$\|\overline{W}_T e_j - \overline{W}_0 e_j\|_2 \le \eta N_j(T), \qquad N_j(T) := \sum_{t<T} \mathbf{1}\{q_j(t) > 0\}. \tag{36}$$

For independent mini-batches, the indicators are i.i.d. Bernoulli with parameter

$$r_j = 1 - (1 - p_j)^B,$$

and hence $N_j(T)/T \to r_j$ almost surely. The expansion $r_j = Bp_j + O((Bp_j)^2)$ in the sparse regime gives the trajectory-level frequency clock used in the main text.

Next define the empirical block masses

$$\pi_{\mu,t} = \sum_{j\in\mu} q_j(t)$$

and the block-averaged count matrix

$$Q_t^{\parallel} = \mathrm{diag}\Big(\frac{\pi_{1,t}}{C} I_C, \ldots, \frac{\pi_{M,t}}{C} I_C\Big).$$

Starting from $\widetilde{W}_0 = 0$, consider

$$\widetilde{W}_{t+1} = \widetilde{W}_t + \eta\,\mathrm{msgn}((I - \widetilde{S}_t)Q_t^{\parallel}).$$

Both the initialization and $Q_t^{\parallel}$ are invariant under permutations within each block, so the trajectory remains block symmetric. The decomposition in the next subsection therefore applies at every step. Whenever $\pi_{\mu,t} > 0$, the contrast singular value on U_μ is positive and the partial polar factor equals the identity there; when $\pi_{\mu,t} = 0$, that contrast block is in the right nullspace. Hence

$$U_\mu^\top(\widetilde{W}_{t+1} - \widetilde{W}_t)U_\mu = \eta \mathbf{1}\{\pi_{\mu,t} > 0\} I_{C-1}, \tag{37}$$

Summing over steps gives

$$U_\mu^\top \widetilde{W}_T U_\mu = \eta N_\mu^{\mathrm{blk}}(T) I_{C-1}, \qquad N_\mu^{\mathrm{blk}}(T) := \sum_{t<T} \mathbf{1}\{\pi_{\mu,t} > 0\}. \tag{38}$$

This sampled block-symmetric process is an exact finite-B benchmark. Appendix D.7 records the stronger finite-horizon continuum refinement obtained after taking large expected task counts and a small step size.

D.2 Block decomposition at a symmetric reference state

For each block μ, let

$$X_\mu \in \mathbb{R}^{C\times(C-1)}$$

have orthonormal columns spanning the zero-sum subspace

$$\left\{x \in \mathbb{R}^C : \mathbf{1}_C^\top x = 0\right\}.$$

Let $U_\mu \in \mathbb{R}^{K\times(C-1)}$ denote the embedding of X_μ into block μ. Define the normalized block-average vectors

$$(y_\mu)_i = \begin{cases} 1/\sqrt{C}, & i \in \mu, \\ 0, & \text{otherwise}, \end{cases}$$

and set

$$Y = (y_1, \dots, y_M), \qquad \mathcal{U} = (U_1, \dots, U_M, Y).$$

Then $\mathcal{U}$ is orthogonal.

Consider a block-symmetric softmax matrix $\widetilde{S}$, and write

$$A := I - \widetilde{S}.$$

Block symmetry implies that each within-block contrast subspace is invariant under A. Hence there exist scalars a_μ and a matrix $L \in \mathbb{R}^{M\times M}$ such that

$$\mathcal{U}^\top A \mathcal{U} = \begin{bmatrix} \operatorname{diag}(a_\mu I_{C-1}) & 0 \\ 0 & L \end{bmatrix}. \tag{39}$$

Equivalently, if d_μ denotes the eigenvalue of $\widetilde{S}$ on the within-block zero-sum subspace of block μ, then

$$a_\mu = 1 - d_\mu.$$

Along the block-symmetric full-batch trajectory considered below, $a_\mu > 0$.

Now let

$$Q^{\parallel} = \operatorname{diag}\left(r_1 I_C, \dots, r_M I_C\right), \qquad r_\mu > 0.$$

In the basis $\mathcal{U}$,

$$\mathcal{U}^\top Q^{\parallel} \mathcal{U} = \begin{bmatrix} \operatorname{diag}(r_\mu I_{C-1}) & 0 \\ 0 & R \end{bmatrix}, \qquad R := \operatorname{diag}(r_1, \dots, r_M).$$

Therefore the block-symmetric reference gradient

$$\widetilde{G} := A Q^{\parallel}$$

takes the form

$$\mathcal{U}^\top \widetilde{G} \mathcal{U} = \begin{bmatrix} \operatorname{diag}(\sigma_\mu I_{C-1}) & 0 \\ 0 & H \end{bmatrix}, \qquad \sigma_\mu := a_\mu r_\mu, \qquad H := LR. \tag{40}$$

Since $\widetilde{S}$ is a positive column-stochastic matrix, $A = I - \widetilde{S}$ has rank $K - 1$. Because $a_\mu > 0$, it follows from equation 39 that L, and hence H, has rank $M - 1$.

D.3 Full-batch balanced contrast dynamics

Under full-batch training,

$$Q = P = \operatorname{diag}\left(\frac{\tilde{p}_1}{C} I_C, \dots, \frac{\tilde{p}_M}{C} I_C\right).$$

Starting from $\overline{W}_0 = 0$, the full-batch dynamics preserve block symmetry. Thus $(I - S_t)P$ has the form equation 40 at every iteration, with

$$r_\mu = \frac{\tilde{p}_\mu}{C}.$$

Since every contrast singular value σ_μ is positive, the matrix sign acts as the identity on every within-block contrast subspace:

$$\mathrm{msgn}((I - S_t)P) = \sum_{\mu=1}^{M} U_\mu U_\mu^\top + Y\,\mathrm{msgn}(H_t)Y^\top. \tag{41}$$

Using

$$\sum_{\mu=1}^{M} U_\mu U_\mu^\top = I_K - YY^\top,$$

we may equivalently write

$$\mathrm{msgn}((I - S_t)P) = I_K + Y\left(\mathrm{msgn}(H_t) - I_M\right)Y^\top. \tag{42}$$

Since

$$\|\mathrm{msgn}(H_t) - I_M\|_{\mathrm{op}} \le 2,$$

every entry of the correction term in equation 42 is $O(1/C)$. Hence

$$\mathrm{msgn}((I - S_t)P) = I_K + O_{\mathrm{entry}}\left(\frac{1}{C}\right).$$

More strongly, the within-block statement is exact:

$$U_\mu^\top\,\mathrm{msgn}((I - S_t)P)\,U_\nu = \delta_{\mu\nu} I_{C-1}. \tag{43}$$

Therefore

$$U_\mu^\top\left(\overline{W}_{t+1} - \overline{W}_t\right)U_\mu = \eta I_{C-1}.$$

Thus full-batch Muon exactly equalizes the update on all within-block contrast directions; all remaining nontrivial structure is confined to the M-dimensional block-average subspace.

D.4 Mini-batch perturbation decomposition around the symmetric reference

We analyze mini-batch sampling fluctuations around a fixed block-symmetric reference state $\widetilde{S}$. The calculation is local and one-step: it quantifies the instantaneous response of the polar map rather than the full departure of S_t from the block-symmetric manifold.

Let

$$Q = \mathrm{diag}(q_1, \ldots, q_K),$$

where Bq is the empirical frequency vector of a mini-batch of size B. Define the empirical block frequencies

$$\pi_\mu := \sum_{j \in \mu} q_j.$$

Decompose

$$Q = Q^{\parallel} + Q^{\perp},$$

where

$$Q^{\parallel} = \mathrm{diag}\left(\frac{\pi_1}{C} I_C, \ldots, \frac{\pi_M}{C} I_C\right). \tag{44}$$

For $j \in \mu$, define the within-block residual

$$(\epsilon_\mu)_j := q_j - \frac{\pi_\mu}{C}.$$

Then

$$\mathbf{1}_C^\top \epsilon_\mu = 0,$$

and the restriction of $Q^\perp$ to block μ is

$$\mathrm{diag}(\epsilon_\mu).$$

Conditioned on π_μ, the $B\pi_\mu$ samples falling in block μ are uniformly distributed among its C tasks. Therefore

$$\mathbb{E}[\epsilon_\mu \mid \pi_\mu] = 0, \tag{45}$$

and

$$\mathrm{Cov}(\epsilon_\mu \mid \pi_\mu) = \frac{\pi_\mu}{BC}\left(I_C - \frac{1}{C}J_C\right). \tag{46}$$

Define

$$z_\mu := X_\mu^\top \epsilon_\mu.$$

Since $X_\mu^\top J_C = 0$,

$$\mathrm{Cov}(z_\mu \mid \pi_\mu) = \frac{\pi_\mu}{BC} I_{C-1}, \tag{47}$$

and hence

$$\mathbb{E}\big[\|z_\mu\|_2^2 \mid \pi_\mu\big] = \frac{(C-1)\pi_\mu}{BC}. \tag{48}$$

Let

$$A := I - \widetilde{S}, \qquad \widetilde{G} := AQ^{\|}, \qquad E := AQ^{\perp}.$$

For the fixed-rank calculation below, work on the event

$$\mathcal{E}_B := \{\pi_\mu > 0 \text{ for every } \mu\}.$$

For fixed M,

$$\mathbb{P}(\mathcal{E}_B^c) \le \sum_{\mu=1}^{M} (1-\tilde{p}_\mu)^B \le M e^{-B\tilde{p}_{\min}},$$

so $\mathbb{P}(\mathcal{E}_B) \to 1$ in the well-sampled regime considered below. On $\mathcal{E}_B$, $Q^{\|}$ is invertible. Moreover, for sufficiently small $|s|$, the diagonal matrix

$$Q^{\|} + sQ^{\perp}$$

remains invertible, so the path

$$G(s) := A\left(Q^{\|} + sQ^{\perp}\right) = \widetilde{G} + sE$$

has constant rank $K-1$ in a neighborhood of $s = 0$.

Because msgn is not an ambient-smooth map at a rank-deficient matrix, we define its derivative here along this constant-rank path:

$$\mathcal{D}_{\widetilde{G}}[E] := \frac{d}{ds}\,\mathrm{msgn}\Big(\widetilde{G} + sE\Big)\Big|_{s=0}. \tag{49}$$

We will use the following standard fixed-rank derivative formula.

Lemma D.1 (Derivative of the partial polar factor along a fixed-rank path)**.** *Let $G(s)$ be a differentiable square-matrix path of constant rank r near $s = 0$, and write*

$$G(0) = U\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_r), \qquad \sigma_i > 0.$$

Let

$$E := G'(0), \qquad \widehat{E} := U^\top E V,$$

and let

$$\dot{P} := \frac{d}{ds}\,\mathrm{msgn}(G(s))\Big|_{s=0}.$$

Partition $U = (U_+, U_0)$ and $V = (V_+, V_0)$ according to the positive and zero singular subspaces. Then

$$\left(U_+^\top \dot{P} V_+\right)_{ij} = \frac{(\widehat{E}_{++})_{ij} - (\widehat{E}_{++})_{ji}}{\sigma_i + \sigma_j}, \tag{50}$$

$$U_+^\top \dot{P} V_0 = \Sigma^{-1}\widehat{E}_{+0}, \tag{51}$$

$$U_0^\top \dot{P} V_+ = \widehat{E}_{0+}\Sigma^{-1}, \tag{52}$$

$$U_0^\top \dot{P} V_0 = 0. \tag{53}$$

Proof. Because the rank is constant, the column and row spaces vary smoothly locally. Choose differentiable orthonormal frames $U_+(s)$ and $V_+(s)$ for these spaces and write

$$G(s) = U_+(s)R(s)V_+(s)^\top,$$

where $R(s) \in \mathbb{R}^{r\times r}$ is invertible. We choose the frames so that $R(0) = \Sigma$. If $P_R(s)$ denotes the polar factor of $R(s)$, then

$$\operatorname{msgn}(G(s)) = U_+(s)P_R(s)V_+(s)^\top, \qquad P_R(0) = I_r.$$

Let

$$A := U_+^\top \dot{U}_+, \qquad B := V_+^\top \dot{V}_+, \qquad \Omega_R := \dot{P}_R(0).$$

Both A and B are skew-symmetric. Differentiating the polar decomposition of the invertible core $R(s)$ at $s = 0$ gives the Sylvester equation

$$\Sigma\Omega_R + \Omega_R\Sigma = \dot{R} - \dot{R}^\top.$$

On the positive singular subspaces,

$$\widehat{E}_{++} = A\Sigma + \dot{R} - \Sigma B,$$

while

$$U_+^\top \dot{P} V_+ = A + \Omega_R - B.$$

Combining the last two displays yields

$$\Sigma\left(U_+^\top \dot{P} V_+\right) + \left(U_+^\top \dot{P} V_+\right)\Sigma = \widehat{E}_{++} - \widehat{E}_{++}^\top,$$

which gives equation 50 entrywise.

For the cross blocks, differentiating $G(s)V_0(s) = 0$ and $U_0(s)^\top G(s) = 0$, or equivalently projecting the derivative of $G(s) = U_+(s)R(s)V_+(s)^\top$, gives

$$U_+^\top \dot{P} V_0 = \Sigma^{-1}\widehat{E}_{+0}, \qquad U_0^\top \dot{P} V_+ = \widehat{E}_{0+}\Sigma^{-1}.$$

Finally, differentiating $\operatorname{msgn}(G(s)) = U_+(s)P_R(s)V_+(s)^\top$ shows that its derivative has no null-to-null block, proving equation 53. □

We next compute the perturbation $E = AQ^\perp$ in the basis $\mathcal{U}$. By equation 39,

$$U_\mu^\top A U_\mu = a_\mu I_{C-1}, \qquad U_\mu^\top A Y = 0, \qquad Y^\top A U_\mu = 0, \qquad Y^\top A Y = L.$$

Moreover,

$$U_\mu^\top Q^\perp U_\mu = D_\mu, \qquad D_\mu := X_\mu^\top \operatorname{diag}(\epsilon_\mu) X_\mu, \tag{54}$$

$$U_\mu^\top Q^\perp Y = \frac{1}{\sqrt{C}} z_\mu e_\mu^\top, \tag{55}$$

$$Y^\top Q^\perp U_\mu = \frac{1}{\sqrt{C}} e_\mu z_\mu^\top, \tag{56}$$

$$Y^\top Q^\perp Y = 0. \tag{57}$$

Here e_μ is the μ-th standard basis vector of $\mathbb{R}^M$. Hence

$$U_\mu^\top E U_\mu = a_\mu D_\mu, \tag{58}$$

$$U_\mu^\top E Y = \frac{a_\mu}{\sqrt{C}} z_\mu e_\mu^\top, \tag{59}$$

$$Y^\top E U_\mu = \frac{1}{\sqrt{C}} L e_\mu z_\mu^\top, \tag{60}$$

$$Y^\top E Y = 0. \tag{61}$$

For $\mu \neq \nu$,

$$U_\mu^\top E U_\nu = 0.$$

D.5 WELL-SAMPLED REGIME: FIRST-ORDER CANCELLATION AND RESIDUAL COUPLING

We now apply Lemma D.1 to the reference gradient $\widetilde{G}$.

Cancellation inside the contrast subspaces. Recall from equation 40 that the singular value of $\widetilde{G}$ on U_μ is

$$\sigma_\mu = a_\mu r_\mu, \qquad r_\mu = \frac{\pi_\mu}{C}.$$

The matrix D_μ in equation 54 is symmetric. Therefore

$$U_\mu^\top E U_\mu = \left(U_\mu^\top E U_\mu\right)^\top.$$

By equation 50,

$$U_\mu^\top \mathcal{D}_{\widetilde{G}}[E] U_\mu = 0.$$

For $\mu \neq \nu$, both cross blocks $U_\mu^\top E U_\nu$ and $U_\nu^\top E U_\mu$ vanish. Hence

$$U_\mu^\top \mathcal{D}_{\widetilde{G}}[E] U_\nu = 0, \qquad \forall\, \mu, \nu. \tag{62}$$

Thus the direct first-order effect of within-block sampling fluctuations vanishes on every contrast-to-contrast block.

Coupling to block-average directions. Let

$$H = \widehat{U}\,\mathrm{diag}(s_1, \ldots, s_M)\widehat{V}^\top$$

be an SVD of the block-average matrix $H = LR$, with

$$s_1, \ldots, s_{M-1} > 0, \qquad s_M = 0.$$

Write $\widehat{u}_k$ and $\widehat{v}_k$ for the corresponding left and right singular vectors.

Since

$$H = LR,$$

the singular-vector identity

$$\widehat{u}_k^\top H = s_k \widehat{v}_k^\top$$

gives, for every k and μ,

$$\widehat{u}_k^\top L e_\mu = \frac{s_k}{r_\mu}(\widehat{v}_k)_\mu. \tag{63}$$

For $s_k = 0$, this follows because R is invertible and $\widehat{u}_k^\top L R = 0$.

Using equations 59 and 60, and the derivative formulas in Lemma D.1, we obtain

$$U_\mu^\top \mathcal{D}_{\widetilde{G}}[E] Y \widehat{v}_k = \frac{(\widehat{v}_k)_\mu}{\sqrt{C}\, r_\mu} \frac{\sigma_\mu - s_k}{\sigma_\mu + s_k} z_\mu, \qquad k = 1, \ldots, M. \tag{64}$$

When $s_k = 0$, the fraction in equation 64 equals one, and the formula agrees with the positive-to-null derivative in equation 51.

Since

$$\left|\frac{\sigma_\mu - s_k}{\sigma_\mu + s_k}\right| \leq 1$$

and $\widehat{V}$ is orthogonal,

$$\begin{aligned} \left\|U_\mu^\top \mathcal{D}_{\widetilde{G}}[E] Y\right\|_F^2 &\leq \frac{\|z_\mu\|_2^2}{C r_\mu^2} \sum_{k=1}^{M} (\widehat{v}_k)_\mu^2 \\ &= \frac{\|z_\mu\|_2^2}{C r_\mu^2}. \end{aligned} \tag{65}$$

Conditioning on π_μ and using equation 48 together with $r_\mu = \pi_\mu / C$,

$$\mathbb{E}\left[\left\| U_\mu^\top \mathcal{D}_{\widetilde{G}}[E]Y \right\|_F^2 \;\middle|\; \pi_\mu \right] \le \frac{1}{C r_\mu^2} \frac{(C-1)\pi_\mu}{BC} = \frac{C-1}{B\pi_\mu}. \tag{66}$$

Therefore, conditionally on $\pi_\mu > 0$,

$$\left\| U_\mu^\top \mathcal{D}_{\widetilde{G}}[E]Y \right\|_F = O_{\mathbb{P}}\left(\sqrt{\frac{C}{B\pi_\mu}} \right). \tag{67}$$

The empirical block frequency satisfies

$$B\pi_\mu \sim \mathrm{Binom}(B, \tilde{p}_\mu).$$

Hence, whenever

$$B\tilde{p}_\mu \to \infty,$$

we have

$$\frac{\pi_\mu}{\tilde{p}_\mu} \xrightarrow{\mathbb{P}} 1.$$

It follows from equation 67 that

$$\left\| U_\mu^\top \mathcal{D}_{\widetilde{G}}[E]Y \right\|_F = O_{\mathbb{P}}\left(\sqrt{\frac{C}{B\tilde{p}_\mu}} \right). \tag{68}$$

Thus the first-order contrast-to-average coupling vanishes when

$$\frac{B\tilde{p}_\mu}{C} \to \infty. \tag{69}$$

For fixed M, the sufficient uniform condition

$$\frac{B\tilde{p}_{\min}}{C} \to \infty$$

makes the first-order coupling small in every block simultaneously. This condition should be interpreted as a large expected sample count per task. It does not require every individual task to appear in every mini-batch.

D.6 Momentum: averaging across steps and tracking

We next analyze momentum as EMA averaging of the empirical frequency matrix. This subsection also proves Proposition 5.1. The variance-equivalent batch-size notation is retained as a shorthand, but the underlying object is the exponentially weighted observation process derived explicitly below. Let the matrix momentum obey

$$\mathsf{M}_t = \beta \mathsf{M}_{t-1} + (1-\beta) G_t, \qquad G_t = (I - S_t) Q_t, \tag{70}$$

with fresh independent mini-batches and $0 \le \beta < 1$. Here G_t denotes the negative gradient in task coordinates, so the parameter update adds its normalized momentum.

Taskwise stationary EMA and finite-initialization correction. For a single task j, the empirical mini-batch frequency satisfies

$$\mathbb{E} q_{j,t} = p_j, \qquad \mathrm{Var}(q_{j,t}) = \frac{p_j(1-p_j)}{B},$$

and the variables are independent across time. For the stationary infinite-past EMA

$$\overline{q}_{j,t} = (1-\beta) \sum_{r=0}^{\infty} \beta^r q_{j,t-r},$$

independence gives

$$\mathbb{E}\overline{q}_{j,t} = p_j,$$
$$\mathrm{Var}(\overline{q}_{j,t}) = (1-\beta)^2 \sum_{r\geq 0} \beta^{2r} \frac{p_j(1-p_j)}{B} = \frac{1-\beta}{1+\beta}\frac{p_j(1-p_j)}{B}. \tag{71}$$

Thus, with

$$B_{\mathrm{eff}}(\beta) = B\frac{1+\beta}{1-\beta},$$

we have

$$\frac{\mathrm{Var}(\overline{q}_{j,t})}{p_j^2} = \frac{1-p_j}{B_{\mathrm{eff}}(\beta)p_j},$$

which proves the frozen-state variance identity in Proposition 5.1.

For comparison with an implementation initialized at $\mathsf{M}_{-1} = 0$, define the bias-corrected finite-history frequency average

$$\widehat{q}_{j,t} := \frac{1-\beta}{1-\beta^{t+1}} \sum_{r=0}^{t} \beta^r q_{j,t-r}.$$

Then $\mathbb{E}\widehat{q}_{j,t} = p_j$ and

$$\frac{\mathrm{Var}(\widehat{q}_{j,t})}{p_j^2} = \frac{1-p_j}{Bp_j}\frac{1-\beta}{1+\beta}\frac{1+\beta^{t+1}}{1-\beta^{t+1}}. \tag{72}$$

The last factor tends to one as $t \to \infty$, recovering the stationary B_{eff}. Without bias correction, the zero-initialized buffer is exactly the truncated stationary sum and has mean $(1-\beta^{t+1})p_j$.

For an evolving state, the stationary infinite-past form of equation 70 is

$$\mathsf{M}_t = (1-\beta)\sum_{r\geq 0} \beta^r (I - S_{t-r}) Q_{t-r}.$$

Adding and subtracting $(I - S_t)Q_{t-r}$ term by term gives the exact decomposition

$$\mathsf{M}_t = (I - S_t)\overline{Q}_t + R_t, \qquad R_t = (1-\beta)\sum_{r\geq 0} \beta^r (S_t - S_{t-r}) Q_{t-r}, \tag{73}$$

where $\overline{Q}_t = (1-\beta)\sum_{r\geq 0}\beta^r Q_{t-r}$. If $\|S_t - S_{t-r}\|_F \leq L_S \eta r$ and $\|Q_{t-r}\|_{\mathrm{op}} \leq 1$, then

$$\|R_t\|_F \leq (1-\beta)\sum_{r\geq 0}\beta^r \|S_t - S_{t-r}\|_F \, \|Q_{t-r}\|_{\mathrm{op}}$$
$$\leq L_S\eta(1-\beta)\sum_{r\geq 0} r\beta^r = L_S\eta\frac{\beta}{1-\beta}, \tag{74}$$

proving the tracking part of Proposition 5.1. A zero-initialized finite-history buffer has the same decomposition with the sums truncated at $r = t$; its bound is no larger, while bias correction multiplies both the frozen term and the remainder by $(1-\beta^{t+1})^{-1}$.

Frozen-state block representation. First freeze the state at a block-symmetric matrix $\widetilde{S}$, and let

$$A := I - \widetilde{S}.$$

After initialization transients,

$$\mathsf{M}_t = (1-\beta)\sum_{k=0}^{\infty}\beta^k A Q_{t-k} = A\overline{Q}_t,$$

where

$$\overline{Q}_t := (1-\beta)\sum_{k=0}^{\infty}\beta^k Q_{t-k}. \tag{75}$$

Thus, under a frozen state, momentum simply replaces the single-batch frequency matrix Q_t by its exponentially weighted average across steps.

Decompose

$$\overline{Q}_t = \overline{Q}_t^{\parallel} + \overline{Q}_t^{\perp}$$

exactly as before. Define

$$\overline{\pi}_\mu(t) := (1-\beta)\sum_{k=0}^{\infty}\beta^k \pi_\mu(t-k),$$

and let $\overline{\epsilon}_\mu(t)$ be the corresponding within-block residual vector.

The batch variables are independent across time. Since

$$\mathbb{E}[\epsilon_\mu(t)] = 0$$

and, by averaging equation 46 over π_μ,

$$\mathrm{Cov}(\epsilon_\mu(t)) = \frac{\tilde{p}_\mu}{BC}\left(I_C - \frac{1}{C}J_C\right),$$

we obtain

$$\mathrm{Cov}(\overline{\epsilon}_\mu) = \frac{1-\beta}{1+\beta}\frac{\tilde{p}_\mu}{BC}\left(I_C - \frac{1}{C}J_C\right). \tag{76}$$

Similarly,

$$\mathrm{Var}(\overline{\pi}_\mu) = \frac{1-\beta}{1+\beta}\frac{\tilde{p}_\mu(1-\tilde{p}_\mu)}{B}. \tag{77}$$

Define

$$B_{\text{eff}} := B\frac{1+\beta}{1-\beta}. \tag{78}$$

Then equations 76 and 77 are exactly the corresponding single-batch *covariance* formulas with B replaced by B_{eff}. This covariance identity is not a distributional identity: $\overline{Q}_t$ is a weighted average of multinomial batches rather than an empirical frequency vector from B_{eff} independent samples.

Let

$$\overline{z}_\mu := X_\mu^\top \overline{\epsilon}_\mu.$$

Thus equation 76 gives

$$\mathbb{E}\|\overline{z}_\mu\|_2^2 = \frac{C-1}{C}\frac{\tilde{p}_\mu}{B_{\text{eff}}}. \tag{79}$$

Moreover, if

$$B_{\text{eff}}\tilde{p}_\mu \to \infty,$$

then

$$\frac{\overline{\pi}_\mu}{\tilde{p}_\mu} \xrightarrow{\mathbb{P}} 1$$

by equation 77.

Combining the deterministic bound in equation 65 with equation 79 and the concentration of $\overline{\pi}_\mu$ therefore yields

$$\left\| U_\mu^\top \mathcal{D}_{\overline{G}^{\parallel}}[\overline{E}]Y \right\|_F = O_{\mathbb{P}}\left(\sqrt{\frac{C}{B_{\text{eff}}\tilde{p}_\mu}}\right), \tag{80}$$

where

$$\overline{G}^{\parallel} = A\overline{Q}^{\parallel}, \qquad \overline{E} = A\overline{Q}^{\perp}.$$

Hence, in probability at first order around the frozen symmetric state, momentum averaging suppresses the sampling distortion whenever

$$\frac{B_{\text{eff}}\tilde{p}_{\min}}{C} \gg 1. \tag{81}$$

Using equation 78, this can also be written, up to the bounded factor $1+\beta$, as

$$\frac{B\tilde{p}_{\min}}{(1-\beta)C} \gg 1. \tag{82}$$

Tracking a slowly evolving state. We now return to the actual dynamics, where S_t evolves. The exact stationary-form expansion of the momentum is

$$\mathsf{M}_t = (1-\beta)\sum_{k=0}^{\infty}\beta^k(I - S_{t-k})Q_{t-k}.$$

If the reference state is frozen at its current value S_t, the corresponding approximation is

$$(I - S_t)\overline{Q}_t.$$

Their difference is

$$\mathsf{M}_t - (I - S_t)\overline{Q}_t = (1-\beta)\sum_{k=0}^{\infty}\beta^k(S_t - S_{t-k})Q_{t-k}. \tag{83}$$

In the late-time regime, the block-symmetric reference quantities evolve on the parameter timescale $1/\eta$. Locally, if their relative change per iteration is $O(\eta)$, then over k iterations the mismatch is $O(k\eta)$. Since

$$(1-\beta)\sum_{k=0}^{\infty}k\beta^k = \frac{\beta}{1-\beta},$$

equation 83 gives the adiabatic tracking scale

$$O\left(\frac{\eta}{1-\beta}\right). \tag{84}$$

Under the stated local $O(\eta)$-variation assumption, the frozen-state mismatch is therefore small on the adiabatic scale

$$\eta \ll 1-\beta. \tag{85}$$

This is a tracking estimate, not a global stability theorem for the stochastic trajectory.

Thus the local momentum regime is described by two separate requirements:

$$\frac{B\tilde{p}_{\min}}{(1-\beta)C} \gg 1, \tag{86}$$

for small relative sampling fluctuations under momentum, and

$$\frac{\eta}{1-\beta} \ll 1, \tag{87}$$

for adiabatic tracking. The first improves the frozen-state sampling estimate relative to the no-momentum condition $B\tilde{p}_{\min}/C \gg 1$; the second prevents the averaging window from becoming so long that the underlying gradient changes substantially across it.

D.7 Technical refinement: finite-horizon well-sampled continuum limit

The main text does not use convergence to the full-batch continuum flow as the substantive higher-rank claim; large expected sample counts making mini-batch frequencies concentrate is expected. We retain the following result because it formalizes how the local sampling analysis connects to the full-batch trajectory on any fixed continuum-time horizon.

Proposition D.2 (Finite-horizon well-sampled continuum limit)**.** *Fix M, C and positive task frequencies. Let $\overline{W}^{\eta,B}(\tau)$ be the piecewise-linear interpolation of the stochastic hierarchy on slow time $\tau = \eta t$, and let $W^\star$ solve the full-batch continuum flow $\dot{W}^\star = \mathrm{msgn}((I - S(W^\star))P)$ from $W^\star(0) = 0$. For every finite τ_0 and $\varepsilon > 0$,*

$$\lim_{\eta\downarrow 0}\lim_{Bp_{\min}\to\infty}\mathbb{P}\left(\sup_{0\le\tau\le\tau_0}\|\overline{W}^{\eta,B}(\tau) - W^\star(\tau)\|_F > \varepsilon\right) = 0. \tag{88}$$

The same holds along any joint sequence $\eta_n \downarrow 0$, $B_n p_{\min} \to \infty$ satisfying $\log(1/\eta_n) = o(B_n p_{\min})$. Moreover, along the limiting flow,

$$U_\mu^\top W^\star(\tau) U_\nu = \delta_{\mu\nu}\,\tau I_{C-1}, \qquad \tau \ge 0. \tag{89}$$

From large expected task count to the continuum trajectory. We now prove Proposition D.2. Define

$$F(W) := \mathrm{msgn}((I - S(W))P).$$

For every finite W, the softmax matrix $S(W)$ has strictly positive entries and is column stochastic. Perron–Frobenius therefore implies that its eigenvalue 1 is simple, so $I - S(W)$, and hence $(I - S(W))P$, has rank $K - 1$. The partial polar factor is smooth on the fixed-rank manifold in a neighborhood of each such matrix. Moreover

$$\|F(W)\|_F \le \sqrt{K-1},$$

so the ODE

$$\dot{W}^\star(\tau) = F(W^\star(\tau)), \qquad W^\star(0) = 0,$$

has a unique global solution. On every compact time interval $[0, \tau_0]$, the image of $W^\star$ is compact and the smallest nonzero singular value of $(I - S(W^\star(\tau)))P$ has a positive minimum. Thus F is uniformly Lipschitz on a sufficiently small tube around that trajectory.

The ODE preserves permutations within every frequency block because P and the initialization share those symmetries and the polar map is orthogonally equivariant. Therefore the block decomposition of Appendix D.3 applies for every τ. In particular,

$$U_\mu^\top \dot{W}^\star(\tau) U_\nu = \delta_{\mu\nu} I_{C-1}. \tag{90}$$

Integrating from $W^\star(0) = 0$ proves equation 89 for every $\tau \ge 0$.

We first prove the sequential limit in equation 88. Fix $\eta > 0$ and $\tau_0 < \infty$. Then only $N = \lceil \tau_0/\eta \rceil$ mini-batches are involved. As $Bp_{\min} \to \infty$, multinomial concentration and a union bound over these finitely many steps give

$$\max_{t<N} \|Q_t P^{-1} - I\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0.$$

Uniform fixed-rank smoothness of the polar map on a tube around the corresponding full-batch Euler trajectory then implies that the stochastic recursion converges uniformly over these N steps to

$$V_{t+1} = V_t + \eta F(V_t), \qquad V_0 = 0.$$

Finally, sending $\eta \downarrow 0$, the piecewise-linear Euler interpolation converges uniformly on $[0, \tau_0]$ to $W^\star$. This proves the iterated well-sampled/small-step limit without any coupling between B and η.

For completeness, we now prove the convenient joint sufficient scaling stated in the proposition. Let

$$N_n := \left\lceil \frac{\tau_0}{\eta_n} \right\rceil, \qquad \kappa_n = B_n p_{\min}.$$

For a single mini-batch, multiplicative Chernoff bounds give, for $0 < \delta < 1$,

$$\mathbb{P}\left(\max_j \frac{|q_j - p_j|}{p_j} > \delta \right) \le 2K \exp(-c\delta^2 \kappa_n) \tag{91}$$

for a universal constant $c > 0$. Set

$$\delta_n := \left(\frac{\log(2KN_n)}{\kappa_n} \right)^{1/4}.$$

The assumption $\log(1/\eta_n) = o(\kappa_n)$ implies $\delta_n \to 0$ and

$$N_n \, 2K \exp(-c\delta_n^2 \kappa_n) \to 0.$$

Hence, with probability tending to one,

$$\max_{0 \le t < N_n} \|Q_t P^{-1} - I\|_{\mathrm{op}} \le \delta_n. \tag{92}$$

In particular all task counts are positive on this event for all sufficiently large n, so every sampled gradient remains on the same rank-$(K-1)$ manifold as the full-batch gradient.

Let $V_{t+1} = V_t + \eta_n F(V_t)$, $V_0 = 0$, denote the deterministic Euler discretization of the continuum flow. Uniform fixed-rank smoothness on the tube around $W^\star([0, \tau_0])$ gives constants $L_{\mathrm{dyn}}, a_{\mathrm{samp}} < \infty$ such that, on the event in equation 92,

$$\begin{aligned} &\| \operatorname{msgn}((I - S(W))Q_t) - F(W)\|_F \le a_{\mathrm{samp}}\delta_n, \\ &\|F(W) - F(V)\|_F \le L_{\mathrm{dyn}}\|W - V\|_F. \end{aligned} \tag{93}$$

A standard bootstrap then keeps both discrete trajectories inside the same tube and yields

$$\max_{t \le N_n} \|\overline{W}_t - V_t\|_F \le a_{\mathrm{samp}}\tau_0 e^{L_{\mathrm{dyn}}\tau_0}\delta_n = o_{\mathbb{P}}(1).$$

Because F is Lipschitz on the same compact tube, the usual Euler estimate gives

$$\sup_{0 \le \tau \le \tau_0} \|V^{\eta_n}(\tau) - W^\star(\tau)\|_F = O(\eta_n),$$

where V^{η_n} denotes the piecewise-linear interpolation. Combining the two estimates proves equation 88.

Finally, equation 89 implies for every fixed $\tau > 0$,

$$\frac{1}{\tau} U_\mu^\top \overline{W}^{\eta_n, B_n}(\tau) U_\mu \xrightarrow{\mathbb{P}} I_{C-1}.$$

Hence also

$$\lim_{\tau \to \infty} \lim_{n \to \infty} \frac{1}{\tau} U_\mu^\top \overline{W}^{\eta_n, B_n}(\tau) U_\mu = I_{C-1} \quad \text{in probability.} \tag{94}$$

This is an iterated long-time statement rather than a claim of uniform stochastic approximation on the entire half-line: the nonzero singular gap of the gradient may shrink as training proceeds. The important distinction from the previous finite-step bound is that the number of discrete optimizer steps already diverges as $1/\eta_n$ on every fixed continuum interval, while the limiting dynamics themselves are global in τ and have an exact asymptotic contrast rate.

E Stochastic Labels: Exact Statistics, Projected Regimes, and Local Closure

We now derive the stochastic-gradient structure under symmetric label noise, assuming $0 < \alpha < 1$.

Roadmap and status of the results. Appendices E.1–E.3 concern the complete categorical model and are exact: they give the gradient decomposition, covariance/reliability formulas, and the fixed-dimensional local Gaussian limit. Appendix E.4 introduces the longitudinal target-versus-average-nontarget projection; the counting formulas are exact *within that reduced model* and describe the occurrence-limited and corruption-sparse regimes. Appendices E.5–E.6 add a moment-matched isotropic covariance closure (and a local frequency-homogeneous sector) to obtain the explicit $\sqrt{p}$ coefficient in the corruption-resolved local regime. The exact categorical covariance is anisotropic, so the closure result should not be read as an exact formula for the full matrix process.

Define

$$c_\alpha := 1 - \alpha + \frac{\alpha}{K}, \qquad \rho_\alpha := 1 - c_\alpha = \alpha\frac{K-1}{K}.$$

Thus c_α is the probability that the observed label equals the target label and ρ_α is the probability that the observed label differs from the target label. The transition matrix is

$$D_\alpha = (1 - \alpha)I_K + \frac{\alpha}{K}\mathbf{1}\mathbf{1}^\top.$$

E.1 Complete categorical model: gradient dynamics

Let J denote the sampled task and Y the observed label. We have

$$\mathbb{P}(J = j) = p_j, \qquad \mathbb{P}(Y = i \mid J = j) = p_{i|j},$$

with

$$p_{i|j} = \begin{cases} c_\alpha, & i = j, \\ \alpha/K, & i \neq j. \end{cases}$$

The population loss under label noise is

$$\mathcal{L}(W) = -\sum_{j=1}^{K} p_j \sum_{i=1}^{K} p_{i|j} \log S_{ij}(W).$$

In task coordinates, differentiation gives

$$\nabla \mathcal{L}(W) = -\widetilde{E}(D_\alpha - S)PE^\top, \qquad P := \mathrm{diag}(p_1, \ldots, p_K). \tag{95}$$

For a mini-batch of size B, let N_{ij} be the number of samples with task j and observed label i, and define the empirical joint frequency matrix

$$\widehat{T}_{ij} := \frac{N_{ij}}{B}.$$

The empirical task frequencies are

$$q_j := \sum_i \widehat{T}_{ij}, \qquad Q := \mathrm{diag}(q_1, \ldots, q_K).$$

The negative stochastic gradient in task coordinates is therefore

$$\widehat{T} - SQ. \tag{96}$$

If $q_j > 0$, one may equivalently write $\widehat{T}_{\cdot j} = q_j R_{\cdot j}$, where $R_{\cdot j}$ is the empirical conditional label distribution. When $q_j = 0$, the column of R is immaterial because only $RQ = \widehat{T}$ enters the gradient.

A useful decomposition is

$$\widehat{T} - SQ = (D_\alpha - S)P + L + T, \tag{97}$$

where

$$L := \widehat{T} - D_\alpha Q, \tag{98}$$

$$T := (D_\alpha - S)(Q - P). \tag{99}$$

Here L is the conditional label-sampling fluctuation and T is the task-frequency fluctuation. At the noisy population optimum

$$S^* = D_\alpha,$$

the task-frequency term T vanishes exactly, so the persistent local stochasticity comes from L.

E.2 Complete categorical model: exact covariance and reliability

From this point through the local-response analysis we work in task coordinates; the orthogonal embedding transforms preserve all covariance, norm, and polar-equivariance statements used below.

Let

$$d_j := D_\alpha e_j, \qquad \delta_j := Se_j - d_j, \qquad \Gamma_j := \mathrm{diag}(d_j) - d_j d_j^\top.$$

For one observation (J, Y), define the task-j positive-gradient contribution

$$X_j := \mathbf{1}\{J = j\}(Se_j - e_Y).$$

Then

$$\mathbb{E}X_j = p_j \delta_j, \qquad \mathbb{E}[X_j X_j^\top] = p_j(\Gamma_j + \delta_j \delta_j^\top).$$

Averaging B independent observations gives

$$\mathrm{Cov}(\widehat{g}_j) = \frac{1}{B}\left[p_j \Gamma_j + p_j(1 - p_j)\delta_j \delta_j^\top\right], \tag{100}$$

and for $j \neq k$,

$$\mathrm{Cov}(\widehat{g}_j, \widehat{g}_k) = -\frac{p_j p_k}{B} \delta_j \delta_k^\top. \tag{101}$$

Therefore

$$\mathrm{SNR}_j^2 = \frac{B p_j \|\delta_j\|_2^2}{\mathrm{tr}\,\Gamma_j + (1-p_j)\|\delta_j\|_2^2}. \tag{102}$$

At $S^* = D_\alpha$, all $\delta_j = 0$. Writing $\widehat{G} := SQ - \widehat{T}$ for the positive stochastic gradient, the population Hessian is block diagonal across task columns with blocks $p_j \Gamma_j$, so

$$\mathrm{Cov}\Big(\mathrm{vec}\,\widehat{G}(W^*)\Big) = \frac{1}{B} H^* \tag{103}$$

on the columnwise zero-sum subspace. This likelihood identity motivates a local structured-noise viewpoint, but it does not imply that applying a polar map implements $(H^*)^{-1/2}$.

We now isolate the conditional label-sampling term. Let $\ell_j := L e_j$. From equation 98,

$$\ell_j = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{J_b = j\}\big(e_{Y_b} - d_j\big).$$

Hence

$$\mathbb{E}[\ell_j] = 0.$$

Different task columns are uncorrelated, and

$$\mathrm{Cov}(\ell_j) = \frac{p_j}{B} \Gamma_j. \tag{104}$$

In particular,

$$\mathrm{Var}(L_{jj}) = \frac{p_j}{B} c_\alpha \rho_\alpha, \tag{105}$$

$$\mathrm{Var}(L_{ij}) = \frac{p_j}{B} \frac{\alpha}{K} \left(1 - \frac{\alpha}{K}\right), \qquad i \neq j, \tag{106}$$

and, for $i \neq k$,

$$\mathrm{Cov}(L_{ij}, L_{kj}) = -\frac{p_j}{B} p_{i|j} p_{k|j}.$$

The geometry of Γ_j is especially simple on the zero-sum subspace. Let

$$P_0 := I_K - \frac{1}{K} \mathbf{1}\mathbf{1}^\top.$$

A direct calculation gives

$$\Gamma_j = \frac{\alpha}{K} P_0 + \alpha(1-\alpha)(P_0 e_j)(P_0 e_j)^\top. \tag{107}$$

Define the correct-versus-average-incorrect contrast

$$v_j := e_j - \frac{1}{K-1} \sum_{i \neq j} e_i.$$

The covariance eigenvalue along the normalized version of v_j is

$$\lambda_\parallel = \alpha c_\alpha,$$

whereas the other $K-2$ zero-sum directions have eigenvalue

$$\lambda_\perp = \frac{\alpha}{K}.$$

Equivalently, if ζ_j is the coefficient of ℓ_j along the unnormalized vector v_j, then

$$\mathrm{Var}(\zeta_j) = \frac{p_j}{B} c_\alpha \rho_\alpha. \tag{108}$$

These formulas identify the two basic expected-count scales. A batch contains $B p_j$ task-j examples on average and

$$B p_j \rho_\alpha = B p_j \alpha \frac{K-1}{K}$$

incorrectly labeled task-j examples on average. The latter scale resolves the aggregate correct-versus-incorrect fluctuation. If one instead requires each individual incorrect class to be well sampled, the stronger scale $B p_j \alpha / K$ is relevant.

E.3 Complete categorical model: local Gaussian limit and structured response

The covariance calculation above determines the complete local sampling limit before any isotropic closure. The complete categorical model admits a fixed-dimensional limit that retains every zero-sum fluctuation mode; this separates exact local sampling behavior from the auxiliary closures introduced later only to obtain explicit scalar rates. Let

$$\mathcal{S} = \{A \in \mathbb{R}^{K\times K} : \mathbf{1}^\top A = 0\}$$

with the Frobenius inner product. Choose the zero-column-mean representative W^* of a population minimizer and let the Hessian operator satisfy

$$\mathcal{H}^*[A]e_j = p_j \Gamma_j A e_j.$$

Theorem E.1 (Local sampling limit with complete categorical noise)**.** *Fix K, $p_j > 0$, and $0 < \alpha < 1$. For deterministic $A \in \mathcal{S}$, let $W_B = W^* + A/\sqrt{B}$. Then the positive mini-batch gradient obeys*

$$\sqrt{B}\,\widehat{G}_B(W_B) \xrightarrow{d} \mathcal{H}^*[A] + Z, \tag{109}$$

where $Z \in \mathcal{S}$ is centered Gaussian with vectorized covariance $\mathcal{H}^*$. Moreover,

$$\lim_{B\to\infty} \mathbb{E}\ \mathrm{msgn}(\widehat{G}_B(W_B)) = \mathbb{E}\ \mathrm{msgn}(\mathcal{H}^*[A] + Z). \tag{110}$$

The limit keeps K, p, α fixed and is not uniform over vanishing task frequencies or vanishing label noise.

Proof. Smoothness of the finite-dimensional softmax loss gives $\sqrt{B}\,\nabla\mathcal{L}(W_B) = \mathcal{H}^*[A] + o(1)$. Per-sample gradients are uniformly bounded near W^* and their covariance converges to $\mathcal{H}^*$, so the triangular-array central limit theorem gives equation 109. Choose an orthonormal basis $X \in \mathbb{R}^{K\times(K-1)}$ for the zero-sum subspace. In these coordinates the Gaussian covariance is positive definite, so $X^\top(\mathcal{H}^*[A] + Z)$ has full row rank almost surely; equivalently, $\mathcal{H}^*[A] + Z$ has rank $K-1$ almost surely. Every stochastic gradient also lies in $\mathcal{S}$. Writing any $G \in \mathcal{S}$ as $G = XB$ gives $\mathrm{msgn}(G) = X\,\mathrm{msgn}(B)$; the rectangular partial polar map is continuous on the open set of full-row-rank $(K-1)\times K$ matrices. Hence the square partial polar map is continuous at every rank-$(K-1)$ point relative to $\mathcal{S}$, and the continuous mapping theorem applies on this support. Since every entry of a partial polar factor is bounded by one, convergence in distribution upgrades to convergence of expectations, and positive scale invariance removes the factor $\sqrt{B}$. □

The theorem isolates the remaining question: how does the polar map respond to an anisotropic Gaussian law? In orthonormal coordinates on $\mathcal{S}$, let Z be centered nondegenerate Gaussian with covariance Σ and define

$$F_\Sigma(A) = \mathbb{E}\ \mathrm{msgn}(Z + A).$$

Gaussian density differentiation gives the exact linear-response operator

$$DF_\Sigma(0)[A] = \mathbb{E}\big[\mathrm{msgn}(Z)\,\langle \Sigma^{-1}\,\mathrm{vec}_{\mathcal{S}}\,A, \mathrm{vec}_{\mathcal{S}}\,Z\rangle\big]\,. \tag{111}$$

The inverse is taken only on the zero-sum support, not on softmax gauge directions. The same Gaussian-smoothing argument as Appendix A.1 implies a cubic fixed-dimensional remainder after this linear term. Unlike the fully isotropic setting, however, no left–right orthogonal symmetry forces equation 111 to be a scalar multiple of A. Anisotropy is therefore necessary for nontrivial directional response in this local Gaussian picture, but is not by itself sufficient to guarantee improved conditioning.

The exact Gaussian limit above retains the full anisotropic response but does not simplify to a scalar closed form. The following subsections therefore study complementary reductions: first a longitudinal counting model that is exact within its projection, and then an isotropic covariance closure that preserves total label-noise energy while sacrificing anisotropy.

E.4 PROJECTED COUNTING MODEL: TASK AND CORRUPTION COUNTS

We quantify the occurrence-limited and corruption-sparse regimes in a projected target-versus-average-nontarget model. The projection retains the longitudinal component along v_j and removes the transverse categorical fluctuations in equation 107. Within this reduced model, the counting argument below is exact; the Gaussian approximation enters only later.

Consider a symmetry-equivalent task direction with frequency p and margin Δ. Write

$$u(\Delta) := 1 - S_{jj} = \frac{K-1}{e^{\Delta} + K - 1}.$$

Let N be the number of samples of this task in a batch and let M be the number of those samples whose observed label is incorrect. For each of the B draws, the three possibilities have probabilities

$$pc_\alpha, \qquad p\rho_\alpha, \qquad 1-p,$$

corresponding to a correctly labeled occurrence of the task, an incorrectly labeled occurrence, and another task, respectively. Hence

$$\mathbb{P}(N=n, M=m) = \frac{B!}{(n-m)!\, m!\, (B-n)!}(pc_\alpha)^{n-m}(p\rho_\alpha)^m(1-p)^{B-n}, \tag{112}$$

for $0 \le m \le n \le B$.

The correct-coordinate residual in the stochastic gradient column is

$$\widehat{T}_{jj} - S_{jj}q_j = \frac{N - M - NS_{jj}}{B}.$$

Since the residual column has zero sum, this is also its coefficient along the unnormalized contrast v_j. Therefore the longitudinal random coefficient is

$$A(\Delta) := \frac{Nu(\Delta) - M}{B}. \tag{113}$$

The corresponding scalar sign response is

$$m_{B,p}(\Delta) := \mathbb{E}[\operatorname{sgn}(A(\Delta))] = \sum_{n=0}^{B}\sum_{m=0}^{n} \operatorname{sgn}(nu(\Delta) - m)\, \mathbb{P}(N=n, M=m), \tag{114}$$

with the convention $\operatorname{sgn}(0) = 0$.

To connect this scalar response to the projected Muon dynamics, consider s exchangeable directions with coefficients $A_1, \dots, A_s$, and let

$$X \in \mathbb{R}^{s\times(s-1)}$$

have orthonormal columns spanning the zero-sum subspace. Define

$$\mathcal{A}_s(\Delta) := X^\top \operatorname{diag}(A_1(\Delta), \dots, A_s(\Delta))X.$$

Permutation symmetry implies

$$\mathbb{E}[\operatorname{msgn}(\mathcal{A}_s(\Delta))] = r_s(\Delta) I_{s-1} \tag{115}$$

for a scalar $r_s(\Delta)$. By Cauchy interlacing between the eigenvalues of $\mathcal{A}_s$ and the diagonal entries $A_1, \dots, A_s$,

$$|r_s(\Delta) - m_{B,p}(\Delta)| \le \frac{2}{s-1}. \tag{116}$$

Thus $m_{B,p}$ gives the large-sector mean Muon response in the projected model, with an explicit $O(s^{-1})$ correction.

Sparse-task limit. Suppose $Bp \to 0$. With probability $1 - O((Bp)^2)$, the task appears at most once. Whenever it appears once, an observation matching the target label gives $A(\Delta) > 0$, whereas a mismatched observation gives $A(\Delta) < 0$, for every finite Δ. Therefore

$$m_{B,p}(\Delta) = Bp(c_\alpha - \rho_\alpha) + O\big((Bp)^2\big) = Bp(1-2\rho_\alpha) + O\big((Bp)^2\big). \tag{117}$$

This expansion is uniform for Δ in compact sets. Hence, for $\rho_\alpha < 1/2$, the scalar counting dynamics have positive mean margin drift proportional to Bp. For the finite-sector matrix response r_s, the additive bound in equation 116 alone does not establish the same leading rate as $Bp \to 0$; transferring this asymptotic requires additional control, such as $s^{-1} = o(Bp)$.

Large-margin response. We next let $\Delta \to \infty$, so that $u(\Delta) \downarrow 0$. For each fixed batch realization,

$$\operatorname{sgn}(A(\Delta)) \longrightarrow \mathbf{1}\{N > 0, M = 0\} - \mathbf{1}\{M > 0\}.$$

Now

$$\mathbb{P}(M = 0) = (1 - p\rho_\alpha)^B, \qquad \mathbb{P}(N = 0) = (1 - p)^B.$$

Therefore dominated convergence gives

$$\lim_{\Delta \to \infty} m_{B,p}(\Delta) = \chi_{B,p} := 2(1 - p\rho_\alpha)^B - (1 - p)^B - 1. \tag{118}$$

Since $u(\Delta)$ decreases with Δ, every summand in equation 114 is nonincreasing in Δ. Thus $m_{B,p}(\Delta)$ is nonincreasing and $\chi_{B,p}$ is its minimum tail value.

Consider the large-sector projected mean-field recursion

$$\Delta_{t+1}^{\mathrm{mf}} - \Delta_t^{\mathrm{mf}} = \eta m_{B,p}(\Delta_t^{\mathrm{mf}}). \tag{119}$$

If $\chi_{B,p} > 0$, then $m_{B,p}(\Delta) \geq \chi_{B,p} > 0$ for every finite Δ, so the limiting mean-field recursion in equation 119 satisfies $\Delta_t^{\mathrm{mf}} \to \infty$. Since $m_{B,p}(\Delta) \to \chi_{B,p}$ along that trajectory,

$$\frac{\Delta_t^{\mathrm{mf}}}{\eta t} \longrightarrow \chi_{B,p}. \tag{120}$$

For a finite projected sector, equation 116 additionally guarantees positive mean drift whenever $2/(s-1) < \chi_{B,p}$. Consequently,

$$1 - S_{jj}(t) \asymp (K - 1)e^{-\eta \chi_{B,p} t}$$

along this target-interpolating mean-field trajectory.

Corruption-sparse window and corruption threshold. We use $\ll$ and $\gg$ in this paragraph as scale-separation notation. For fixed nonzero ρ_α, the corruption-sparse case is the intermediate window

$$1 \ll Bp \ll \rho_\alpha^{-1}.$$

Within this window, $(1 - p)^B \leq e^{-Bp}$ and a binomial expansion gives

$$\chi_{B,p} = 1 - 2Bp\rho_\alpha + O\big((Bp\rho_\alpha)^2 + e^{-Bp}\big). \tag{121}$$

Thus the projected large-margin growth rate is close to the unit deterministic-label full-batch rate and is no longer occurrence-limited.

A sharper crossover is obtained in the Poisson scaling

$$Bp \to \infty, \qquad p\rho_\alpha \to 0, \qquad Bp\rho_\alpha \to \lambda.$$

Then

$$\chi_{B,p} \longrightarrow 2e^{-\lambda} - 1. \tag{122}$$

Hence the large-margin regime has positive drift for $\lambda < \log 2$ and negative drift for $\lambda > \log 2$. Equivalently, once task absence is negligible, the crossover batch scale is

$$B_{\mathrm{noise}} \simeq \frac{\log 2}{p\rho_\alpha}. \tag{123}$$

The regime $Bp \gg 1$ but $Bp\rho_\alpha \ll 1$ therefore has a quantitatively stable large-margin regime in the reduced model, while resolved corruption reverses the projected large-margin drift. The finite population optimum exists for every $0 < \alpha < 1$, independently of batch size; the local response around it is analyzed separately below.

E.5 MOMENT-MATCHED ISOTROPIC COVARIANCE CLOSURE

A Gaussian proxy with the exact column covariance of L can be represented as

$$L_{\mathrm{G}} = \frac{1}{\sqrt{B}} P_0 \left[\sqrt{\frac{\alpha}{K}} Z + \sqrt{\alpha(1-\alpha)}\, \mathrm{diag}(g) \right] P^{1/2}, \tag{124}$$

where Z has i.i.d. standard Gaussian entries and g has i.i.d. standard Gaussian coordinates, independently. This proxy makes the anisotropic categorical covariance explicit: it contains both a projected dense component and a projected diagonal component.

For a closed-form polar response, we now replace this anisotropic covariance by a moment-matched isotropic covariance on the zero-sum subspace. The exact covariance in equation 107 has one longitudinal eigenvalue αc_α and $K-2$ transverse eigenvalues α/K; the closure preserves their total covariance energy while removing the directional anisotropy. This is the additional approximation that allows us to reuse the Gaussian polar response from Section 3.2.

From equation 107,

$$\mathrm{tr}(\Gamma_j) = \alpha(2-\alpha)\frac{K-1}{K}. \tag{125}$$

The zero-sum subspace has dimension $K-1$. We therefore define

$$\tau_\alpha^2 := \frac{\mathrm{tr}(\Gamma_j)}{K-1} = \frac{\alpha(2-\alpha)}{K}, \tag{126}$$

and make the moment-matched covariance approximation

$$\Gamma_j \approx \tau_\alpha^2 P_0. \tag{127}$$

This approximation preserves the total label-noise variance on the zero-sum subspace but removes the longitudinal anisotropy.

Let

$$X \in \mathbb{R}^{K\times(K-1)}$$

have orthonormal columns spanning the zero-sum subspace, so that

$$X^\top X = I_{K-1}, \qquad XX^\top = P_0.$$

Since every stochastic-gradient column sums to zero, we may write

$$G = X\widehat{G}, \qquad \widehat{G} := X^\top G \in \mathbb{R}^{(K-1)\times K}.$$

The partial polar factors are related by

$$\mathrm{msgn}(G) = X\,\mathrm{msgn}(\widehat{G}). \tag{128}$$

For an explicit closed-form response, we additionally use a local frequency-homogeneous approximation: the symmetry-equivalent columns under consideration are assigned a common frequency scale p. This is a local spectral proxy, not an exact reduction of a heterogeneous matrix: the polar factor couples columns, and their unequal noise scales generally break right-orthogonal symmetry. Varying p compares proxies at different local frequency scales; it does not vary the common probability of all K tasks in one normalized uniform distribution. Under equation 127, the projected Gaussian label noise is then approximated by

$$\widehat{L} \approx \sigma_{p,\alpha}\Xi, \qquad \sigma_{p,\alpha}^2 := \frac{p\tau_\alpha^2}{B} = \frac{p\alpha(2-\alpha)}{BK}, \tag{129}$$

where

$$\Xi \in \mathbb{R}^{(K-1)\times K}$$

has i.i.d. standard Gaussian entries.

We use the rectangular analogue of Proposition 3.1. For a rectangular matrix $A = U\Sigma V^\top$ with singular values σ_i, we use the same partial-polar definition

$$\mathrm{msgn}(A) := U\,\mathrm{diag}(\mathbf{1}\{\sigma_i > 0\})V^\top.$$

Thus positive singular values are mapped to one and zero singular values remain zero. If $\Xi \in \mathbb{R}^{m\times n}$ has i.i.d. standard Gaussian entries, define

$$c_{m,n} := \frac{1}{mn}\mathbb{E}\|\Xi\|_*. \tag{130}$$

For fixed m, n, Gaussian convolution and $O(m) \times O(n)$ equivariance give

$$\mathbb{E}[\mathrm{msgn}(H + \sigma\Xi)] = \frac{c_{m,n}}{\sigma} H + O_{m,n}\left(\frac{\|H\|_F^3}{\sigma^3}\right) \tag{131}$$

as $\|H\|_F/\sigma \to 0$. The coefficient follows by taking the trace of the derivative:

$$mn\, c_{m,n} = \mathbb{E}\langle \mathrm{msgn}(\Xi), \Xi\rangle_F = \mathbb{E}\|\Xi\|_*.$$

Thus equation 131 is the direct rectangular counterpart of the square Gaussian response derived in Appendix A.1.

E.6 Local Muon rate under the covariance closure

Under the equal-incorrect-logit reduction in a symmetry-equivalent sector, define

$$\Delta_j := \overline{W}_{jj} - \frac{1}{K-1}\sum_{i\neq j} \overline{W}_{ij}.$$

Then

$$S_{jj} = \frac{e^{\Delta_j}}{e^{\Delta_j} + K - 1}.$$

The noisy population optimum satisfies

$$S^*_{jj} = c_\alpha,$$

which gives

$$\Delta^*_j = \log\left(\frac{K}{\alpha} - (K-1)\right). \tag{132}$$

Define

$$r_j := c_\alpha - S_{jj}.$$

Under the equal-incorrect-logit reduction,

$$(D_\alpha - S)e_j = r_j v_j, \qquad v_j = e_j - \frac{1}{K-1}\sum_{i\neq j} e_i. \tag{133}$$

For a task of local frequency scale p, the projected deterministic signal is therefore

$$\widehat{H}_j = p r_j x_j e_j^\top, \qquad x_j := X^\top v_j. \tag{134}$$

Because v_j lies in the zero-sum subspace,

$$\|x_j\|_2^2 = \|v_j\|_2^2 = \frac{K}{K-1}. \tag{135}$$

Combining equations 129 and 131 with $m = K-1$ and $n = K$ gives

$$\mathbb{E}[\mathrm{msgn}(\widehat{H}_j + \widehat{L})] \approx \frac{c_{K-1,K}}{\sigma_{p,\alpha}}\widehat{H}_j \tag{136}$$

in the small-signal regime near the population optimum under label noise.

The margin functional is

$$\Delta_j = v_j^\top \overline{W} e_j.$$

Using equations 128, 134, and 135, the expected Muon margin increment becomes

$$\mathbb{E}[\Delta_j(t+1) - \Delta_j(t) \mid \Delta_j(t)] \approx \eta \frac{c_{K-1,K}}{\sigma_{p,\alpha}} p r_j \frac{K}{K-1}. \tag{137}$$

Finally,

$$\frac{\partial S_{jj}}{\partial \Delta_j} = S_{jj}(1 - S_{jj}),$$

and at the optimum,

$$S^*_{jj}(1 - S^*_{jj}) = c_\alpha \rho_\alpha.$$

Hence

$$r_j = c_\alpha \rho_\alpha \big(\Delta^*_j - \Delta_j\big) + O\big((\Delta_j - \Delta^*_j)^2\big)\,. \tag{138}$$

Since

$$\rho_\alpha = \alpha \frac{K-1}{K},$$

substituting equations 129 and 138 into equation 137 yields

$$\mathbb{E}[\Delta_j(t+1) - \Delta_j(t) \mid \Delta_j(t)] \approx \eta\, c_{K-1,K} \sqrt{BKp}\, \frac{c_\alpha \sqrt{\alpha}}{\sqrt{2-\alpha}} \big(\Delta^*_j - \Delta_j(t)\big). \tag{139}$$

Thus the covariance and locally homogeneous frequency approximations predict a local Muon conditional-mean contraction coefficient proportional to $\sqrt{p}$.

For comparison, under the same symmetric-logit reduction, the conditional mean SGD update obeys

$$\mathbb{E}[\Delta_j(t+1) - \Delta_j(t) \mid W_t] = \eta \frac{K}{K-1} p_j r_j(t),$$

so its exact local conditional-mean contraction coefficient is

$$\eta \frac{K}{K-1} p_j c_\alpha \rho_\alpha, \tag{140}$$

which scales linearly with p_j.

For large K, the quarter-circle law gives

$$c_{K-1,K}\sqrt{K} \longrightarrow \frac{8}{3\pi}, \tag{141}$$

while $c_\alpha \to 1 - \alpha$. Therefore equation 139 reduces to

$$\mathbb{E}[\Delta_j(t+1) - \Delta_j(t) \mid \Delta_j(t)] \approx \eta\, \frac{8}{3\pi} \frac{(1-\alpha)\sqrt{\alpha}}{\sqrt{2-\alpha}} \sqrt{Bp}\, \big(\Delta^*_j - \Delta_j(t)\big). \tag{142}$$

Scope of the approximation. The covariance formulas in equations 104–107 are exact. The explicit rate equation 139 uses two additional approximations: the anisotropic categorical covariance is replaced by the isotropic covariance equation 127 with matching trace, and the columns in the local spectral sector are assigned a common frequency scale p. These approximations restore the left-right isotropy needed for a closed-form polar response while retaining the total label-noise energy. They are intended for the corruption-resolved local regime; the occurrence-limited and corruption-sparse regimes are instead described by the counting analysis in Appendix E.4.

F Muon–SignSGD Sparse-Support Contrast: Complete Derivation

This appendix derives the two-task SignSGD dynamics and makes explicit how the result depends on rotations of the embedding space. We use the elementwise convention $\mathrm{sgn}(0) = 0$.

F.1 Rank-one reduction

In task coordinates, the two-task negative stochastic gradient is

$$v g_t^\top, \qquad v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad g_t = \begin{bmatrix} q_1(t) u_1(t) \\ -q_2(t) u_2(t) \end{bmatrix}.$$

Hence, in the original parameter coordinates, the negative gradient is

$$\widetilde{E} v g_t^\top E^\top = (\widetilde{E} v)(E g_t)^\top.$$

For arbitrary vectors a, b,

$$\operatorname{sgn}(ab^\top) = \operatorname{sgn}(a)\operatorname{sgn}(b)^\top,$$

where the sign is applied elementwise. Therefore the SignSGD update in task coordinates is

$$\overline{W}_{t+1} - \overline{W}_t = \eta \widetilde{E}^\top \operatorname{sgn}(\widetilde{E}v)\operatorname{sgn}(Eg_t)^\top E.$$

Define the margin vector

$$h_t := \begin{bmatrix} \Delta_1(t) \\ -\Delta_2(t) \end{bmatrix} = \overline{W}_t^\top v.$$

Then

$$\begin{aligned} h_{t+1} - h_t &= \eta E^\top \operatorname{sgn}(Eg_t)\operatorname{sgn}(\widetilde{E}v)^\top \widetilde{E}v \\ &= \eta \widetilde{c}\, E^\top \operatorname{sgn}(Eg_t), \end{aligned}$$

where

$$\widetilde{c} := \operatorname{sgn}(\widetilde{E}v)^\top \widetilde{E}v = \|\widetilde{E}v\|_1 > 0. \tag{143}$$

Thus the answer embedding contributes only the positive scalar $\widetilde{c}$ to the two margins, whereas the query embedding determines their coupling.

F.2 Rotation-dependent margin recursion

Consider the representative rotation

$$E = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \qquad c := \cos\theta, \qquad s := \sin\theta, \qquad 0 < \theta < \frac{\pi}{2}.$$

Then

$$Eg_t = \begin{bmatrix} cq_1u_1 - sq_2u_2 \\ -sq_1u_1 - cq_2u_2 \end{bmatrix}.$$

The second coordinate is strictly negative because $q_1 + q_2 = 1$ and $u_1, u_2 > 0$. Define

$$\xi_t := \operatorname{sgn}\left(cq_1(t)u_1(t) - sq_2(t)u_2(t)\right). \tag{144}$$

Hence

$$\operatorname{sgn}(Eg_t) = \begin{bmatrix} \xi_t \\ -1 \end{bmatrix},$$

and

$$E^\top \operatorname{sgn}(Eg_t) = \begin{bmatrix} c\xi_t + s \\ s\xi_t - c \end{bmatrix}.$$

Since the second entry of h_t is $-\Delta_2(t)$, the exact margin recursion is

$$\Delta_1(t+1) - \Delta_1(t) = \eta\widetilde{c}(c\xi_t + s), \tag{145}$$

$$\Delta_2(t+1) - \Delta_2(t) = \eta\widetilde{c}(c - s\xi_t). \tag{146}$$

Equivalently, for

$$D_t := \Delta_1(t) - \Delta_2(t),$$

we have

$$D_{t+1} - D_t = \eta\widetilde{c}\left((c+s)\xi_t + s - c\right). \tag{147}$$

In particular,

$$\xi_t = 1 \quad \Longrightarrow \quad D_{t+1} - D_t = 2\eta\widetilde{c}\, s,$$

and

$$\xi_t = -1 \quad \Longrightarrow \quad D_{t+1} - D_t = -2\eta\widetilde{c}\, c.$$

If $\xi_t = 0$, then

$$D_{t+1} - D_t = \eta\widetilde{c}(s - c).$$

F.3 FULL-BATCH REFERENCE DYNAMICS

Set $q_i(t) = p_i$. Define the sum and gap

$$A_t := \Delta_1(t) + \Delta_2(t), \qquad D_t := \Delta_1(t) - \Delta_2(t).$$

From equations 145–146,

$$A_{t+1} - A_t = \begin{cases} 2\eta\widetilde{c}\,c, & \xi_t = 1, \\ \eta\widetilde{c}(c+s), & \xi_t = 0, \\ 2\eta\widetilde{c}\,s, & \xi_t = -1. \end{cases}$$

Therefore

$$A_t \geq 2\eta\widetilde{c}\min\{c, s\}\, t, \tag{148}$$

so $A_t \to \infty$.

For fixed A, write

$$\Delta_1 = \frac{A+D}{2}, \qquad \Delta_2 = \frac{A-D}{2}.$$

The switching variable in equation 144 has the same sign as

$$\Lambda(A, D) := \log\frac{p_1 c}{p_2 s} + \log\left(1 + e^{(A-D)/2}\right) - \log\left(1 + e^{(A+D)/2}\right). \tag{149}$$

For each fixed A, $\Lambda(A, D)$ is strictly decreasing in D, with limits $+\infty$ and $-\infty$ as $D \to -\infty$ and $D \to +\infty$, respectively. Hence there is a unique switching threshold $T(A)$ satisfying

$$\Lambda(A, T(A)) = 0.$$

Moreover,

$$T(A) \longrightarrow D_\theta := \log\frac{p_1}{p_2} + \log\cot\theta \qquad \text{as } A \to \infty. \tag{150}$$

On either strict side of the switching threshold, the recursion moves the gap toward that threshold:

$$D_t < T(A_t) \quad \Longrightarrow \quad \xi_t = 1,$$

while

$$D_t > T(A_t) \quad \Longrightarrow \quad \xi_t = -1.$$

At equality, $\xi_t = 0$ and

$$D_{t+1} - D_t = \eta\widetilde{c}(s - c),$$

so the one-step displacement is uniformly bounded. Since $T(A)$ is bounded for $A \geq 0$, these one-step bounds imply that D_t is bounded. Combining this with equation 148 gives

$$\Delta_1(t) \to \infty, \qquad \Delta_2(t) \to \infty.$$

Consequently, equation 150 applies along the trajectory.

More precisely, for every $\varepsilon > 0$, for all sufficiently large t,

$$D_\theta - \varepsilon \leq T(A_t) \leq D_\theta + \varepsilon.$$

An upward crossing of the threshold changes D_t by at most $2\eta\widetilde{c}\,s$, while a downward crossing changes it by at most $2\eta\widetilde{c}\,c$. Therefore

$$D_\theta - 2\eta\widetilde{c}\,c \leq \liminf_{t\to\infty} D_t \leq \limsup_{t\to\infty} D_t \leq D_\theta + 2\eta\widetilde{c}\,s. \tag{151}$$

Thus the discrete full-batch gap need not converge to a single limit; rather, it remains in an $O(\eta)$ switching band around the embedding-dependent location D_θ. In the small-step limit this band collapses to D_θ.

We next compute the asymptotic growth rates. Rearranging equation 147 gives

$$\frac{1}{t}\sum_{k=0}^{t-1}\xi_k = \frac{c-s}{c+s} + \frac{D_t - D_0}{\eta\widetilde{c}(c+s)t}.$$

Since D_t is bounded,

$$\frac{1}{t}\sum_{k=0}^{t-1}\xi_k \longrightarrow \frac{c-s}{c+s}. \tag{152}$$

Averaging equations 145 and 146 then yields

$$\frac{\Delta_1(t)}{\eta t} \longrightarrow \frac{\widetilde{c}}{c+s}, \qquad \frac{\Delta_2(t)}{\eta t} \longrightarrow \frac{\widetilde{c}}{c+s}. \tag{153}$$

Hence full-batch SignSGD balances the two tasks at leading order, but both the switching location and the common growth rate depend on the embedding rotation.

F.4 Sparse-support negative transfer

Consider the event

$$\mathcal{A}_t := \{q_2(t) = 0\},$$

meaning that the batch contains no examples from task 2. Since every sample is then task 1,

$$\mathbb{P}(\mathcal{A}_t) = p_1^B.$$

On $\mathcal{A}_t$,

$$\xi_t = 1,$$

and hence

$$\Delta_2(t+1) - \Delta_2(t) = \eta\widetilde{c}(c-s). \tag{154}$$

Therefore, whenever

$$\theta > \frac{\pi}{4},$$

a batch containing no task 2 examples decreases the task 2 margin.

A stronger statement follows by bounding all remaining updates. For every batch,

$$\Delta_2(t+1) - \Delta_2(t) \le \eta\widetilde{c}(c+s).$$

Consequently,

$$\Delta_2(t) \le \Delta_2(0) + \eta\widetilde{c}\left[(c-s)\sum_{k<t}\mathbf{1}_{\mathcal{A}_k} + (c+s)\sum_{k<t}\mathbf{1}_{\mathcal{A}_k^c}\right].$$

By the strong law of large numbers,

$$\frac{1}{t}\sum_{k<t}\mathbf{1}_{\mathcal{A}_k} \longrightarrow p_1^B \qquad \text{a.s.}$$

Therefore

$$\limsup_{t\to\infty}\frac{\Delta_2(t)}{\eta\widetilde{c}\,t} \le c + s\left(1 - 2p_1^B\right) \qquad \text{a.s.} \tag{155}$$

If

$$p_1^B > \frac{1+\cot\theta}{2}, \tag{156}$$

then the right-hand side of equation 155 is strictly negative, and therefore

$$\Delta_2(t) \to -\infty$$

linearly almost surely. Condition (156) can only hold when $\theta > \pi/4$. In this regime, equation 145 also gives

$$\Delta_1(t+1) - \Delta_1(t) \ge \eta\widetilde{c}(s-c) > 0.$$

Thus insufficiently small relative task-frequency fluctuations can be converted by the embedding rotation into persistent negative transfer from the frequent task to the rare task.

For Muon, orthogonal equivariance allows the dynamics to be analyzed directly in task coordinates. When $q_2(t) = 0$, the exact two-task Muon recursion in equation 3 gives

$$\Delta_2(t+1) - \Delta_2(t) = 0.$$

The rotation-induced negative-transfer mechanism above is therefore specific to the coordinate-sensitive elementwise sign operation.

G SUPPLEMENTARY EMA BIAS–VARIANCE CALCULATION

This optional scalar calculation makes the familiar EMA bias–variance trade-off explicit for an exogenous affine drifting signal. It is included only to interpret the tracking side of Proposition 5.1; it is not a separate Muon convergence result or a method-level contribution. For reference, write

$$R(\beta; v, b) := v\frac{1-\beta}{1+\beta} + b^2\frac{\beta^2}{(1-\beta)^2}, \tag{157}$$

and denote the first-order optimality condition by

$$b^2\beta(1+\beta)^2 = v(1-\beta)^3. \tag{158}$$

G.1 AFFINE-DRIFT CALCULATION

Let $y_t = a_t + \varepsilon_t$, where $a_t = a_0 + bt$ and the noise variables are independent with mean zero and variance $v > 0$. Define $m_t = (1-\beta)\sum_{r\geq 0}\beta^r y_{t-r}$ for $0 \leq \beta < 1$. The deterministic weighted sum converges absolutely, and the noise weights are square summable, so the infinite-past EMA is well defined in mean square. Its mean is

$$\mathbb{E}m_t = (1-\beta)\sum_{r\geq 0}\beta^r(a_t - br) = a_t - \frac{\beta}{1-\beta}b,$$

while independence gives

$$\mathrm{Var}(m_t) = v(1-\beta)^2\sum_{r\geq 0}\beta^{2r} = v\frac{1-\beta}{1+\beta}.$$

Adding squared bias and variance proves equation 157. For $0 < \beta < 1$,

$$\partial_\beta R = -\frac{2v}{(1+\beta)^2} + \frac{2b^2\beta}{(1-\beta)^3}.$$

The function $\beta(1+\beta)^2/(1-\beta)^3$ increases strictly from zero to infinity, so for $b \neq 0$ the derivative has exactly one root and changes sign from negative to positive. This gives equation 158. Writing $\epsilon = 1-\beta^*$ gives

$$\epsilon^3 = (b^2/v)(1-\epsilon)(2-\epsilon)^2,$$

which yields $1-\beta^* \sim (4b^2/v)^{1/3}$ as $b^2/v \to 0$.

Corollary G.1 (Strict advantage over a shared constant memory)**.** *For finitely many directions, let $v_j > 0$, $b_j \neq 0$, and $w_j > 0$, and let β_j^* minimize $R(\beta; v_j, b_j)$. If two ratios b_j^2/v_j differ, then*

$$\sum_j w_j R(\beta_j^*; v_j, b_j) < \min_{0\leq\beta<1}\sum_j w_j R(\beta; v_j, b_j).$$

Proof. The strict monotonicity in the optimality equation gives distinct unique minimizers whenever the ratios b_j^2/v_j differ. Every shared coefficient is therefore strictly suboptimal for at least one positive-weight direction. The shared objective is continuous and diverges as $\beta \uparrow 1$, so it attains its minimum on a compact subset of $[0, 1)$. Evaluating the strict pointwise inequality at that minimizer proves the claim. □

The corollary compares constant coefficients in an exogenous affine-signal estimation problem. It is not a lower bound against every adaptive scalar-momentum method and does not establish end-to-end acceleration for Muon.

H NUMERICAL EXPERIMENTS AND REPRODUCIBILITY DETAILS

The synthetic experiments in this appendix diagnose the displayed theoretical mechanisms using the exact partial polar factor. The preliminary language-model pretraining comparison is documented separately in Appendix H.8.

H.1 COMMON PROTOCOL AND MAP FROM MAIN-TEXT FIGURES

Where each main-text figure is documented.

Main-text figure	Appendix protocol	Role in the paper
Fig. 1	App. H.2	Direct check of Proposition 3.1.
Fig. 2(a)	App. H.3	Direct check of Theorem 4.1.
Fig. 2(b)	App. H.1, H.4	Trajectory-level regime summary plus a frozen-state check of the perturbation rate underlying the well-sampled regime in Proposition 4.2.
Fig. 3	App. H.5, H.6	Full-matrix transfer/stress tests kept separate from the projected counting law and covariance closure used in Proposition 4.4.
Fig. 4	App. H.7	Paired long-time trajectory check of the sparse-support negative-transfer mechanism in Proposition 4.5.
Fig. 5	App. H.3	Direct check of the variance-equivalent sampling and tracking scales in Proposition 5.1.
Fig. 6	App. H.8	Single-run validation-loss comparison in language-model pretraining.

Two replication notions. Unless specified otherwise, synthetic trajectory experiments use $R_{\mathrm{traj}} = 64$ independent mini-batch streams per configuration, and frozen-state checks use $N_{\mathrm{MC}} = 8192$ independent Monte Carlo replicates. Experiment-specific sample counts are given below, including larger samples for Gaussian coefficient estimates. Trajectory runs measure variability of an entire optimization path, whereas frozen-state replicates estimate a conditional expectation or coupling norm. For antithetic estimators, uncertainty should be computed across independent pairs, because the two evaluations within each pair are dependent. Unless stated otherwise, curves show means, trajectory ribbons show interquartile ranges, and confidence intervals are 1.96 standard errors over independent replicates. Log–log exponents are ordinary least-squares fits to the displayed grid points. The pretraining comparison contains one run per method and has no repeated-run uncertainty estimate.

Exact Muon map and numerical invariances. Every synthetic Muon simulation uses the exact SVD partial polar factor rather than an approximate Newton–Schulz implementation. If $A = U\,\mathrm{diag}(\sigma_1, \ldots)V^\top$ and $\sigma_{\max} > 0$, singular values are treated as numerical zeros only relative to the matrix scale, using a cutoff proportional to $\sigma_{\max}$; the zero matrix maps to zero. This preserves positive-scale invariance. Deterministic regression tests verify positive-scale invariance, left–right orthogonal equivariance, the rank-one polar formula, and agreement between the full 2×2 matrix update and the reduced two-task recursion, all to errors below 4×10^{-15}.

Hierarchical trajectory experiment in Figure 2. We simulate the deterministic-label associative-memory recursion directly in task coordinates,

$$W_{t+1} = W_t + \eta\ \mathrm{msgn}\big((I - S(W_t))Q_t\big), \qquad Bq_t \sim \mathrm{Multinomial}(B, p), \qquad W_0 = 0.$$

There are $M = 4$ frequency blocks of size $C = 6$ ($K = 24$ tasks) with block masses $(0.55, 0.25, 0.13, 0.07)$, so each task in block μ has probability $\widetilde{p}_\mu / C$. We use $B \in \{8, 64, 256, 1024\}$, step size $\eta = 0.02$, and $T = 800$ updates. For block μ with index set $\mathcal{I}_\mu$, the plotted margin is

$$m_\mu(t) = \frac{1}{C} \sum_{j \in \mathcal{I}_\mu} \left(W_{jj}(t) - \frac{1}{C-1} \sum_{\substack{i \in \mathcal{I}_\mu \\ i \neq j}} W_{ij}(t) \right).$$

The late growth rate is the OLS slope of m_μ against continuous time $\tau = \eta t$ over the last 40% of the recorded trajectory. The main text summarizes the resulting sampling collapse in panel (b) of Figure 2; the raw small- and large-expected-count trajectories are retained below for reference.

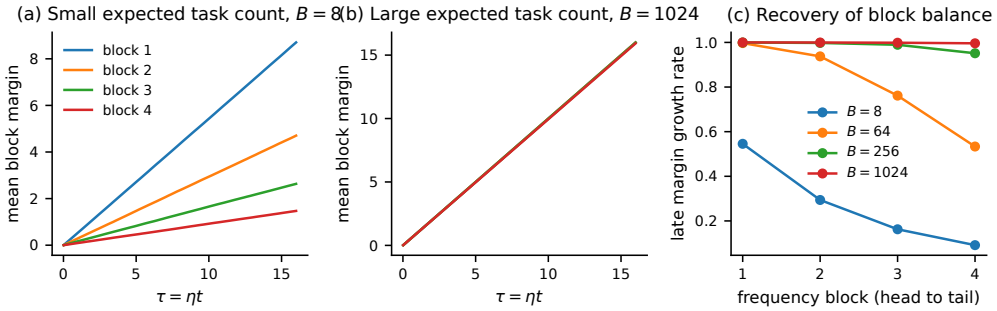


Figure 7: **Higher-rank trajectory view underlying Figure 2(b).** At small expected task count ($B = 8$), block margins grow at visibly different rates; at large expected task count ($B = 1024$), all four blocks approach the unit full-batch rate. The main text uses the rate-versus-Bp_j collapse because it exposes the controlling expected-count variable more directly.

Diagnostic	measured	prediction
Isotropic anisotropic-signal median matrix error	3.05%	$\to 0$ as $\varepsilon \to 0$
Two-task equilibrium-gap bias exponent	-1.011	-1
Two-task gap-fluctuation exponent	-0.531	$-1/2$
Hierarchical spatial-coupling exponent	-0.512	$-1/2$
Hierarchical momentum-coupling exponent	-0.509	$-1/2$
Projected large-margin drift RMSE	0.00618	0
Direct isotropic-closure frequency exponent	0.501	$1/2$
Direct isotropic-closure mean coefficient error	0.61%	$\to 0$ in closure limit
Full-categorical stress-test exponent	0.494	$1/2$ under auxiliary closure

Table 1: **Quantitative summary of the controlled numerical checks.** Exact and reduced-model validations are kept separate from transfer tests. The isotropic and closure errors are finite-signal/finite-Monte-Carlo quantities whose limiting prediction is zero; the full-categorical label-noise exponent is a stress test rather than a direct validation of the isotropic closure.

H.2 Isotropic Gaussian mean response with an anisotropic signal

The first-order theorem is directional, so we use a deliberately non-isotropic signal rather than a multiple of the identity. For each $d \in \{4, 8\}$, define

$$H_d = \frac{\operatorname{diag}(s_1, \dots, s_d)}{\|\operatorname{diag}(s_1, \dots, s_d)\|_F}, \qquad (s_1, \dots, s_d) = \operatorname{geomspace}(1, 0.08; d).$$

For $\varepsilon \in \{0.04, 0.07, 0.12, 0.20\}$ we estimate

$$F_d(\varepsilon) = \mathbb{E}\ \operatorname{msgn}(\varepsilon H_d + \Xi)$$

using 32,768 antithetic Gaussian pairs. The coefficient $c_d = d^{-2}\mathbb{E}\|\Xi\|_*$ is estimated independently from 65,536 Gaussian matrices, so agreement does not arise from sharing Monte Carlo noise between response and theory. We report both the scalar projection $\langle \widehat{F}_d(\varepsilon), H_d\rangle/\varepsilon$ and the full matrix residual

$$\frac{\|\widehat{F}_d(\varepsilon) - c_d\varepsilon H_d\|_F}{\|c_d\varepsilon H_d\|_F}.$$

Across the eight settings, the median full-matrix relative error is 3.05%. For the directional-gain panel in Figure 1(a), we additionally repeat the $d = 8$ experiment with 131,072 antithetic pairs and an independent 131,072-sample estimate of c_d, then report the gain along each singular direction divided by c_d.

The corresponding response curves are shown in Figure 1 in the main text.

H.3 Two-task finite-batch transition, asymptotics, and momentum

Direct finite-batch transition. To test Theorem 4.1 itself, we use $p_{\min} \in \{0.02, 0.05, 0.10\}$ and choose batch sizes spanning both sides of B_{bal}. For each configuration we estimate the late-time

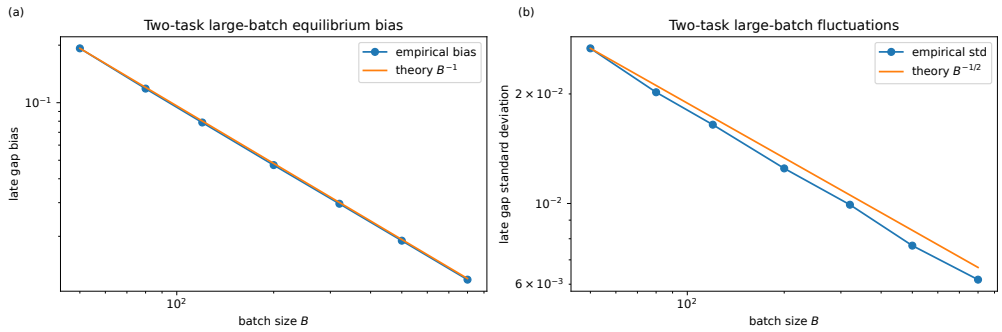


Figure 8: **Two-task large-batch asymptotics.** Both panels use $p_2 = 0.05$, $\eta = 0.003$, and 64 trajectories. (a) Absolute late-time bias of $\Delta_1 - \Delta_2$ from the full-batch target $\log(p_1/p_2)$ versus batch size B on log–log axes; the fitted exponent is -1.011, compared with the B^{-1} prediction. (b) Late-time stationary standard deviation of the gap versus B; the fitted exponent is -0.531, compared with the $B^{-1/2}$ diffusion prediction.

growth rate of the gap $D(t) = \Delta_1(t) - \Delta_2(t)$ across 64 trajectories and plot it against B/B_{bal}. In the small-batch phase,

$$\frac{D(t)}{\eta t} \longrightarrow \sqrt{2}\big(2p_1^B - 1\big),$$

whereas the balanced phase has zero asymptotic gap slope. The three frequency settings collapse around $B/B_{\mathrm{bal}} = 1$ and agree with the piecewise theoretical curve.

The direct finite-batch transition plot is panel (a) of Figure 2 in the main text.

Large-batch asymptotics. We use the exact two-task recursion of Appendix C with $p_2 = 0.05$, $p_1 = 0.95$, $\eta = 0.003$, $T = 30{,}000$, and

$$B \in \{50, 80, 120, 200, 320, 500, 800\}.$$

For each of the 64 trajectories we discard the first 60% of updates. Over the remaining segment we estimate the time-averaged gap $\Delta_1 - \Delta_2$ and its stationary standard deviation (sampling every five iterations). The equilibrium target is $\log(p_1/p_2)$. Regressing the mean equilibrium bias and the stationary standard deviation against B gives exponents -1.011 and -0.531, respectively, close to the predicted -1 and $-1/2$.

Variance-equivalent sampling under momentum. For the momentum-collapse check we keep $p_2 = 0.05$, use $\eta = 0.002$ and $T = 30{,}000$, and target $B_{\mathrm{eff}} \in \{40, 80, 160, 320\}$. For every $B \in \{2, 4, 8, 16, 32\}$ with $B \le B_{\mathrm{eff}}$, we choose

$$\beta = \frac{B_{\mathrm{eff}}/B - 1}{B_{\mathrm{eff}}/B + 1},$$

so that $B(1+\beta)/(1-\beta) = B_{\mathrm{eff}}$. The same 60% burn-in convention is used. For the tracking check we instead fix $B = 4$ and $\beta = 0.95$, sweep

$$\eta \in \{5\times10^{-4}, 0.002, 0.01, 0.03, 0.05, 0.10, 0.20\},$$

and choose

$$T = \mathrm{round}(50/\eta),$$

so that the continuous-time horizon satisfies $\tau_{\max} = \eta T \approx 50$ for every learning rate. We again discard the first 60% of updates and compute the mean gap over the retained late-time window; the plotted mean-gap error is its absolute deviation from the full-batch target $\log(p_1/p_2)$. The measured error grows once $\eta/(1-\beta)$ is no longer small.

The corresponding collapse/tracking summary is shown directly in Figure 5 of the main text.

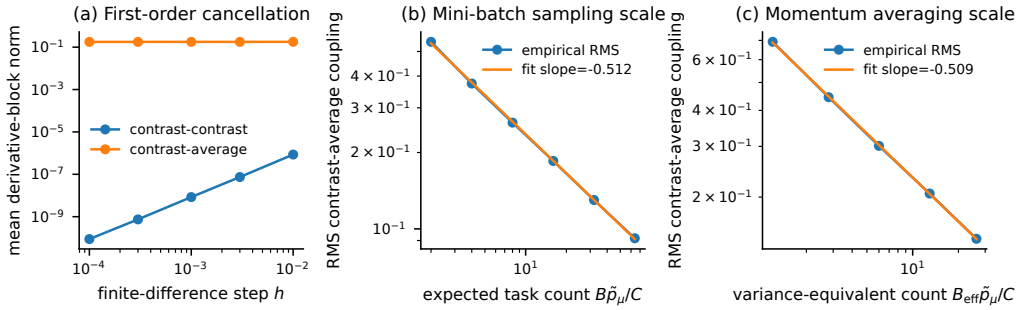


Figure 9: **Higher-rank perturbation checks at a frozen block-symmetric state.** (a) Mean finite-difference derivative norm versus step h: the contrast-to-contrast block decays to numerical zero while contrast-to-average coupling remains $O(1)$, verifying first-order cancellation. (b) RMS contrast-to-average coupling versus expected task count $B\widetilde{p}_\mu/C$; the fitted log–log slope is -0.512, matching the predicted $-1/2$. (c) The same coupling versus variance-equivalent task count $B_{\text{eff}}\widetilde{p}_\mu/C$ under EMA momentum; the fitted slope is -0.509. Each point uses $N_{\text{MC}} = 8192$ draws.

H.4 Hierarchical well-sampled perturbation scaling

These checks are deliberately *frozen-state* Monte Carlo experiments rather than optimization trajectories. We set $M = 3$, $C = 8$, block masses $(0.5, 0.3, 0.2)$, and test the third block. To remove any unspecified initialization, the reference logit matrix is fixed deterministically as follows. If column j belongs to block ν and row i to block μ, then

$$W_{ij} = \begin{cases} 2 + 0.1\nu, & i = j, \\ 0.3 + 0.05\nu, & \mu = \nu,\ i \neq j, \\ -0.2 - 0.05\mu + 0.03\nu, & \mu \neq \nu, \end{cases}$$

with zero-based block indices $\mu, \nu \in \{0, 1, 2\}$. We set $A = I - S(W)$, draw a multinomial frequency vector q, decompose it into its block-average and within-block parts $q = q^{\parallel} + q^{\perp}$, and estimate the derivative

$$\frac{\operatorname{msgn}(A \operatorname{diag}(q^{\parallel} + hq^{\perp})) - \operatorname{msgn}(A \operatorname{diag}(q^{\parallel} - hq^{\perp}))}{2h}.$$

Each point below averages N_{MC} independent draws of q.

For cancellation we fix $B = 640$ and use $h \in \{10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}\}$. For spatial scaling we fix $h = 10^{-4}$ and sweep $B \in \{80, 160, 320, 640, 1280, 2560\}$. For momentum scaling we freeze the same state, set $B = 16$, and sweep $\beta \in \{0.50, 0.75, 0.875, 0.94, 0.97\}$. For each β, the EMA frequency vector is initialized at p and run for $\lceil 12/(1-\beta) \rceil$ mini-batches before measuring the same finite-difference derivative. The resulting spatial and momentum exponents are -0.512 and -0.509, respectively.

H.5 Label-noise large-margin drift: projected validation and full-matrix transfer

Projected large-margin formula. The large-margin drift $\chi_{B,p}$ in equation 118 is exact for the longitudinal projected counting model, not for the complete categorical matrix. We validate that reduced model directly before testing transfer to full matrix Muon. We use $K = 8$, $p_j = 1/8$, margin 12, $\alpha \in \{0.05, 0.10, 0.20\}$, and expected corrupted-sample counts

$$\lambda = Bp_j\rho_\alpha \in \{0.20, 0.35, 0.50, \log 2, 1.0, 1.4, 2.0\}$$

(up to integer rounding of B). Each point uses 16,384 samples. Direct projected Monte Carlo has RMSE 0.00618 from $\chi_{B,p}$. The complete matrix update shows the same sign transition around $\lambda \simeq \log 2$ but is not expected to equal the projected formula; its RMSE relative to the projected prediction is 0.0233.

The full-matrix transfer summary is shown in panel (a) of Figure 3; the direct projected-model validation is isolated below so that reduced-model validation and full-matrix transfer are not conflated.

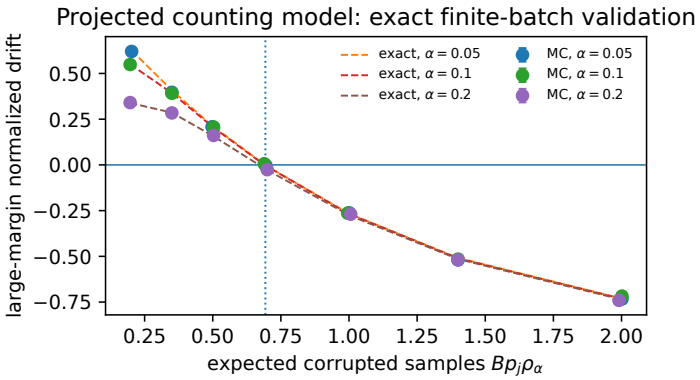


Figure 10: **Direct validation of the projected large-margin counting law.** Points are Monte Carlo estimates in the longitudinal projected model and dashed curves are the exact finite-B formula. The full-matrix transfer test is shown in Figure 3(a), where deviations from the projected formula are expected because transverse categorical modes are retained.

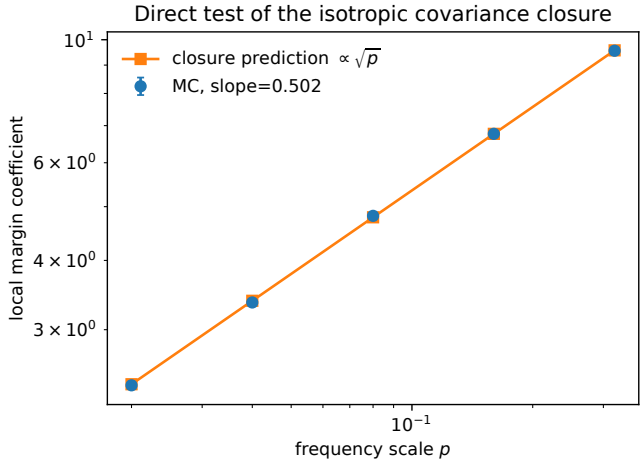


Figure 11: **Direct check of the moment-matched isotropic covariance closure.** The Monte Carlo coefficient follows the predicted $p^{1/2}$ scaling with fitted exponent 0.502. This panel validates the stated Gaussian proxy itself; the main text uses the full-categorical simulation as the more demanding stress test.

H.6 Label-noise local response: direct closure test and categorical stress test

Direct test of the stated isotropic covariance closure. We simulate exactly the Gaussian proxy used to derive equation 139. We take $K = 6$, $\alpha = 0.2$, $B = 5000$, local displacement $h = 0.01$, and frequency scales

$$p \in \{0.02, 0.04, 0.08, 0.16, 0.32\}.$$

The projected signal and isotropic Gaussian covariance follow Appendix E.6, and each point uses $32{,}768$ antithetic samples. Here p is the frequency scale of the local proxy, not a common probability assigned to all six tasks in a normalized categorical distribution. The fitted frequency exponent is 0.502 and the mean relative error of the explicit finite-K coefficient is 0.61%.

The direct Gaussian-closure check is retained below; panel (b) of Figure 3 instead shows the stronger full-categorical frequency-scaling stress test.

Full-categorical heterogeneous-frequency stress test. We separately retain the stronger complete-categorical simulation. It uses $K = 6$, $\alpha = 0.2$, $B = 5000$, frequencies $p = (0.35, 0.25, 0.16, 0.11, 0.08, 0.05)$, and a local margin perturbation $\delta = 0.03$. Each target task is evaluated with $N_{\mathrm{MC}} = 8192$ complete mini-batches. The empirical Muon frequency exponent is 0.49397, while the corresponding SGD exponent is 1; the mean relative error against the auxiliary isotropic-closure coefficient is 3.25%. Because the exact categorical covariance is anisotropic, this agreement is reported as a stress test beyond the closure assumptions, not as theorem-level validation.

The full-categorical frequency-scaling result is the quantity plotted in Figure 3(b).

H.7 Muon–SignSGD sparse-support contrast

Figure 4 uses the exact two-task recursions from Appendices C.1 and F. We set $p_1 = 0.95$, $p_2 = 0.05$, $B = 1$, $\eta = 0.01$, $T = 3000$, and $\theta = 1.15$ radians, and run 64 independent paired mini-batch streams. The same task draw is fed to Muon and SignSGD at every step. For SignSGD, the positive answer-embedding scalar $\widetilde{c}$ in equation 143 is absorbed into the learning rate so that the experiment isolates the rotation-dependent coupling.

For Muon, a $B = 1$ batch updates only the sampled task, giving asymptotic rates $\sqrt{2}p_1$ and $\sqrt{2}p_2$ per unit optimization time. The rare-task margin therefore remains nonnegative and grows on its occurrence clock. For SignSGD, $\theta > \pi/4$ makes every task-1-only step decrease the task-2 margin, and the sufficient condition in Proposition 4.5 holds. We record both margins every 20 steps and plot the median and interquartile range across the 64 paired runs.

The final recorded medians at $\tau = \eta T = 30$ are approximately 40.28 and 2.15 for the frequent and rare Muon margins, versus 38.40 and -12.35 for SignSGD. These trajectories are the two panels of Figure 4.

H.8 Language-model pretraining comparison

We compare standard Muon with the fast–slow construction of Zhu et al. (2026) on a 0.12B-parameter dense language model trained for 30,000 steps. The remaining training settings are matched. The metric is validation loss, with one recorded trajectory per method. The export contains 140 checkpoints; the first two have no matching Muon value, leaving 138 paired checkpoints at 200-step intervals from step 2,600 to 30,000. We use these paired observations without interpolation or smoothing.

The fast–slow run has lower validation loss at every paired checkpoint. Final losses are 3.009736 for Muon and 2.988844 for fast–slow momentum, a difference of 0.020892. The final-ten-checkpoint means are 3.010568 and 2.990669, respectively. Figure 6 displays the 51 paired checkpoints from step 20,000 to 30,000. Checkpoints within a trajectory are not independent training replications; these observations do not establish variability across seeds or isolate the contribution of each algorithmic component.