REACHABILITY AND CONTROL OF DIRECTIONAL STEPS IN SCALE-INVARIANT SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

In scale-invariant models, the learning rate controls optimization only indirectly. Weight decay changes the parameter norm and thus also influences gradient updates in directions that subsequently affect prediction. We study this coupling through *directional forcing* B_t, which is the input that learning rate and decay jointly supply to the effective-step dynamics. This viewpoint turns schedule design into an inverse path-control problem, requiring a choice of a 'forcing path' and a characterization of how bounded weight decay can realize it. We characterize exactly the paths permitted by bounded decay multipliers, derive their optimal approximation, and solve exactly a noisy alignment problem whose optimal policy is shared by every finite horizon. We then establish a sharp stochastic classification of directional steps and prove almost-sure stationarity on the sphere under weak moment assumptions. The classification reveals that weight decay can change the decay exponent and summability class of effective steps even under a fixed harmonic learning-rate shape. The resulting design perspective has practical consequences. Across datasets, we find that a narrow forcing ramp improves mean accuracy in different tested settings, even over popular cosine learning rate setup. In a separate exact-symmetry experiment with signed decay, merely reversing an identical forcing multiset results in a massive accuracy drop of 30.34 percentage points. Together, the results set foundations for making directional motion a tractable object of schedule design.

1 INTRODUCTION

A normal learning rate schedule does not always measure how far a model moves in function space. Consider a parameter block whose sample loss is scale-invariant and does not change with positive rescaling. Its direction affects the loss, whereas its norm does not. Nevertheless, a larger norm reduces both the gradient magnitude and the directional change produced by an optimizer update, like that using Stochastic Gradient Descent (SGD). Weight decay changes this norm and therefore changes subsequent directional steps, even when its immediate radial action leaves the loss unchanged. This distinction underlies effective-rate analyses of scale-invariant networks (van Laarhoven, 2017; Arora et al., 2019; Wan et al., 2021; Lobacheva et al., 2021).

An exact recurrence makes the coupling between weight decay and learning rate explicit (Amin et al., 2026). Let Φ_t denote the effective directional-step coefficient, U_t the squared norm of the scale-free stochastic gradient, and $a_t = 1 - \eta_t\lambda_t > 0$ the radial multiplier. Then

$$\Phi_{t+1} = \frac{B_t\Phi_t}{1 + U_t\Phi_t^2}, \qquad B_t = \frac{\eta_{t+1}}{\eta_t a_t a_{t+1}}. \tag{1}$$

The schedule determines B_t, which we call *directional forcing*, while the denominator is the response of the norm to the gradient update.

This coupling matters when interpreting a schedule's effect on optimization, since conditions imposed on the normal learning rate need not describe the steps taken by the parameter direction. In stochastic approximation, divergent step sums permit continued progress, while summable squared steps control accumulated stochastic error. Whether these conditions hold must therefore be checked for the directional coefficients themselves. For example, $\eta_t = c/(t+\tau)$ satisfies $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. The coefficients of the corresponding directional updates need not satisfy the same conditions. We later show how constant decay can change their power-law exponent and move them

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.

across a square-summability boundary. Thus a convergence argument based only on the nominal learning rate can miss the dynamics of the scale-invariant block.

The above representation turns schedule design into a path-control problem. A preferred sequence of directional inputs cannot be chosen pointwise (since consecutive values share the same decay multipliers), while at the same time the cumulative path determines the long horizon scale of directional steps. This separates two specific questions: which forcing paths bounded weight decay can realize, and what directional behavior those paths produce. We address the first through an exact feasibility characterization, and the second through an optimal alignment benchmark and a stochastic classification. Our contributions are:

- **Feasible schedule design.** We characterize the entire set of forcing paths realizable at fixed learning rates and an initial radial multiplier (Theorem 1). Each multiplier appears in two adjacent transitions, so pointwise bounds are not sufficient. Our characterization gives unique synthesis and a global projection for infeasible requests.
- **An optimal reference path.** We solve a noisy spherical alignment benchmark with freely chosen positive directional steps (Theorem 2), and obtain one policy that is optimal at every finite horizon. Its inducing forcing crosses 1 at most once and approaches $1 - 1/t$, sustaining harmonic progress under the specified noise geometry.
- **Directional-step asymptotics.** We classify effective steps through the cumulative forcing product (Theorem 3). The classification identifies a square-summability boundary and yields almost-sure stationarity under bounded conditional second gradient moments. For harmonic learning rates, constant decay changes the directional-step exponent and produces the explicit threshold $c\lambda = 1/4$.
- **Finite-horizon empirical comparisons.** A common narrow forcing request improves mean performance across 14 model–dataset–budget settings, with significant performance improvement. A separate experiment with exact scale symmetry and signed decay compares a forcing path with its exact reverse, whose 30.34-point difference demonstrates extreme order dependence, at least in that specific setting.

2 Related work

Length–direction decoupling and effective-rate analyses establish how normalization changes optimization (Ioffe & Szegedy, 2015; Li & Arora, 2020; Li et al., 2020; Kohler et al., 2019; Chiley et al., 2019; Cho & Lee, 2017; Cai et al., 2019). Spherical and rotational descriptions study the resulting motion (Roburin et al., 2022; Kosson et al., 2024; Hoffer et al., 2018), including periodicity and distinct training regimes (Kodryan et al., 2022; Li et al., 2022). The forward recurrence in (1) and its finite-horizon stochastic drift decomposition are due to Amin et al. (2026). We use that law to study the inverse problem: which forcing paths are realizable under bounds on weight decay for a fixed learning-rate schedule and how infeasible paths can be approximated. We then derive an optimal reference path for a noisy spherical problem. Motivated by its forcing tail, we determine more generally how cumulative forcing determines long-run stochastic steps. Schedule search and empirical trajectory-invariance studies also expose learning-rate–decay coupling (Naganuma et al., 2026; Li et al., 2025; d'Angelo et al., 2024), and our trajectory equivalence follows algebraically from scale symmetry. Direct norm constraints and angular reparameterizations offer another route to directional control (Wen et al., 2026; Xie et al., 2026). AngularMuown makes an angular multiplier schedulable by optimizing normalized directions separately from radial magnitudes (Hübler et al., 2026). Our question concerns the coupling retained by native SGD: which directional inputs its existing learning rate and weight decay can realize, how to approximate an infeasible request, and what the realized tail implies for stochastic motion.

3 Directional dynamics under SGD

Separating norm and direction. Write $w_t = r_t u_t \neq 0$, where $r_t = \|w_t\|$ and $u_t \in \mathbb{S}^{D-1}$. Assume each differentiable sample loss satisfies $\ell_t(cw) = \ell_t(w)$ for $c > 0$. Differentiating gives $\langle w, \nabla \ell_t(w) \rangle = 0$ and $\nabla \ell_t(cw) = c^{-1} \nabla \ell_t(w)$. Thus the scale-free gradient $\bar{g}_t := r_t \nabla \ell_t(w_t) = \nabla \ell_t(u_t)$ is tangent to the sphere.

For momentum-free SGD with learning rate $\eta_t > 0$ and coupled weight-decay coefficient λ_t, define

$$w_{t+1} = a_t w_t - \eta_t \nabla \ell_t(w_t), \qquad a_t = 1 - \eta_t \lambda_t > 0, \qquad \Phi_t = \frac{\eta_t}{a_t r_t^2}, \quad U_t = \|\bar{g}_t\|^2. \tag{2}$$

We call U_t the *gradient energy*: it is the squared Euclidean norm of the scale-free gradient, not the objective value. The coefficient Φ_t sets the scale of the tangent update after factoring out $a_t r_t$. Nonnegative decay gives $a_t \leq 1$, while negative decay gives $a_t > 1$ and is used only where explicitly stated. Using $\bar{g}_t \perp u_t$, the update separates exactly into

$$u_{t+1} = \frac{u_t - \Phi_t \bar{g}_t}{\sqrt{1 + \Phi_t^2 U_t}}, \qquad r_{t+1}^2 = a_t^2 r_t^2 (1 + \Phi_t^2 U_t). \tag{3}$$

Although the loss is unchanged by rescaling, the evolving norm therefore changes the coefficient of subsequent directional updates.

Directional forcing. Substituting the norm update into $\Phi_{t+1} = \eta_{t+1}/(a_{t+1} r_{t+1}^2)$ yields (1), with $B_t = \eta_{t+1}/(\eta_t a_t a_{t+1})$. Its logarithmic form is

$$\log \Phi_{t+1} - \log \Phi_t = \log B_t - \log(1 + \Phi_t^2 U_t). \tag{4}$$

We call B_t the *directional forcing*. The term $\log B_t$ records the schedule contribution, while $\log(1 + \Phi_t^2 U_t)$ records the reduction caused by gradient-induced norm growth. When $B_t = 1$, only the latter contribution remains; $B_t \leq 1$ guarantees $\Phi_{t+1} \leq \Phi_t$. When $B_t > 1$, the coefficient increases only if $B_t > 1 + \Phi_t^2 U_t$. The turning angle also depends on gradient magnitude through $\theta_t = \arctan(\Phi_t \sqrt{U_t})$. Forcing therefore specifies an input to the directional dynamics, rather than the realized motion itself.[1]

Why temporal order matters. Two schedules with the same initial state (u_0, Φ_0), ordered forcing sequence, scale-free gradient rule, and sampled randomness generate identical (u_t, Φ_t) trajectories, even if their learning-rate and decay sequences differ. This follows by induction from (3) and (1); see Appendix A. Matching only the forcing product is insufficient. Although the product fixes the cumulative schedule contribution to $\log \Phi_t$, the gradient contribution depends on the intermediate coefficients. For two transitions with common gradient energy $U > 0$, define $T_b(h) = bh/(1 + Uh^2)$ for $h > 0$. Then

$$\frac{T_c(T_b(h))}{T_b(T_c(h))} = \frac{(1 + Uh^2)^2 + Uc^2 h^2}{(1 + Uh^2)^2 + Ub^2 h^2}. \tag{5}$$

If $b > c$, applying the larger forcing first produces a larger intermediate coefficient and a stronger subsequent reduction, leaving a smaller final coefficient despite the same product bc. In training, changing the order can also change the directions and subsequent gradient energies.

The ordered forcing path therefore determines the schedule input, while its interaction with gradients determines the response. We next characterize which requested paths SGD can realize under bounds on weight decay.

4 Feasible forcing sequences and schedule construction

Given a requested forcing sequence, can SGD produce it while respecting bounds on weight decay? We fix the learning-rate sequence and initial multiplier, then solve for the remaining decay coefficients. Because B_t depends on both a_t and a_{t+1}, the feasibility of consecutive requests must be checked jointly. We consider N forcing transitions. For T parameter updates, only $N = T - 1$ forcing values, $B_0, \ldots, B_{T-2}$, affect the directional endpoint u_T: Φ_t determines u_{t+1}, whereas B_t first affects u_{t+2}. The initial coefficient is specified separately through $\Phi_0 = \eta_0/(a_0 r_0^2)$.

With this convention, fix $\eta_0, \ldots, \eta_N > 0$, an admissible a_0, and finite positive bounds $\underline{a}_t \leq a_t \leq \overline{a}_t$. Write $x_t = \log a_t$, $b_t = \log B_t$, $\nu_t = \log \eta_t$, and $[L_t, H_t] = [\log \underline{a}_t, \log \overline{a}_t]$. Equation (1) requires

$$x_{t+1} = \nu_{t+1} - \nu_t - b_t - x_t. \tag{6}$$

Each multiplier participates in two adjacent transitions. Solving for the next one reverses the sign of the previous one, so the realized multipliers depend on alternating sums of the requested log-forcing values. The resulting constraint concerns the entire ordered path.

[1] These identities use exact tangency and do not require small-step approximation. Momentum, adaptive preconditioning, and departures from scale invariance introduce additional state or radial terms (see Appendix E for discussion on these extensions.

Theorem 1 (Exact feasibility and unique synthesis)**.** *A requested $b \in \mathbb{R}^N$ is feasible if and only if, for $1 \le t \le N$,*

$$L_t \le X_t(b) := (-1)^t x_0 + \sum_{k=0}^{t-1} (-1)^{t-1-k}(\nu_{k+1} - \nu_k - b_k) \le H_t. \tag{7}$$

Its realization is unique: $a_t = \exp X_t(b)$. The feasible set $\mathcal{P}$ is an affine image of $\prod_{t=1}^{N}[L_t, H_t]$.

Since the learning rates, initial multiplier, and requested forcing are prescribed, this is a deterministic feasibility test. For an infeasible target $\hat{b}$, choose positive weights ω_t and solve

$$\min_{b,x} \ \frac{1}{2}\sum_{t=0}^{N-1} \omega_t (b_t - \hat{b}_t)^2 \quad \text{s.t.} \quad b_t + x_t + x_{t+1} = \nu_{t+1} - \nu_t, \quad L_t \le x_t \le H_t, \tag{8}$$

with x_0 fixed. The problem yields a strictly convex, box-constrained quadratic with a tridiagonal Hessian. Thus every requested path has a unique closest feasible approximation under the chosen weighted log-error. The projection distributes approximation error across the schedule, accounting for the multiplier shared by adjacent transitions. Here, we provide one example: For constant learning rate, $a_0 = 1$, and $a_1, a_2 \in [1/2, 1]$, the request $(B_0, B_1) = (2, 1)$ requires $(a_1, a_2) = (1/2, 2)$ and is therefore infeasible. Sequential clipping realizes $(2, 2)$, whereas the global projection realizes $(\sqrt{2}, \sqrt{2})$ with equal projection weights, halving the squared log-error (Figure 1). Therefore, for the global construction, convert decay bounds to bounds on $x_t = \log(1 - \eta_t \lambda_t)$, test feasibility, and project infeasible requests using Equation (8). Recover the decay coefficients as $\lambda_t = (1 - e^{x_t})/\eta_t$. (Appendix B gives the proof, evaluation costs, and numerical optimality residual)

5 An optimal alignment benchmark

Section 4 determines which forcing paths bounded decay can realize. To obtain an optimal reference, we now temporarily remove these constraints and solve a noisy alignment problem over positive directional steps. We then discuss whether SGD can realize the resulting path.

To obtain a reference forcing path with a known optimization objective, we study alignment with a target direction on the sphere. We choose a noise geometry that preserves stochastic directional motion while making the alignment dynamics exactly solvable. Let $D \ge 3$, $u_* \in \mathbb{S}^{D-1}$, and $f(u) = -\langle u, u_* \rangle$. Assume $q_0 = \langle u_0, u_* \rangle \in (0, 1)$. Write $q_t = \langle u_t, u_* \rangle$ and $S_t = 1 - q_t^2$. Before each update, $\mathcal{F}_t$ contains the past and the current state. Use $\bar{g}_t = q_t u_t - u_* + \xi_t$, where $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$, $\|\xi_t\|^2 = v > 0$, and $\xi_t \perp \operatorname{span}\{u_t, u_*\}$. The noise is orthogonal to both the current and target directions. For a freely selected step $\phi > 0$,

$$q_{t+1} = F_v(q_t, \phi) := \frac{q_t + \phi(1 - q_t^2)}{\sqrt{1 + \phi^2(1 - q_t^2 + v)}}. \tag{9}$$

The map F_v gives the next alignment for every noise draw, not only in expectation.

Theorem 2 (Alignment-optimal steps)**.** *For every finite positive horizon, the unique optimal actions along the visited states of* (9)*, maximizing terminal alignment over unrestricted positive steps, are*

$$\phi_t^* = \frac{S_t}{q_t(S_t + v)}, \qquad S_t = \frac{v}{v/S_0 + t}. \tag{10}$$

Their inducing forcing is

$$B_t^* = \frac{q_{t+1}}{q_t}\frac{S_t + v}{2S_t + v} = 1 - \frac{1}{t} + O(t^{-2}). \tag{11}$$

Let $S_c = [3 - 3v + \sqrt{(3v-3)^2 + 32v}]/8$. Then $S_c \in (0, 1)$ and $B_t^* > 1$ exactly when $S_t > S_c$. Thus the forcing crosses one at most once, from above to below.

The derivative of F_v in ϕ has numerator $S - q\phi(S + v)$, so (10) uniquely maximizes the next alignment. Crucially, the best attainable next alignment, $\max_\phi F_v(q, \phi) = \sqrt{1 - v(1 - q^2)/(1 - q^2 + v)}$, is

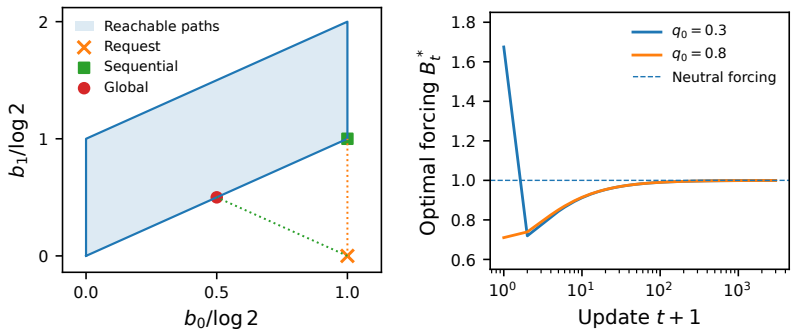

Figure 1: **Native feasibility and directional optimality.** Left: feasible log-forcing pairs for $a_0 = 1$, $a_1, a_2 \in [1/2, 1]$, and constant learning rate. Global projection shares the approximation error across time, whereas sequential clipping matches the first request at the expense of the second. Right: exact optimal forcing from Theorem 2 with $v = 0.2$. Initial alignment determines whether an expansionary prefix is present; both paths approach one from below. The curves are analytical.

strictly increasing in q. The optimal error obeys $S_{t+1}^{-1} = S_t^{-1} + v^{-1}$, which yields (10). Appendix C supplies the derivatives, induction, and forcing algebra.

The optimal error and directional-step coefficient both decay harmonically. Initial alignment determines whether an expansionary prefix occurs, while the tail approaches neutral forcing from below. This one-crossing statement does not assert that the entire forcing path is monotone. A fixed $B < 1$ would instead give $\Phi_t \le \Phi_0 B^t$ and a finite total coefficient sum. The rate at which forcing approaches one is therefore essential to sustained motion.

To implement this reference with native SGD, both $\Phi_0 = \phi_0^*$ and Theorem 1 must hold. Constant learning rate with nonnegative decay forces $B_t \ge 1$ and cannot realize the eventual contracting tail. Projection supplies a feasible approximation in forcing space, without preserving the unconstrained alignment optimum. The next section examines what this tail behavior implies under more general stochastic gradients.

6 How forcing controls long-run directional steps

Section 5 gives an optimal reference whose forcing approaches one at rate $1/t$. We now examine how such forcing affects directional steps under more general stochastic gradients. The central question is whether the cumulative forcing determines the step decay, or whether gradient-induced norm growth changes that rate.

From forcing to effective steps. For a deterministic positive forcing sequence, define $Q_0 = 1$ and $Q_t = \prod_{j<t} B_j$. Without gradients, the recurrence gives $\Phi_t = \Phi_0 Q_t$ exactly, so Q_t describes the cumulative schedule contribution. We consider paths satisfying

$$Q_t \sim K t^{-\beta}, \qquad K \in (0, \infty), \quad \beta \in \mathbb{R}, \tag{12}$$

where $\sim$ means that the ratio tends to one. For example, with $\tau > 0$,

$$B_t = \left(\frac{t+\tau}{t+\tau+1}\right)^{\beta} \quad \Longrightarrow \quad Q_t = \left(\frac{\tau}{t+\tau}\right)^{\beta}.$$

More generally, summing logarithms shows that

$$B_t = 1 - \frac{\beta}{t+\tau} + O(t^{-2}), \qquad \tau > 0, \tag{13}$$

is sufficient for (12). Nonzero gradients introduce the denominator $1 + \Phi_t^2 U_t$ in the effective-step recurrence. To separate their contribution from forcing, set $Y_t = Q_t/\Phi_t$ and $k_t = Q_t^2$. Then

$$Y_{t+1} = Y_t + \frac{k_t U_t}{Y_t}, \qquad Y_{t+1}^2 - Y_t^2 = 2k_t U_t + \frac{k_t^2 U_t^2}{Y_t^2}. \tag{14}$$

Equation (14) shows how gradient energy accumulates in Y_t^2: its leading increment is $2Q_t^2 U_t$, with a nonnegative remainder. The relevant deterministic series is therefore $\sum_t Q_t^2$. Since $Q_t^2 \sim K^2 t^{-2\beta}$, this series is finite precisely when $\beta > 1/2$. Under the assumptions below, Y_t then approaches a finite positive limit, so $\Phi_t = Q_t/Y_t$ retains the forcing exponent. When $\beta \le 1/2$ and the conditional gradient energy approaches a positive limit, the accumulated contribution instead satisfies $Y_t^2 \sim 2v \sum_{j<t} Q_j^2$, which changes the effective-step rate.

Stochastic rates and stationarity. Let $\mathcal{F}_t$ contain the history and current state before sampling the gradient at step t. Then Φ_t is $\mathcal{F}_t$-measurable, while U_t is the realized squared norm of the scale-free gradient. The quantity $\mathbb{E}[U_t \mid \mathcal{F}_t]$ is its conditional mean; a bound on $\mathbb{E}[U_t^p \mid \mathcal{F}_t]$ controls larger fluctuations. The following conclusions hold almost surely, meaning with probability one over the gradient draws.

Theorem 3 (Step asymptotics and stationarity)**.** *Let $\Phi_0 > 0$ be fixed, assume* (1) *and* (12)*, and take all moment bounds below to hold almost surely at every time with finite deterministic constants.*

1. *If $\beta > 1/2$ and $\mathbb{E}[U_t \mid \mathcal{F}_t] \le M$, then $\Phi_t \sim Ct^{-\beta}$ almost surely, where C is finite, strictly positive, and possibly random.*
2. *If $\beta \le 1/2$, $\mathbb{E}[U_t \mid \mathcal{F}_t] \to v > 0$ almost surely for deterministic v, and $\mathbb{E}[U_t^p \mid \mathcal{F}_t] \le M_p$ for some $1 < p \le 2$, then almost surely*
$$\Phi_t \sim \begin{cases} (2vt\log t)^{-1/2}, & \beta = 1/2, \\ \sqrt{(1-2\beta)/(2v)}\, t^{-1/2}, & \beta < 1/2. \end{cases} \tag{15}$$
3. *Suppose also that* (3) *holds, $U_t = \|\bar{g}_t\|^2$, $\bar{g}_t \perp u_t$, $f \in C^2(\mathbb{S}^{D-1})$, $\mathbb{E}[\bar{g}_t \mid \mathcal{F}_t] = \operatorname{grad} f(u_t)$, and $\mathbb{E}[\|\bar{g}_t\|^2 \mid \mathcal{F}_t] \le M$. If $1/2 < \beta \le 1$, then almost surely*
$$\sum_t \Phi_t = \infty, \qquad \sum_t \Phi_t^2 < \infty, \qquad \|\operatorname{grad} f(u_t)\| \longrightarrow 0. \tag{16}$$

Part (iii) requires only a bounded conditional second gradient moment, without the positive limiting energy of part (ii). In part (ii), choosing $1 < p < 2$ permits infinite fourth gradient moments. Appendix D provides the proofs. The first two parts distinguish three regimes. Above $\beta = 1/2$, the effective step retains the forcing exponent, although its leading constant can depend on the gradient history. Below $\beta = 1/2$, persistent gradient energy instead determines a $t^{-1/2}$ rate. At $\beta = 1/2$, the series $\sum_{j<t} Q_j^2$ grows logarithmically, producing the additional $(\log t)^{-1/2}$ factor.

Part (iii) connects the rates to optimization. The divergent sum $\sum_t \Phi_t$ permits continued progress, while square-summability, $\sum_t \Phi_t^2 < \infty$, controls accumulated second-order errors in the stochastic descent argument (Robbins & Monro, 1951; Bonnabel, 2013). For $\Phi_t \sim Ct^{-\beta}$, both conditions hold when $1/2 < \beta \le 1$. By contrast, $\beta > 1$ makes the step sum finite, and the boundary rate in part (ii) gives $\Phi_t^2 \sim 1/(2vt\log t)$, whose sum diverges. Under the additional assumptions in part (iii), the gradient norm tends to zero and establish stationarity.

Harmonic learning rates with constant decay. The forcing exponent can be computed directly from ordinary learning-rate and decay schedules. The following consequence shows that changing weight decay can change the directional-step regime even when the nominal learning-rate shape remains fixed.

Corollary 1 (Harmonic learning rate and constant decay)**.** *Let $\eta_t = c/(t+\tau)$, $\lambda_t = \lambda \ge 0$, $c > 0$, and $\tau > c\lambda$. Then* (12) *holds with $\beta = 1 - 2c\lambda$. Under the respective assumptions of Theorem 3,*
$$\Phi_t \sim \begin{cases} Ct^{-(1-2c\lambda)}, & c\lambda < 1/4, \\ (2vt\log t)^{-1/2}, & c\lambda = 1/4, \\ \sqrt{(4c\lambda-1)/(2v)}\, t^{-1/2}, & c\lambda > 1/4. \end{cases} \tag{17}$$
Stationarity follows for $c\lambda < 1/4$ under part (iii)'s assumptions.

To identify the exponent, put $z = t + \tau$ and $d = c\lambda$. The forcing factor is
$$B_t = \frac{z^2}{(z-d)(z+1-d)} = 1 - \frac{1-2d}{z} + O(z^{-2}).$$

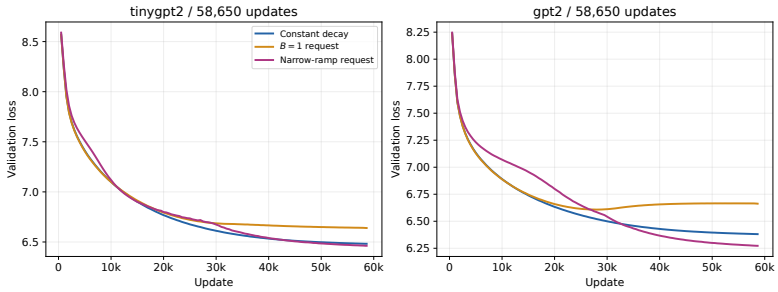

Figure 2: **Validation trajectories on WikiText-103.** Curves show mean validation loss over three seeds at an initial learning rate of 0.005. The narrow ramp initially reduces loss more slowly but reaches a lower final mean than both reference schedules.

Thus zero decay gives exponent one, whereas $c\lambda = 0.2$ gives exponent 0.6. Both use a harmonic nominal learning rate, but their effective directional steps decay differently. The threshold $c\lambda = 1/4$ marks the transition between preserving this forcing exponent and the rates determined by persistent gradient energy. We next examine forcing-based schedules at finite training horizons and test the predicted stochastic regimes in controlled experiments.

7 Experiments

We first test whether forcing-based schedule design improves conventional training, then examine the mechanisms identified by the theory. All comparisons use momentum-free SGD and paired runs with matched initialization and sampling. Full experiments details are in Appendix F.

Accuracy gains at a fixed learning-rate schedule. We study ResNet-18/ResNet-34/ResNet-50 and Tiny GPT-2/GPT-2 with conventional training on CIFAR-10/CIFAR-100 and WikiText-103. Here, we compare among three different choices of schedules. First, the original cosine path with weight decay fixed. Second, for an exactly scale-invariant block, realized forcing $B_t = 1$ removes the schedule term from the effective-step recurrence. Finally, a decreasing forcing ramp is designed to offset this decrease early in training and strengthen it later. These comparisons test whether explicitly shaping the directional coefficient improves performance relative to the reference and neutral-forcing schedules. Learning rates (CIFAR: 0.5 to 0.05 and WikiText: 0.005 to 0.0005) follows the same cosine path in every arm to isolate the effect of decay scheduling. The common narrow ramp, $\delta = 0.002$, improves mean accuracy and validation loss drop in all 14 model/dataset settings (Table 1). Gains range from 0.37 to 6.83 percentage for vision models and 0.02 to 0.108 for language models. (See detailed discussion regarding reachability in F.5 and F.4)

How much forcing intervention is needed? We next investigate how the degree of forcing intervention relates to performance. We measure intervention relative to neutral forcing ($B_t = 1$) by the cumulative absolute log-forcing, $\mathcal{A}_N = \sum_{t=0}^{N-1} |\log B_t|$, over the N operative forcing transitions. This quantity measures the magnitude of schedule contributions without cancellation between expansionary and contractive phases.

We conduct a separate ResNet-50/CIFAR-100 experiment with 58,650 updates and five paired seeds. Schedule changes are applied only to convolutional weights, while complementary parameters retain baseline decay. This scope was selected through comparison of whole-model, convolution-only, and complement-only interventions, reported in Appendix F.6. The following comparisons are therefore exploratory and conditional on that selection.

Within this fixed scope, we construct two lower-action requests. The first is inspired by the alignment benchmark's $\beta = 1$ tail. The second controls cumulative forcing across training horizons: extending a fixed forcing range over more updates increases its total action, so we instead scale the log-forcing amplitude inversely with the horizon. For a continuous profile $\zeta : [0, 1] \to \mathbb{R}$, set $B_t^{(N)} =$

Table 1: **A common narrow ramp improves mean accuracy/validation loss in all 14 settings.** Final test accuracy (higher is better) for image task and validation loss (lower is better) for language tasks, mean±SD over five seeds for CIFAR and three for WikiText. Learning rates, optimizer, and training budget are fixed within each row, and only decay scheduling changes. The narrow request is $1.002 \to 0.998$. Δ denotes narrow ramp minus constant decay, measured in percentage points for CIFAR accuracy and in cross-entropy units for WikiText loss.

Dataset	Model	Updates	$B = 1$	Constant decay	Narrow ramp	Δ
CIFAR-10	R18	10000	88.64 ± 1.86	92.03 ± 0.38	92.60 ± 0.38	$+0.57$
CIFAR-10	R34	10000	88.14 ± 0.95	91.36 ± 0.49	92.47 ± 0.21	$+1.12$
CIFAR-10	R50	10000	86.15 ± 1.81	90.16 ± 0.64	91.09 ± 0.67	$+0.93$
CIFAR-100	R18	10000	69.31 ± 0.42	72.02 ± 0.47	72.75 ± 0.44	$+0.73$
CIFAR-100	R34	10000	65.34 ± 1.63	70.29 ± 1.06	71.52 ± 0.89	$+1.23$
CIFAR-100	R50	10000	60.48 ± 2.16	67.36 ± 1.32	67.73 ± 2.34	$+0.37$
CIFAR-10	R18	58650	92.98 ± 0.88	93.16 ± 0.50	95.10 ± 0.19	$+1.94$
CIFAR-10	R34	58650	92.95 ± 0.39	92.78 ± 0.39	95.37 ± 0.11	$+2.59$
CIFAR-10	R50	58650	91.19 ± 1.05	92.32 ± 0.56	94.80 ± 0.38	$+2.48$
CIFAR-100	R18	58650	73.65 ± 0.47	70.34 ± 1.11	77.17 ± 0.30	$+6.83$
CIFAR-100	R34	58650	72.86 ± 0.35	72.07 ± 0.67	77.26 ± 0.45	$+5.18$
CIFAR-100	R50	58650	70.68 ± 2.64	71.44 ± 0.88	76.29 ± 1.23	$+4.85$
WikiText-103	Tiny GPT-2	58650	6.639 ± 0.017	6.482 ± 0.026	6.462 ± 0.008	-0.02
WikiText-103	GPT-2	58650	6.663 ± 0.016	6.381 ± 0.001	6.273 ± 0.018	-0.108

Table 2: **Different action schedules in the selected convolutional scope.** ResNet-50/CIFAR-100 with 58,650 updates and five runs. We record accuracy, $\sum |\log B_t|$ and $\sum \eta_t \lambda_t$ for comparison of intervention across different schedules and its performance. (The normalized and benchmark-tail requests are globally projected)

Schedule	Final accuracy (%)	$\sum \lvert \log B_t \rvert$	$\sum \eta_t \lambda_t$
Constant decay	70.43 ± 2.21	29.96	16.13
Narrow ramp	77.01 ± 0.76	31.00	14.95
Benchmark-tail request	75.87 ± 0.62	6.18	3.56
Horizon-normalized request	76.58 ± 0.69	6.60	2.85

$\exp(\zeta(t/N)/N)$. Then

$$\sum_{t=0}^{N-1} \log B_t^{(N)} \longrightarrow \int_0^1 \zeta(s)\, ds, \qquad \sum_{t=0}^{N-1} |\log B_t^{(N)}| \longrightarrow \int_0^1 |\zeta(s)|\, ds. \tag{18}$$

The profile therefore controls cumulative forcing while the horizon determines how it is distributed across updates. We use a decreasing linear profile calibrated to the earlier short-horizon ramp. Both requests are projected onto the native feasible set using the full ordered path, and we report their realized forcing action.

The ramp reaches 77.01%, compared with 76.58% for the horizon-normalized schedule and 75.87% for the benchmark-tail schedule. Their absolute log-forcing values are $31.00, 6.60$, and 6.18, respectively(Table 2). These results suggest a tradeoff between forcing intervention and predictive performance. Within the selected convolutional scope, the higher-action ramp achieves the highest mean accuracy, while the lower-action schedules achieve lower means with less cumulative forcing.

Effect of reversing the forcing sequence. The preceding comparisons change both schedule shape and total action. To compare temporal order while matching the forcing values, we use a ResNet-18 with scale-invariant block implemented by affine-free BatchNorm and the forcing schedules act on convolutional weights and complementary parameters retain constant decay. On CIFAR-10, we synthesize a forcing ramp $1.003 \to 0.997$ and its exact reverse $0.997 \to 1.003$ over 10,000 updates. The results are shown in Table 3. The mean difference is 30.34 percentage points over eight pairs.

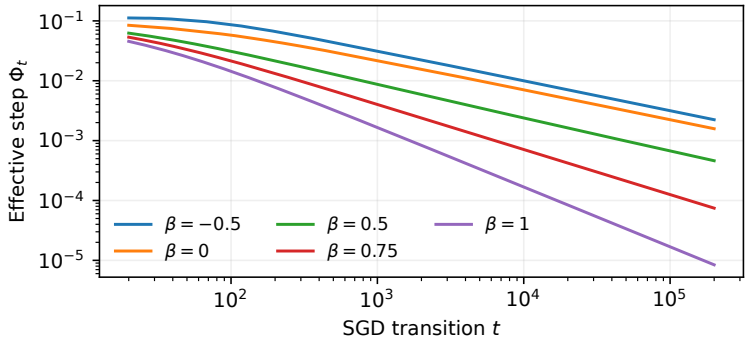

Figure 3: **Forcing determines the step exponent until gradient energy takes over.** Median effective steps over 64 replicates of the exact radial recurrence, with a common energy law and $B_t = ((t+20)/(t+21))^\beta$. At $\beta = 1/2$, the logarithmic correction separates the boundary from a pure $t^{-1/2}$ power law. For $\beta > 1/2$, leading constants depend on the sample path. The prediction holds despite persistent energy fluctuations.

Their different outcomes show why a total forcing budget cannot describe an optimization trajectory and the effect of order dependence established in Section 3. The complete protocol and results appear in Appendix F.2.

Table 3: **Reversing a matched forcing budget.** Mean±SD over eight paired runs.

Forcing order	Final accuracy (%)	Difference (pp)
Forward, $1.003 \to 0.997$	91.23 ± 0.36	$\mathbf{+30.34}$
Reverse, $0.997 \to 1.003$	60.89 ± 4.46	—

The predicted stochastic regimes. Finally, we test Theorem 3 using an unbiased tangent oracle on $\mathbb{S}^2$ whose gradient energy fluctuates indefinitely under synthetic setup. Across five forcing exponents, 200,000 transitions, and 64 replicates, Figure 3 shows the transition from forcing-controlled $t^{-\beta}$ decay to energy-controlled $t^{-1/2}$ decay, including the logarithmic correction at the boundary. The fitted log–log slopes over the final decade of simulated time are -0.49965, -0.49940, -0.54710, -0.75177, and -0.99969 for $\beta = -0.5, 0, 0.5, 0.75, 1$, respectively. These measurements agree closely with the predicted exponent $-1/2$ below the boundary and $-\beta$ above it. At $\beta = 1/2$, the more negative finite-horizon slope is consistent with the logarithmic correction in $(t \log t)^{-1/2}$. Appendix F.1 compares the trajectories with the predicted leading constants, including the logarithmic correction.

We also sample the squared gradient norm U_t from a Pareto distribution with tail index $3/2$, giving finite second but infinite fourth gradient moments. At the final time, median ratios to the predicted leading asymptotics are within 4% of one for all three energy-controlled exponents. The gradient norm also decreases on a nonconstant alignment objective. The experiments agree with the predicted asymptotics and illustrate stationarity under heavy-tailed noise.

Discussion and conclusion. Learning rate and weight decay jointly shape directional steps through the evolving parameter norm. Our feasibility characterization identifies which forcing paths bounded decay can realize, the alignment benchmark provides an optimal reference under specified noise geometry, and the stochastic analysis connects cumulative forcing to effective-step rates and stationarity. Experiments show benefits from a common narrow ramp in the tested settings and demonstrate that temporal order matters even when the forcing multiset is fixed. The exact theory applies to scale-invariant blocks, while conventional-network experiments add to validating its usefulness beyond strict assumptions.

REPRODUCIBILITY STATEMENT

Detailed proofs are provided in appendices. Detailed experimental setup is shared in Appendix F. Supplementary material includes code that implements the training procedures, schedule construction, and synthetic experiments. Datasets are downloaded when needed, and results and diagnostic histories are generated during execution.

AI USE STATEMENT

AI tools assisted with literature retrieval and writing refinement. They also assisted with the implementation of synthetic stochastic experiments and numerical validation. The authors remain responsible for the correctness, originality, and reproducibility of the submitted work.

REFERENCES

Hasan Amin, Wei-Kai Chang, and Rajiv Khanna. When does scale-invariant optimization become unstable? an exact schedule law with weight decay, 2026. URL https://arxiv.org/abs/2609.09116.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkxQ-nA9FX.

Silvëre Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619. URL https://ieeexplore.ieee.org/abstract/document/6487381?casa_token=iO4VpP9nPe4AAAAA:M_U3o4FBZRb3NCjQAQhT9QHBAC8xuvVQwKEgZey92Mjc9wE57P5akOkLzt86JVrZCYTtCrL1Mg.

Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. In *International Conference on Machine Learning*, pp. 882–890. PMLR, 2019. URL https://proceedings.mlr.press/v97/cai19a.html.

Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cb3ce9b06932da6faaa7fc70d5b5d2f4-Abstract.html.

Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3a0844cee4fcf57de0c71e9ad3035478-Abstract.html.

Francesco d'Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *Advances in Neural Information Processing Systems*, 37:23191–23223, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29496c942ed6e08ecc469f4521ebfff0-Paper-Conference.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a0160709701140704575d499c997b6ca-Abstract.html.

Florian Hübler, Kai Lion, Antonio Orvieto, and Niao He. Muown implicitly performs angular step-size decay. arXiv preprint arXiv:2606.23637, 2026. URL https://arxiv.org/abs/2606.23637.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry P Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. *Advances in Neural Information Processing Systems*, 35:14058–14070, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5aea56eefab60e06f35016478e21aae6-Abstract-Conference.html.

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 806–815. PMLR, 2019. URL https://proceedings.mlr.press/v89/kohler19a/kohler19a.pdf.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25333–25369. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kosson24a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Bingrui Li, Jiaxin Wen, Zhanpeng Zhou, Jun Zhu, and Jianfei Chen. Efficient hyperparameter tuning via trajectory invariance principle. *arXiv preprint arXiv:2509.25049*, 2025. URL https://arxiv.org/abs/2509.25049.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJg8TeSFDH.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/a7453a5f026fb6831d68bdc9cb0edcae-Abstract.html.

Zhiyuan Li, Tianhao Wang, and Dingli Yu. Fast mixing of stochastic gradient descent with normalization and weight decay. *Advances in Neural Information Processing Systems*, 35:9233–9248, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3c215225324f9988858602dc92219615-Abstract-Conference.html.

Ekaterina Lobacheva, Maxim Kodryan, Nadezhda Chirkova, Andrey Malinin, and Dmitry P Vetrov. On the periodic behavior of neural network training with batch normalization and weight decay. *Advances in Neural Information Processing Systems*, 34:21545–21556, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/b433da1b32b5ca96c0ba7fcb9edba97d-Abstract.html.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL https://arxiv.org/abs/1608.03983.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Hiroki Naganuma, Atish Agarwala, Priya Kasimbeg, and George E. Dahl. What do near-optimal learning rate schedules look like? *arXiv preprint arXiv:2603.10301*, 2026. URL https://arxiv.org/abs/2603.10301.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. URL https://www.jstor.org/stable/2236626?casa_token=0NwRMOLZjIQAAAAA%3AeSlrWENYjHT2II6B2snHdL9E6GnrxyirE2-7LZd-n4o_FIynLAaQQqJe-85HAjL0pVX4wVSOvcIUxq4FEideI6qJxVHCl6QEUqMTlNLrlGheY8ekYQI8&seq=1.

Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Perez, and Mathieu Aubry. Spherical perspective on learning with normalization layers. *Neurocomputing*, 487:66–74, 2022. URL https://www.sciencedirect.com/science/article/pii/S092523122200159X?casa_token=Awwy7tkb5NUAAAAA:_E6oH3xKHTnVVu0nx3n0drhWJWk94VuvEar66ZAnTfbG4jA3qv944dzQFjA9nBY-mpILn6gXxw.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017. URL https://arxiv.org/abs/1706.05350.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using SGD and weight decay. *Advances in Neural Information Processing Systems*, 34:6380–6391, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/326a8c055c0d04f5b06544665d8bb3ea-Abstract.html.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization. *arXiv preprint arXiv:2606.16899*, 2026. URL https://arxiv.org/abs/2606.16899.

Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, Chong Luo, and Baining Guo. Controlled LLM training on spectral sphere. *arXiv preprint arXiv:2601.08393*, 2026. URL https://arxiv.org/abs/2601.08393.

A EXACT QUOTIENT DYNAMICS AND SCHEDULE EQUIVALENCE

We derive the quotient state equation, establish the conditions for pathwise schedule equivalence, and then verify the order dependence used in the main text. Each argument is an exact discrete-time identity.

A.1 SCALE INVARIANCE AND RADIAL DECOMPOSITION

Fix a differentiable loss $\ell : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}$ satisfying $\ell(cw) = \ell(w)$ for every $c > 0$. Differentiating in c at $c = 1$ gives $\langle w, \nabla\ell(w)\rangle = 0$. Differentiating in w gives $\nabla\ell(cw) = c^{-1}\nabla\ell(w)$. Thus, for $w = ru$, the field $\bar{g}(u) := r\nabla\ell(ru) = \nabla\ell(u)$ is independent of r and tangent to the unit sphere. These identities hold for each realized sample loss, before any averaging over stochastic gradients.

Factoring $a_t r_t$ from (2) yields

$$w_{t+1} = a_t r_t(u_t - \Phi_t \bar{g}_t), \qquad \Phi_t = \frac{\eta_t}{a_t r_t^2}. \tag{19}$$

Orthogonality gives $\|u_t - \Phi_t\bar{g}_t\|^2 = 1 + \Phi_t^2 U_t$, proving (3) and ensuring that the next iterate is nonzero. Substitution into the next effective step gives

$$\Phi_{t+1} = \frac{\eta_{t+1}}{a_{t+1}a_t^2 r_t^2(1 + \Phi_t^2 U_t)} \tag{20}$$

$$= \frac{\eta_{t+1}}{\eta_t a_t a_{t+1}} \frac{\Phi_t}{1 + \Phi_t^2 U_t}. \tag{21}$$

Moreover, $\langle u_{t+1}, u_t\rangle = (1 + \Phi_t^2 U_t)^{-1/2} > 0$. The turning angle lies in $[0, \pi/2)$ and equals $\arctan(\Phi_t\sqrt{U_t})$, separating the directional coefficient from the realized angular magnitude. Suppose the two schedules generate the same (u_t, Φ_t). Under the same scale-free gradient rule and sampled randomness, their gradients coincide. Equations Equation (1) and Equation (3), together with the common B_t, then give identical (u_{t+1}, Φ_{t+1}). Induction from the common initial state proves trajectory equivalence.

A.2 TWO-TRANSITION ORDER DEPENDENCE

For $T_b(h) = bh/(1 + Uh^2)$, direct substitution gives

$$T_c(T_b(h)) = \frac{bc\,h(1 + Uh^2)}{(1 + Uh^2)^2 + Ub^2h^2}. \tag{22}$$

Exchanging b and c and taking the ratio proves (5). For $U, h > 0$, the ratio is strictly below one exactly when $b > c$. This establishes noncommutativity along a common energy path. The calculation compares effective steps, rather than training losses, and makes no assertion that different controlled trajectories retain the same gradient energy.

B REACHABILITY AND GLOBAL SYNTHESIS

B.1 PROOF OF THEOREM 1

Taking logarithms of $B_t = \eta_{t+1}/(\eta_t a_t a_{t+1})$ gives (6). Recursively substituting for x_t yields

$$x_t = (-1)^t x_0 + \sum_{k=0}^{t-1} (-1)^{t-1-k}(\nu_{k+1} - \nu_k - b_k) = X_t(b). \tag{23}$$

With x_0 fixed, the recurrence uniquely determines every subsequent actuator. If all $X_t(b)$ lie in their allowed intervals, exponentiation gives an admissible realization of the requested B_t. If any interval constraint fails, no alternative realization exists.

Conversely, each $x_1, \ldots, x_N$ in the actuator box determines

$$b_t = \nu_{t+1} - \nu_t - x_t - x_{t+1}. \tag{24}$$

The linear part of this map from $(x_1, \dots, x_N)$ to $(b_0, \dots, b_{N-1})$ is triangular with diagonal entries -1, hence invertible. The reachable set $\mathcal{P}$ is therefore precisely the affine image of the box. It is nonempty, compact, and convex, and has the same dimension as the box. If every interval has positive length, $\mathcal{P}$ is a full-dimensional parallelotope.

□

B.2 Global projection and its optimality certificate

Let $z = (x_1, \dots, x_N)^\top$, and let A be the lower-bidiagonal matrix with diagonal and subdiagonal entries one. With x_0 fixed, (8) becomes

$$\min_{L_{1:N} \le z \le H_{1:N}} \frac{1}{2} \|W^{1/2}(Az - r)\|^2, \qquad W = \operatorname{diag}(\omega_0, \dots, \omega_{N-1}), \tag{25}$$

where $r_0 = \nu_1 - \nu_0 - \widehat{b}_0 - x_0$ and $r_t = \nu_{t+1} - \nu_t - \widehat{b}_t$ for $t \ge 1$. Since A is nonsingular and $W \succ 0$, the Hessian $A^\top W A$ is positive definite and tridiagonal. Compactness gives existence of a minimizer, and strict convexity gives uniqueness. The bijection (24) then yields unique projected forcing and actuator paths. Intervals of zero width simply fix coordinates and preserve this conclusion.

An objective and gradient evaluation uses $O(N)$ arithmetic and storage. This structure alone does not give a horizon-independent iteration bound for a generic solver, since the smallest eigenvalue can deteriorate with horizon. The implementation retains the projected-gradient residual

$$\left\| z - \Pi_{[L,H]}(z - \nabla J(z)) \right\|_\infty, \tag{26}$$

where J is the objective in (25). This residual vanishes exactly at the minimizer of a differentiable convex box problem and provides a numerical optimality diagnostic for each new projection audit.

B.3 Reachable reversals and matched budgets

Let $\mathcal{R}$ reverse a length-N vector. The pair $(b, \mathcal{R}b)$ is jointly reachable exactly when $b \in \mathcal{P}$ and $\mathcal{R}b \in \mathcal{P}$. Since $\mathcal{R}^2 = I$, this is equivalent to $b \in \mathcal{P} \cap \mathcal{R}\mathcal{P}$. Reversal preserves the multiset of b_t, hence also that of $B_t = e^{b_t}$, the signed action $\sum b_t$, and the absolute action $\sum |b_t|$.

The actuator multiset and native dose $\sum \eta_t \lambda_t = \sum (1 - a_t)$ need not be preserved. Once η, a_0, and the forcing are fixed, synthesis uniquely determines each a_t and $\lambda_t = (1 - a_t)/\eta_t$. Independently matching native dose requires additional design freedom or a different intervention. The exact-mirror experiment accordingly compares forcing order at matched forcing budgets, with the induced native doses reported separately.

C The optimal alignment benchmark

C.1 An exact stochastic realization

Let $D \ge 3$, $u_* \in \mathbb{S}^{D-1}$, and $f(u) = -\langle u, u_* \rangle$. At a state with $q = \langle u, u_* \rangle \in (0, 1)$, the spherical gradient is $g = qu - u_*$, with $\|g\|^2 = 1 - q^2 =: S$. The subspace $\operatorname{span}\{u, u_*\}^\perp$ has dimension $D - 2 \ge 1$. Draw ξ from a conditionally symmetric distribution on its radius-$\sqrt{v}$ sphere. Then

$$\mathbb{E}[\xi \mid u] = 0, \qquad \|\xi\|^2 = v, \qquad \langle \xi, u \rangle = \langle \xi, u_* \rangle = \langle \xi, g \rangle = 0. \tag{27}$$

For $\bar{g} = g + \xi$, this gives $\mathbb{E}[\bar{g} \mid u] = \operatorname{grad} f(u)$, $U = S + v$, and $\langle u - \phi \bar{g}, u_* \rangle = q + \phi S$. The normalized tangent update satisfies (9) for every draw.

C.2 FINITE-HORIZON OPTIMALITY

For $q \in (0,1)$ and $S = 1 - q^2$, differentiation gives

$$\frac{\partial}{\partial \phi} F_v(q,\phi) = \frac{\partial}{\partial \phi}\left(\frac{q + \phi S}{\sqrt{1 + \phi^2(S+v)}}\right) \tag{28}$$

$$= \frac{S}{\sqrt{1 + \phi^2(S+v)}} - \frac{q + \phi S}{2} \frac{2\phi(S+v)}{[1 + \phi^2(S+v)]^{\frac{3}{2}}} \tag{29}$$

$$= \frac{S - q\phi(S+v)}{[1 + \phi^2(S+v)]^{\frac{3}{2}}} \tag{30}$$

To maximize the alignment, we take $\phi = S/[q(S+v)]$ where the derivative becomes 0. This is the unique one-step optimum, with ϕ substituted to its maximum representation to have:

$$M_v(q) := \max_{\phi > 0} F_v(q,\phi) = \sqrt{q^2 + \frac{S^2}{S+v}} = \sqrt{1 - \frac{vS}{S+v}} \in (0,1). \tag{31}$$

Every feasible next alignment is positive and at most $M_v(q) < 1$ as long as S is not 0. Since

$$M_v'(q) = \frac{v^2 q}{M_v(q)(1 - q^2 + v)^2} > 0,$$

the optimal one-step map is strictly increasing on $(0,1)$. Let $V_h(q)$ denote the maximum terminal alignment after h transitions, with $V_0(q) = q$. The Bellman recursion is

$$V_h(q) = \max_{\phi > 0} V_{h-1}(F_v(q,\phi)).$$

If V_{h-1} is strictly increasing, this maximization has the same unique optimizer $\phi^\star(q)$ as $F_v(q,\phi)$. Consequently, $V_h = V_{h-1} \circ M_v$ is also strictly increasing. Starting from $V_0(q) = q$, induction proves that the policy $\phi^\star(q)$ is optimal at every finite horizon.

C.3 ERROR TRAJECTORY, INDUCING FORCING, AND ONE CROSSING

Based on maximized selection of the ϕ, equation (31) gives

$$q_{t+1} = \sqrt{1 - \frac{vS_t}{S_t + v}} \tag{32}$$

$$S_{t+1} = 1 - q_{t+1}^2 = \frac{vS_t}{S_t + v} \tag{33}$$

We can rearrange the representation to have the following harmonic recurrence relationship between S_{t+1} and S_t under maximizing choice of $\phi = \frac{S}{q(S+v)}$:

$$S_{t+1} = \frac{vS_t}{S_t + v}, \qquad \frac{1}{S_{t+1}} = \frac{1}{S_t} + \frac{1}{v}. \tag{34}$$

Furthermore, we can unroll the S_{t+1} as following:

$$\frac{1}{S_{t+1}} = \frac{1}{S_t} + \frac{1}{v} = \left(\frac{1}{S_{t-1}} + \frac{1}{v}\right) + \frac{1}{v} = \frac{1}{S_0} + \frac{t+1}{v} = \mathcal{O}(t). \tag{35}$$

This provides the trajectory stated in (10). For the ease of derivation, we list several identities so far in the following:

$$\phi_{t+1} = B_t \frac{\phi_t}{1 + \phi_t^2 U_t} = B_t \frac{\phi_t}{1 + \phi_t^2 (S_t + v)} \tag{36}$$

$$S_{t+1} = \frac{vS_t}{S_t + v} \tag{37}$$

$$\phi_t = \frac{S_t}{q_t(S_t + v)} \tag{38}$$

$$q_{t+1}^2 = 1 - \frac{vS_t}{S_t + v} = \frac{S_t + v - S_t v}{S_t + v} \tag{39}$$

The exact scalar law requires the forcing

$$B_t = \frac{\phi_{t+1}}{\phi_t^*}\big[1 + (\phi_t)^2(S_t + v)\big]. \tag{40}$$

We know that

$$\phi_{t+1} = \frac{S_{t+1}}{q_{t+1}(S_{t+1} + v)} \tag{41}$$

$$= \frac{\frac{vS_t}{S_t+v}}{q_{t+1}(\frac{vS_t}{S_t+v} + v)} \tag{42}$$

$$= \frac{S_t}{q_{t+1}(2S_t + v)} \tag{43}$$

Substituting ϕ_{t+1} in the B_t representation:

$$B_t = \frac{\phi_{t+1}}{\phi_t}(1 + \phi_t^2(S_t + v)), \tag{44}$$

$$= \frac{q_t}{q_{t+1}}\frac{v + S_t}{v + 2S_t}(1 + \frac{S_t^2}{q_t^2(S_t + v)}), \tag{45}$$

$$= \frac{q_t}{q_{t+1}}\frac{v + S_t}{v + 2S_t}(\frac{q_t^2(S_t + v) + S_t^2}{q_t^2(S_t + v)}), \tag{46}$$

$$= \frac{q_t}{q_{t+1}}\frac{v + S_t}{v + 2S_t}(\frac{(1 - S_t)(S_t + v) + S_t^2}{q_t^2(S_t + v)}), \tag{47}$$

$$= \frac{q_t}{q_{t+1}}\frac{v + S_t}{v + 2S_t}\frac{S_t + v - S_t v}{q_t^2(S_t + v)}, \tag{48}$$

$$= \frac{q_{t+1}}{q_t}\frac{v + S_t}{v + 2S_t}, \tag{49}$$

where we use the identities listed above for the derivation in each step and yield (11). Squaring that expression and simplifying gives

$$(B_t)^2 - 1 = (\frac{q_{t+1}}{q_t}\frac{v + S_t}{v + 2S_t})^2 - 1, \tag{50}$$

$$= \frac{1 - S_{t+1}}{1 - S_t}(\frac{v + S_t}{v + 2S_t})^2 - 1, \tag{51}$$

$$= \frac{S_t\left[4S_t^2 + (3v - 3)S_t - 2v\right]}{(1 - S_t)(2S_t + v)^2}. \tag{52}$$

The denominator is positive. The quadratic $P_v(S) = 4S^2 + (3v - 3)S - 2v$ has one positive and one negative root. Its values $P_v(0) = -2v < 0$ and $P_v(1) = 1 + v > 0$ place the positive root S_c strictly in $(0, 1)$, with the value in Theorem 2. Hence $B_t > 1$ exactly when $S_t > S_c$. Since S_t decreases strictly to zero, the threshold is crossed at most once, from above to below. Equality can occur at one integer time, and there is no expansionary prefix when $S_0 \le S_c$.

Here, we aim to analyze the how B_t behaves under this optimal reference trajectory. First, we analyze the $\frac{v+S_t}{v+2S_t}$ Writing $\kappa = v/S_0$ gives

$$\frac{S_t + v}{2S_t + v} = \frac{t + \kappa + 1}{t + \kappa + 2} = 1 - \frac{1}{t + \kappa + 2} = 1 - \frac{1}{t} + \mathcal{O}(t^{-2}). \tag{53}$$

Next, we analyze the $\frac{q_{t+1}}{q_t}$ by first calculating its square:

$$(\frac{q_{t+1}}{q_t})^2 = \frac{1 - S_{t+1}}{1 - S_t} = 1 + \frac{S_t - S_{t+1}}{1 - S_t}. \tag{54}$$

We know that

$$\frac{1}{S_t} = \frac{t}{v} + \frac{1}{S_0} \implies S_t = \frac{vS_0}{tS_0 + v}. \tag{55}$$

and therefore,

$$(\frac{q_{t+1}}{q_t})^2 = 1 + \frac{\frac{vS_0}{tS_0+v} - \frac{vS_0}{(t+1)S_0+v}}{1 - \frac{vS_0}{tS_0+v}}, \tag{56}$$

$$= 1 + \frac{vS_0^2}{(tS_0 + v + S_0)(tS_0 + v - vS_0)}, \tag{57}$$

$$= 1 + \mathcal{O}(t^{-2}), \tag{58}$$

$$\implies \frac{q_{t+1}}{q_t} = 1 + \mathcal{O}(t^{-2}), \tag{59}$$

where the last step follows from Taylor expansion on the square root term $\sqrt{1 + \mathcal{O}(t^{-2})}$. It follows that $B_t = 1 - t^{-1} + O(t^{-2})$ and $\phi_t = t^{-1} + O(t^{-2})$. The error formula holds at every finite time; the asymptotic expressions describe its tail.

D Stochastic forcing tails and almost-sure stationarity

Throughout, Φ_t and u_t are $\mathcal{F}_t$-measurable, while $\bar{g}_t$ and U_t are $\mathcal{F}_{t+1}$-measurable. The initial coefficient $\Phi_0 > 0$ is fixed, and conditional moment bounds use finite deterministic constants. The forcing is deterministic, positive, and satisfies (12). Any finite prefix of positive forcing values is allowed. We first establish the product estimate and radial transformation shared by all three parts of Theorem 3.

D.1 Forcing products and radial rescaling

Lemma 1 (Forcing product)**.** *Let $B_t = 1 - \frac{\beta}{t+\tau} + \mathcal{O}(t^{-2})$, $Q_0 = 1$ and $Q_t = \prod_{j=0}^{t-1} B_j$. There exists a constant $K \in (0, \infty)$ such that $Q_t \sim Kt^{-\beta}$. With $k_t = Q_t^2$ and $A_t = 1 + \sum_{j<t} k_j$,*

$$\begin{aligned} \sum_t k_t < \infty \qquad & \text{if } \beta > 1/2, \\ A_t \sim K^2 \log t \qquad & \text{if } \beta = 1/2, \\ A_t \sim \frac{K^2}{1-2\beta} t^{1-2\beta} \qquad & \text{if } \beta < 1/2. \end{aligned} \tag{60}$$

For $\beta \le 1/2$,

$$\sum_t \left(\frac{k_t}{A_{t+1}}\right)^p < \infty \quad (p > 1), \qquad \frac{k_t^2}{A_t A_{t+1}} = O((t+1)^{-2}). \tag{61}$$

Proof. Under the sufficient local condition (13), for large t, Taylor expansion of the logarithm ($\log(1-x) \sim -x$ for small x) gives $\log B_t = -\beta/(t+\tau) + O(t^{-2})$. The remainder has an absolutely convergent sum, and $\sum_{j<t}(j+\tau)^{-1} - \log t$ has a finite limit. Therefore,

$$\log Q_t + \beta \log t \to L \tag{62}$$

Exponentiation yields the positive finite limit of $Q_t t^\beta$. Under the product hypothesis (12), this limit is assumed directly. Standard integral comparison, or summation of upper and lower multiplicative bounds on $k_t \sim K^2 t^{-2\beta}$, gives (60).

For $\beta < 1/2$, $k_t/A_{t+1} = O(t^{-1})$ and $A_{t+1}/A_t \to 1$. At $\beta = 1/2$, $k_t/A_{t+1} = O((t \log t)^{-1})$ and the same ratio tends to one. These estimates give both statements in (61). □

To separate this deterministic forcing product from gradient-driven norm growth, define

$$Y_t = \frac{Q_t}{\Phi_t} > 0, \qquad Z_t = Y_t^2. \tag{63}$$

The scalar law and $Q_{t+1} = B_t Q_t$ imply

$$Y_{t+1} = \frac{Q_t}{\Phi_t}(1 + \Phi_t^2 U_t) = Y_t + \frac{k_t U_t}{Y_t}, \tag{64}$$

$$Z_{t+1} = Z_t + 2k_t U_t + \frac{k_t^2 U_t^2}{Z_t}. \tag{65}$$

In particular, $Y_t \geq Y_0 = \Phi_0^{-1}$ and

$$0 < \Phi_t \leq \Phi_0 Q_t. \tag{66}$$

This deterministic envelope controls the endogenous step despite its dependence on past gradients. It supplies the summable error bound needed for stationarity.

D.2 Proof of Theorem 3(i)

Suppose $\beta > 1/2$ and $\mathbb{E}[U_t \mid \mathcal{F}_t] \leq M$. By Tonelli's theorem and Lemma 1,

$$\mathbb{E}\left[\sum_t k_t U_t\right] \leq M \sum_t k_t < \infty. \tag{67}$$

Hence $\sum_t k_t U_t < \infty$ almost surely. Equation (64) yields

$$Y_0 \leq Y_t = Y_0 + \sum_{j<t} \frac{k_j U_j}{Y_j} \leq Y_0 + Y_0^{-1} \sum_{j<t} k_j U_j. \tag{68}$$

The monotone sequence Y_t therefore converges almost surely to a finite $Y_\infty \geq Y_0 > 0$. Since $Q_t \sim Kt^{-\beta}$, $\Phi_t \sim (K/Y_\infty)t^{-\beta}$. Set $C = K/Y_\infty$, which is finite and strictly positive. Only the conditional first energy moment enters this argument. □

D.3 Proof of Theorem 3(ii)

When the squared forcing product is nonsummable, the leading radial increment must be averaged and its nonlinear remainder controlled separately. Two summability facts allow both steps under a conditional p-th energy moment with any $p > 1$. We first prove one technical lemma that will be utilized later

Lemma 2 (Moment addition upper bound)**.** *For $x, y \in \mathbb{R}$ and $p \in (1, 2]$, there exist $C_p > 0$ such that:*

$$|x + y|^p \leq |x|^p + p|x|^{p-1} \mathit{sign}(x) y + C_p |y|^p \tag{69}$$

Proof. Fix $1 < p \leq 2$ and set $\alpha = p - 1 \in (0, 1]$. Define

$$h(x) := |x|^\alpha \operatorname{sgn}(x), \qquad h(0) := 0.$$

We first show that

$$|h(a) - h(b)| \leq 2^{1-\alpha} |a - b|^\alpha = 2^{2-p} |a - b|^{p-1}.$$

If a and b have the same sign, subadditivity of x^α on $[0, \infty)$ gives

$$|h(a) - h(b)| = \big||a|^\alpha - |b|^\alpha\big| \leq \big||a| - |b|\big|^\alpha \leq |a - b|^\alpha.$$

If they have opposite signs, concavity of x^α gives

$$|h(a) - h(b)| = |a|^\alpha + |b|^\alpha \leq 2^{1-\alpha}(|a| + |b|)^\alpha = 2^{1-\alpha} |a - b|^\alpha.$$

Since the derivative of $|x|^p$ is $ph(x)$, the fundamental theorem of calculus yields

$$|x + y|^p - |x|^p - ph(x)y = py \int_0^1 \big(h(x + sy) - h(x)\big)\, ds.$$

Therefore,

$$\begin{aligned} \big||x + y|^p - |x|^p - ph(x)y\big| &\leq p|y| \int_0^1 2^{2-p} |sy|^{p-1}\, ds \\ &= 2^{2-p} |y|^p. \end{aligned}$$

In particular,

$$|x + y|^p \leq |x|^p + p|x|^{p-1} \operatorname{sgn}(x) y + 2^{2-p} |y|^p,$$

which proves the claim with $C_p = 2^{2-p}$. □

Lemma 3 (Two summability facts)**.** *Fix $p \in (1, 2]$.*

1. *If martingale differences D_t satisfy $\mathbb{E}[|D_t|^p \mid \mathcal{F}_t] \le C$ and $\mathbb{E}[D_t \mid \mathcal{F}_t] = 0$, then for deterministic weights c_t with $\sum_t |c_t|^p < \infty$, the series $\sum_t c_t D_t$ converges almost surely and in L^p.*
2. *If $U_t \ge 0$ and $\sup_t \mathbb{E}[U_t^p] < C$, then*

$$\sum_{t \ge 0} \frac{U_t^2}{(t+1)^2} < \infty \quad \textit{almost surely}. \tag{70}$$

Proof. For the first claim, let $S_n = \sum_t^n c_t D_t$ which we aim to upper bound and we can write it in the following way:

$$\begin{aligned} |S_n|^p &= |\sum_{t=1}^n c_t D_t|^p, && (71)\\ &= |S_{n-1} + c_n D_n|^p, && (72)\\ &\le |S_{n-1}|^p + p|S_{n-1}|^{p-1}\text{sign}(S_{n-1})c_n D_n + C_p|c_n D_n|^p, && (73)\\ &\le \sum_{t=1}^n C_p|c_t D_t|^p + p\sum_{t=1}^{n-1} |S_t|^{p-1}\text{sign}(S_t)c_{t+1}D_{t+1} && (74) \end{aligned}$$

where the first inequality follows from the Lemma 2 and the second follows from unrolling the whole expression. By taking expectations on both side, the linear term vanishes, so their p-th moments are bounded by $C_p C \sum_t |c_t|^p$. Thus the partial sums form an L^p-bounded martingale. Since $p > 1$, the martingale convergence theorem gives convergence almost surely and in L^p.

For the second claim, Markov's inequality gives

$$\sum_t \Pr\{U_t > t+1\} \le C \sum_t (t+1)^{-p} < \infty \tag{75}$$

Therefore, with the first Borel–Cantelli lemma, it leaves only finitely many such exceedances almost surely, i.e.,

$$\sum_t \frac{U_t^2}{(t+1)^2} \mathbf{1}_{\{U_t > t+1\}} < \infty \tag{76}$$

On their complement,

$$\mathbb{E}[\sum_t \frac{U_t^2}{(t+1)^2} \mathbf{1}_{\{U_t \le t+1\}}] \le C \sum_t (t+1)^{-p} < \infty, \tag{77}$$

since $U_t^2 \mathbf{1}_{\{U_t \le t+1\}} = U_t^{2-p} U_t^p \mathbf{1}_{\{U_t \le t+1\}} \le (t+1)^{2-p} U_t^p$. Tonelli's theorem proves summability of the truncated terms, and the finite set of exceedances does not affect it. The conclusion therefore holds even if $\mathbb{E}U_t^2$ is infinite. □

The first fact yields the required weighted average of conditional energy.

Lemma 4 (Weighted conditional-energy averaging)**.** *Suppose $\beta \le 1/2$, $\mu_t = \mathbb{E}[U_t \mid \mathcal{F}_t] \to v > 0$ almost surely, and $\mathbb{E}[U_t^p \mid \mathcal{F}_t] \le M_p$ for some $p \in (1, 2]$. Then*

$$\frac{\sum_{j<t} k_j U_j}{A_t} \longrightarrow v \quad \textit{almost surely}. \tag{78}$$

Proof. Let $D_t := U_t - \mu_t$. Decompose

$$\frac{1}{A_t}\sum_{j<t} k_j U_j = \frac{1}{A_t}\sum_{j<t} k_j \mu_j + \frac{1}{A_t}\sum_{j<t} k_j D_j.$$

Our goal is to prove that the first term tends to v and second term tends to 0. We first start with the second term. we can upper bound the $|D_j|^p$, Since $\mu_t = \mathbb{E}[U_t \mid \mathcal{F}_t]$, we have $\mathbb{E}[D_t \mid \mathcal{F}_t] = 0$. Convexity and conditional Jensen's inequality give

$$\begin{aligned}\mathbb{E}[|D_t|^p \mid \mathcal{F}_t] &\le 2^{p-1}\left(\mathbb{E}[U_t^p \mid \mathcal{F}_t] + |\mu_t|^p\right)\\ &\le 2^p \mathbb{E}[U_t^p \mid \mathcal{F}_t]\\ &\le 2^p M_p.\end{aligned}$$

We first try to claim that $\sum_t \frac{k_t D_t}{A_{t+1}}$ converges almost surely. By picking $c_t = \frac{k_t}{A_{t+1}}$, we have that

$$\sum_t \frac{k_t D_t}{A_{t+1}} = \sum_t c_t D_t \tag{79}$$

By Lemma 1 and Lemma 3, we know that $\sum_t \frac{k_t D_t}{A_{t+1}}$ converges almost surely as $\beta \le \frac{1}{2}$. We follows the same procedure as Kronecker's lemma as following (for self-contained purpose). Let $z_j = \frac{k_j D_j}{A_{j+1}}$ and we have that

$$S_n = \sum_{j=0}^{n} z_j \to S, \qquad S_{-1} = 0, \tag{80}$$

$$\because z_j = S_j - S_{j-1}, \tag{81}$$

$$\implies \sum_{j=0}^{t-1} k_j D_j = \sum_{j=0}^{t-1} A_{j+1} z_j, \tag{82}$$

$$= A_t S_{t-1} - \sum_{j=0}^{t-2} (A_{j+2} - A_{j+1}) S_j. \tag{83}$$

Divide both side by A_t and we get:

$$\frac{1}{A_t}\sum_{j=0}^{t-1} k_j D_j = S_{t-1} - \sum_{j=0}^{t-2} \frac{(A_{j+2} - A_{j+1})}{A_t} S_j. \tag{84}$$

We know that $S_t \to S$ and that $\frac{A_t - A_1}{A_t} \to 1$ and we have that both terms tend to S and we get $\frac{1}{A_t}\sum_{j=0}^{t-1} k_j D_j \to 0$. Now, we go on and prove the first term,

$$\frac{\sum_{j<t} k_j \mu_j}{A_t} - v = \frac{\sum_{j<t} k_j (\mu_j - v)}{A_t} - \frac{v}{A_t} \because A_t = 1 + \sum_{j<t} k_j. \tag{85}$$

Work on an event of probability one on which $\mu_j \to v$. For any $\varepsilon > 0$, choose a finite J such that $|\mu_j - v| \le \varepsilon$ for all $j \ge J$. For $t > J$,

$$\left|\frac{1}{A_t}\sum_{j<t} k_j \mu_j - v\right| \le \frac{\sum_{j<J} k_j |\mu_j - v| + |v|}{A_t} + \varepsilon,$$

because $k_j \ge 0$ and $\sum_{j<t} k_j = A_t - 1$. The first term vanishes as $A_t \to \infty$. Letting $\varepsilon \downarrow 0$ proves

$$\frac{1}{A_t}\sum_{j<t} k_j \mu_j \longrightarrow v \qquad \text{almost surely.}$$

□

Controlling the nonlinear remainder. Summing (65) gives

$$Z_t = Z_0 + 2\sum_{j<t} k_j U_j + \sum_{j<t} R_j, \qquad R_j := \frac{k_j^2 U_j^2}{Z_j} \ge 0. \tag{86}$$

By Lemma 4, almost surely there is a finite random index after which $Z_t \geq vA_t$. Therefore, eventually,

$$\frac{R_t}{A_{t+1}} \leq \frac{1}{v}\frac{k_t^2 U_t^2}{A_t A_{t+1}}. \tag{87}$$

By (61), the series on the right is bounded, apart from a finite prefix, by a deterministic multiple of $\sum_t U_t^2/(t+1)^2$. Lemma 3 makes this series finite almost surely, even when its expectation is infinite. Every omitted prefix term is finite because $Z_t \geq Z_0 > 0$ and U_t is finite almost surely. Consequently $\sum_t R_t/A_{t+1} < \infty$ almost surely. A second application of Kronecker's lemma yields

$$\frac{1}{A_t}\sum_{j<t} R_j \longrightarrow 0. \tag{88}$$

Equation (86) now implies the exact asymptotic normalization

$$\frac{Z_t}{A_t} \longrightarrow 2v \quad \text{almost surely.} \tag{89}$$

Since $\Phi_t^2 = k_t/Z_t$, the two divergent cases of (60) prove (15), including their constants. This completes Theorem 3(ii). □

D.4 Proof of Theorem 3(iii)

Write $g_t = \operatorname{grad} f(u_t)$ and $\zeta_t = \bar{g}_t - g_t$. Part (i) gives $\sum \Phi_t = \infty$ and $\sum \Phi_t^2 < \infty$ almost surely for $1/2 < \beta \leq 1$. Weighted descent will first give a zero gradient liminf. Convergence of the accumulated perturbations will then rule out repeated excursions away from the stationary set.

Descent and weighted gradient energy. To begin the proof, we first start with the function $\widetilde{f}$ which extend beyond the sphere by $\widetilde{f}(x) = f(x/\|x\|)$ for $x \neq 0$. Its Hessian scales as $\|x\|^{-2}$ under radial rescaling. Since $f \in C^2$ on the compact sphere, this Hessian is bounded by a constant L on $\{x : \|x\| \geq 1\}$. Now, we observe the Taylor's expansion of this function:

$$f(u_{t+1}) = \widetilde{f}(u_t - \Phi_t \bar{g}_t) \leq \widetilde{f}(u_t) - \Phi_t\langle \bar{g}_t, g_t\rangle + \frac{L}{2}\Phi_t^2 U_t, \tag{90}$$

$$\implies f(u_{t+1}) \leq f(u_t) - \Phi_t\langle \bar{g}_t, g_t\rangle + \frac{L}{2}\Phi_t^2 U_t, \tag{91}$$

$$\because \widetilde{f}(x) = f(x), \qquad \forall \|x\| = 1. \tag{92}$$

Conditional expectation, unbiasedness, and (66) imply

$$\mathbb{E}[f(u_{t+1}) \mid \mathcal{F}_t] \leq f(u_t) - \Phi_t\|g_t\|^2 + \frac{LM\Phi_0^2}{2}k_t. \tag{93}$$

Let $f_{\min}$ be the minimum of f on the sphere and $c_0 = LM\Phi_0^2/2$. Then define

$$V_t = f(u_t) - f_{\min} + c_0 \sum_{j=t}^{\infty} k_j \tag{94}$$

is an integrable nonnegative supermartingale with $\mathbb{E}[V_{t+1} \mid \mathcal{F}_t] \leq V_t - \Phi_t\|g_t\|^2$ due to Equation (93). It converges almost surely. Summing expectations shows

$$\mathbb{E}\sum_t \Phi_t\|g_t\|^2 \leq \mathbb{E}V_0 < \infty \qquad \because \sum_j k_j < \infty \text{ with } \frac{1}{2} < \beta \leq 1. \tag{95}$$

Hence

$$f(u_t) \text{ converges}, \qquad \sum_t \Phi_t\|g_t\|^2 < \infty \quad \text{almost surely.} \tag{96}$$

Together with $\sum \Phi_t = \infty$, this gives $\liminf_t \|g_t\| = 0$.

Accumulated noise and normalization error. Unbiasedness implies $\mathbb{E}[\|\zeta_t\|^2 \mid \mathcal{F}_t] = \mathbb{E}[\|\bar{g}_t\|^2 \mid \mathcal{F}_t] - \|g_t\|^2 \le M$ due to assumption used. The vector martingale $\sum_{j<t} \Phi_j \zeta_j$ can be upper bounded as following

$$\sum_t \mathbb{E}\|\Phi_t \zeta_t\|^2 \le M\Phi_0^2 \sum_t k_t < \infty \tag{97}$$

and has total expected squared increment at most $M\Phi_0^2 \sum_j k_j < \infty$, so it converges in L^2 and almost surely.

Now, we estimate the error caused by update and normalization. Define $e_t = u_{t+1} - (u_t - \Phi_t \bar{g}_t)$ satisfies

$$\|e_t\| = \sqrt{1 + \Phi_t^2 U_t} - 1 \le \frac{1}{2}\Phi_t^2 U_t, \qquad \because \sqrt{1+x} - 1 = \frac{x}{\sqrt{1+x}+1} \le \frac{1}{2}x. \tag{98}$$

By (66) and Tonelli's theorem, $\sum_t \Phi_t^2 U_t < \infty$ almost surely, hence $\sum_t \|e_t\| < \infty$. Thus, on an event of probability one,

$$u_{t+1} - u_t = -\Phi_t g_t + \delta_t, \qquad \delta_t = -\Phi_t \zeta_t + e_t, \qquad \sum_t \delta_t \text{ converges.} \tag{99}$$

Convergence of this series implies uniformly vanishing tail sums: $\sup_{m \ge n} \|\sum_{j=n}^m \delta_j\| \to 0$. The argument uses convergence of accumulated noise, not absolute summability of the individual stochastic-gradient perturbations.

Excluding repeated gradient excursions. The map $u \mapsto \operatorname{grad} f(u)$ (i.e., $g(u)$ which is the gradient at the unit sphere) is bounded by G and Lipschitz with some finite constant L_g on the sphere, using its ambient representation:

$$\|g(u)\| \le G, \qquad \|g(u) - g(v)\| \le L_g \|u - v\|. \tag{100}$$

Now, suppose the gradient norm does not tend to zero. Since its liminf is zero, there are an $\epsilon > 0$ and infinitely many disjoint index intervals $[s_k, t_k)$ with

$$\|g_{s_k}\| \ge \epsilon, \qquad \|g_{t_k}\| \le \epsilon/2, \qquad \|g_j\| > \epsilon/2 \quad (s_k \le j < t_k). \tag{101}$$

Here t_k is the first subsequent crossing below $\epsilon/2$ after a high-gradient index. We first observe the decrease in the gradient will lead to large change on the sphere through the following:

$$\frac{1}{2}\epsilon \le \|g_{s_k}\| - \|g_{t_k}\| \le \|g_{s_k} - g_{t_k}\| \le L_g \|u_{s_k} - u_{t_k}\|, \tag{102}$$

$$\implies \|u_{s_k} - u_{t_k}\| \ge \frac{\epsilon}{2L_g}, \tag{103}$$

and we know that

$$u_{t_k} - u_{s_k} = -\sum_{j=s_k}^{t_k-1} \Phi_j g_j + \sum_{j=s_k}^{t_k-1} \delta_j. \tag{104}$$

For sufficiently late intervals, we can upper bound the error with at most $\frac{\epsilon}{4L_g}$:

$$\|\sum_{j=s_k}^{t_k-1} \delta_j\| \le \frac{\epsilon}{4L_g}, \tag{105}$$

and therefore, the norm of the summed gradient must at least be of certain magnitude $\frac{\epsilon}{4L_g}$:

$$G\sum_{j=s_k}^{t_k-1} \Phi_j \ge \|-\sum_{j=s_k}^{t_k-1} \Phi_j g_j\| \ge \frac{\epsilon}{4L_g}. \tag{106}$$

Within the interval, we have that

$$\|g_j\| > \frac{\epsilon}{2} \implies \sum_{j=s_k}^{t_k-1} \Phi_j \|g_j\|^2 \ge \frac{\epsilon^2}{4} \sum_{j=s_k}^{t_k-1} \Phi_j \ge \frac{\epsilon^3}{16 G L_g} \tag{107}$$

Each interval therefore contributes at least $\epsilon^3/(16 L_g G) > 0$ to $\sum_j \Phi_j \|g_j\|^2$. Infinitely many disjoint intervals contradict (96). Hence $\|\operatorname{grad} f(u_t)\| \to 0$ almost surely, proving part (iii) under only the bounded second gradient moment and unbiasedness assumptions.

D.5 ACCUMULATED ANGULAR ACTION

Proposition 1 (Angular-action classification)**.** *Under Theorem 3(i), $\sum_t \theta_t^2 < \infty$ almost surely. Under part (ii), almost surely,*

$$\sum_{t<T} \theta_t^2 \sim \begin{cases} \frac{1}{2}\log\log T, & \beta = 1/2, \\ \frac{1-2\beta}{2}\log T, & \beta < 1/2. \end{cases} \tag{108}$$

Proof. The update step is

$$u_{t+1} = \frac{u_t - \Phi_t \bar{g}_t}{\sqrt{1+\Phi_t^2 U_t}}, \qquad \bar{g}_t \perp u_t \tag{109}$$

therefore, on a unit sphere we have

$$\tan\theta_t = \Phi_t\sqrt{U_t}, \qquad \theta_t = \arctan(\Phi_t\sqrt{U_t}). \tag{110}$$

Since $0 \le \arctan(z) \le z,\ \forall z \ge 0$, we have

$$\theta_t^2 \le \Phi_t^2 U_t. \tag{111}$$

Now, we use that

$$\mathbb{E}[\sum_t \theta_t^2] \le \Phi_0^2 \sum_t k_t \mathbb{E}[U_t] \le M\Phi_0^2 \sum_t k_t < \infty, \tag{112}$$

and we prove that it converges almost surely for the first claim.

For the second claim, we start with the inequality

$$z - \arctan(z) = \int_0^z (s - \arctan(s))' ds \tag{113}$$

$$= \int_0^z 1 - \frac{1}{1+s^2} ds \tag{114}$$

$$= \int_0^z \frac{s^2}{1+s^2} ds \tag{115}$$

$$\le \int_0^z s^2 ds = \frac{z^3}{3} \tag{116}$$

Along with the fact that $\arctan(z) \le z$, we have the following:

$$0 \le z^2 - (\arctan z)^2 = (z - \arctan z)(z + \arctan z) \le \tfrac{2}{3} z^4 \qquad (z \ge 0). \tag{117}$$

Now, we set $z = \Phi_t\sqrt{U_t}$ and have the following:

$$0 \le \Phi_t^2 U_t - \theta_t^2 \le \frac{2}{3}\Phi_t^4 U_t^2. \tag{118}$$

We have that for $\beta \le \frac{1}{2}$ with Theorem 3 (ii),

$$\Phi_t^2 \le \frac{C}{t} \implies \Phi_t^4 U_t^2 \le C^2 \frac{U_t^2}{t^2}. \tag{119}$$

Lemma 3 gives $\sum_{t\ge1} U_t^2/t^2 < \infty$ almost surely, and therefore $\sum_t |\Phi_t^2 U_t - \theta_t^2| < \infty$.

The martingale series $\sum_{t\ge2}(U_t - \mu_t)/t$ and $\sum_{t\ge2}(U_t - \mu_t)/(t\log t)$ both converge almost surely by Lemma 3, since their deterministic weights have summable p-th powers. As $\mu_t \to v$,

$$\sum_{2\le t<T} \frac{U_t}{t} \sim v\log T, \qquad \sum_{2\le t<T} \frac{U_t}{t\log t} \sim v\log\log T \tag{120}$$

almost surely, by the same tail-squeezing argument as in Lemma 4. Multiplying these comparisons by the asymptotic constants in (15), with upper and lower multiplicative bounds on the eventually positive terms, proves (108). □

Under the conditional-energy and moment assumptions of Theorem 3(ii), considered across all $\beta \in \mathbb{R}$, $\sum \Phi_t = \infty$ and $\sum \Phi_t^2 < \infty$ hold simultaneously exactly for $1/2 < \beta \le 1$. At $\beta = 1/2$, squared steps accumulate at order $\log\log T$; for $\beta < 1/2$, they accumulate at order $\log T$. Divergent squared step or angular action is compatible with optimization convergence for some objectives and noise geometries.

D.6 Proof of Corollary 1

Set $d = c\lambda$ and $z = t + \tau$. Positivity of $a_t = 1 - d/z$ follows from $\tau > d$. The native forcing is exactly

$$B_t = \frac{z}{z+1}\frac{z}{z-d}\frac{z+1}{z+1-d} = \frac{z^2}{(z-d)(z+1-d)} = 1 - \frac{1-2d}{z} + O(z^{-2}). \tag{121}$$

Thus $\beta = 1 - 2d$. Substitution into Theorem 3 proves (17). Since $d \geq 0$, the case $d < 1/4$ corresponds precisely to $1/2 < \beta \leq 1$, where part (iii) applies. The constants in the other two cases require persistent conditional energy and the $2p$-th gradient moment for some $p > 1$, as stated in part (ii). □

D.7 Sharpness and scope

The asymptotic constants for $\beta \leq 1/2$ use a positive limiting conditional energy. With $U_t \equiv 0$, one instead has $\Phi_t = \Phi_0 Q_t \asymp t^{-\beta}$; the noise-supported $t^{-1/2}$ law is then false. Likewise, $\beta > 1$ makes $\sum \Phi_t < \infty$ and does not guarantee stationarity from arbitrary initialization. Random forcing products from adaptive feedback, additional momentum state, and noncompact quotients without global smoothness control require separate arguments. The deterministic product estimate and the compact-sphere descent bound are explicit dependencies of the present proof.

A conditional mean does not fix the tail constant. The slightly higher moment in part (ii) cannot simply be dropped from this uniform conditional-moment formulation. Take $B_t \equiv 1$ and independent $U_t = (t+1)I_t$, where $I_t \sim \mathrm{Bernoulli}(1/(t+1))$, with the natural filtration $\mathcal{F}_t = \sigma(I_j : 0 \leq j < t)$. Then $\mathbb{E}[U_t \mid \mathcal{F}_t] = 1$ at every time, but the conclusion $t\Phi_t^2 \to 1/2$ is impossible almost surely. Independence and $\sum_t \Pr\{I_t = 1\} = \infty$ imply infinitely many shocks. If $t\Phi_t^2 \to 1/2$ held, the ratio Φ_{t+1}^2/Φ_t^2 would tend to one, whereas at those shocks the exact recurrence gives

$$\frac{\Phi_{t+1}^2}{\Phi_t^2} = \frac{1}{[1 + (t+1)\Phi_t^2]^2} \longrightarrow \frac{4}{9}, \tag{122}$$

a contradiction. For every $p > 1$, the conditional p-th energy moment in this example grows as $(t+1)^{p-1}$. The failure occurs in the nonsummable squared-forcing regime. Part (iii) uses $\beta > 1/2$, where the summable deterministic envelope controls even such rare shocks through their conditional first energy moment.

E Radial defects, feedback, and finite-horizon control

The quotient law has useful exact counterparts when the update is not tangent or the input is chosen by feedback. These identities clarify which parts of the control description remain available beyond the assumptions of the stochastic theorem.

E.1 An exact correction for radial defects

Consider $w_{t+1} = a_t w_t - \eta_t p_t$ for an arbitrary update vector p_t. Decompose $r_t p_t = \rho_t u_t + \tau_t$ with $\tau_t \perp u_t$, and retain $\Phi_t = \eta_t/(a_t r_t^2)$ and $B_t = \eta_{t+1}/(\eta_t a_t a_{t+1})$. Factoring $a_t r_t$ as in the tangent case gives

$$H_t = (1 - \Phi_t\rho_t)^2 + \Phi_t^2\|\tau_t\|^2, \qquad u_{t+1} = \frac{(1-\Phi_t\rho_t)u_t - \Phi_t\tau_t}{\sqrt{H_t}}, \qquad \Phi_{t+1} = \frac{B_t\Phi_t}{H_t}, \tag{123}$$

provided $H_t > 0$. The identity uses no small-defect approximation. If

$$z_t = \frac{2\Phi_t\rho_t - \Phi_t^2\rho_t^2}{1 + \Phi_t^2\|\tau_t\|^2}, \qquad |z_t| \leq \delta < 1, \tag{124}$$

then $H_t = (1 + \Phi_t^2\|\tau_t\|^2)(1 - z_t)$, and

$$\left|\log\frac{\Phi_{t+1}}{\Phi_t} - \left[\log B_t - \log(1 + \Phi_t^2\|\tau_t\|^2)\right]\right| \leq -\log(1-\delta). \tag{125}$$

Tangency violations can therefore be measured directly in logarithmic step drift. Positive BatchNorm epsilon, additive components, and numerical error can all contribute radial terms. With momentum or preconditioning, p_t also carries optimizer state. The corrected identity remains exact, but B_t alone is no longer a complete control coordinate and Theorem 3 does not apply without further analysis.

E.2 Feedback and shared forcing across blocks

After observing a tangent gradient, the choice

$$B_t = (1 + \Phi_t^2 U_t)\frac{\Phi_{t+1}^{\mathrm{ref}}}{\Phi_t} \tag{126}$$

requests a specified next effective step. The current update $u_t \mapsto u_{t+1}$ is already fixed, and the requested input must still satisfy native bounds. For a fixed learning-rate path, (6) uniquely determines its next multiplier.

For several blocks sharing $b_t = \log B_t$, let their current norm factors be $H_{i,t} > 0$. Their logarithmic step drifts at a fixed current state satisfy

$$d_{i,t} := \log \Phi_{i,t+1} - \log \Phi_{i,t} = b_t - \log H_{i,t}. \tag{127}$$

The difference $d_{i,t} - d_{j,t} = \log H_{j,t} - \log H_{i,t}$ is independent of the common input. For a common target d^{ref} and positive weights w_i, the weighted least-squares input is $b_t = d^{\mathrm{ref}} + \sum_i w_i \log H_{i,t} / \sum_i w_i$. The minimax input is $b_t = d^{\mathrm{ref}} + (\min_i \log H_{i,t} + \max_i \log H_{i,t})/2$, with irreducible worst residual $(\max_i \log H_{i,t} - \min_i \log H_{i,t})/2$. A shared input can shift all current drifts but cannot independently set each one. These are instantaneous statements; future norm factors can change with the chosen input, so the drift differences are not trajectory invariants.

E.3 Horizon normalization and finite-horizon cosine

For the continuous profile ζ and request $B_t^{(N)} = \exp(\zeta(t/N)/N)$ in (18), the signed and absolute log-actions are the left Riemann sums of ζ and $|\zeta|$. Continuity on $[0, 1]$ proves both limits. In contrast, repeating a fixed nonzero log-forcing shape over N steps generally accumulates action of order N. Projection may change either limiting action, which is why the experiments audit realized rather than requested budgets.

For the positive-floor cosine learning rate $\eta_t = \eta_{\min} + (\eta_{\max} - \eta_{\min})(1 + \cos(\pi t/N))/2$ and fixed $\lambda > 0$ satisfying $\lambda\eta_{\max} < 1$, the learning-rate ratio differs from one by $O(1/N)$ uniformly in time. The decay factors $(1 - \eta_t\lambda)^{-1}$ remain nontrivial as N grows. Stretching the same native curve over a longer horizon can therefore produce a qualitatively different forcing path. Whether a late $B_t > 1$ increases the effective step is determined by the full drift (4), including $U_t\Phi_t^2$. This finite-horizon dependence is distinct from the harmonic asymptotics of Corollary 1.

F Experimental protocols and complete results

The experiments comprise stochastic-oracle tests, an exact-symmetry forcing-order intervention, a conventional-network width sweep, and a separately executed scope and lower-action follow-up. We give the oracle constructions and numerical checks first, followed by statistical conventions, neural protocols, and all comparisons. Neural conclusions concern the tested momentum-free training systems; the oracle optimality result and stochastic stationarity theorem are assessed under their own stated assumptions.

F.1 Stochastic simulations and numerical checks

The experiment in Figure 3 uses the exact conditional energy distribution of an unbiased tangent oracle on $\mathbb{S}^2$ for the constant objective. If z is uniform on $\mathbb{S}^2$ and u is any unit vector, then $X = \langle u, z\rangle$ is uniform on $[-1, 1]$ and

$$\bar{g} = \sqrt{3/2}(I - uu^\top)z, \qquad U = \|\bar{g}\|^2 = \tfrac{3}{2}(1 - X^2), \qquad \mathbb{E}U = 1, \quad \mathbb{E}U^2 = \tfrac{6}{5}. \tag{128}$$

The energy distribution is the same at every current direction. Sampling the scalar energy law therefore gives the exact distribution of Φ_t without simulating the direction. This constant-objective experiment measures radial scaling. A separate nonconstant objective tests stationarity below.

We use $\Phi_0 = 0.1$, $B_t = ((t+20)/(t+21))^\beta$, and $\beta \in \{-0.5, 0, 0.5, 0.75, 1\}$. For each of 64 independently sampled energy streams, the same draws are shared across exponents. Each stream has 200,000 transitions. Curves are retained on a logarithmic time grid.

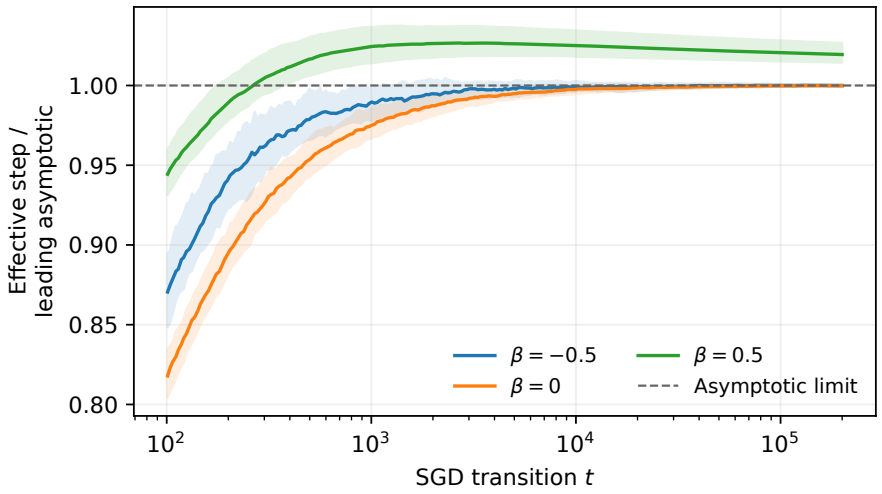

Figure 4: **Leading-constant checks for the stochastic tail theorem.** The ordinate is Φ_t divided by its leading asymptotic equivalent in Theorem 3(ii). Lines are medians and shaded bands are the 10th–90th percentiles across 64 independent streams, not confidence intervals for a population mean. The critical case has slowly decaying finite-horizon corrections.

At the final time, the median asymptotic ratios are 0.99996 for $\beta = -0.5$, 0.99984 for $\beta = 0$, and 1.01942 for $\beta = 0.5$. For $\beta = 0.75$ and 1, the predicted coefficient is path-dependent; the medians of Φ_t/Q_t are 0.07450 and 0.08422, respectively. Last-decade fitted log-slopes are $-0.49965, -0.49940, -0.54710, -0.75177, -0.99969$ in increasing order of β which is aligned with our theoretical prediction. The critical log-slope is still appreciably below $-1/2$ at this horizon, as its logarithmic factor predicts. The asymptotic ratios distinguish that correction from a different limiting power.

Infinite fourth gradient moments. To exercise the weaker moment range, we repeat the recurrence experiment with $U = V/3$, where $\Pr(V > x) = x^{-3/2}$ for $x \geq 1$. Then

$$\mathbb{E}U = 1, \qquad \mathbb{E}U^2 = \infty, \qquad \mathbb{E}U^{5/4} = \frac{6}{3^{5/4}} < \infty. \tag{129}$$

To realize this energy, draw a uniform unit tangent direction and multiply it by an independent $\sqrt{U}$, conditional on the current state. The resulting gradient has zero conditional mean and finite $5/2$-th moment but infinite fourth moment. We use the same five exponents, initial state, horizon, and replicate count as above. The final median ratios for $\beta = -0.5, 0, 0.5$ are $1.00525, 1.00470, 1.03586$ and the median values of Φ_t/Q_t are $0.07617, 0.08614$. Figure 5 retains every stream, showing the larger dispersion and the boundary correction without removing finite-time outliers.

Stationarity under heavy-tailed noise. For a separate stationarity check, take $f(u) = -\langle u, u^\star \rangle$ on $\mathbb{S}^2$, with $q_0 = 0.3$ and $\Phi_0 = 0.1$. Write $S = 1 - q^2$ and use the unbiased oracle $\bar{g} = qu - u^\star + \xi$, where ξ has conditionally symmetric sign in the direction orthogonal to $\mathrm{span}\{u, u^\star\}$ and random energy $E = 0.2V/3$ with the same Pareto law. Thus $\mathbb{E}_t\|\bar{g}_t\|^2 = S_t + 0.2 \leq 1.2$, while the fourth gradient moment is infinite. The exact recurrences are

$$q_{t+1} = \frac{q_t + \Phi_t S_t}{\sqrt{1 + \Phi_t^2(S_t + E_t)}}, \qquad \Phi_{t+1} = \frac{B_t \Phi_t}{1 + \Phi_t^2(S_t + E_t)}. \tag{130}$$

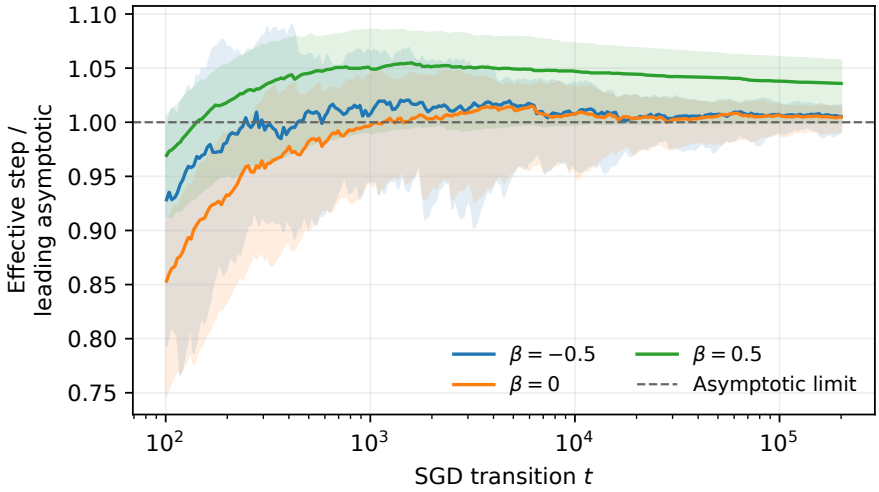

Figure 5: **Tail constants beyond fourth-moment assumptions.** The same asymptotic ratios as Figure 4, now with Pareto energy of tail index $3/2$ and infinite second energy moment. Medians and 10th–90th percentile bands use all 64 streams.

We use $\beta \in \{0.75, 1\}$, $B_t = ((t+20)/(t+21))^\beta$, 64 paired energy streams, and 200,000 updates. The forcing tails are deterministic, as required by Theorem 3(iii), rather than chosen by oracle-optimal feedback. The final median squared gradients S_t are 8.8394×10^{-6} and 1.2478×10^{-6}. The nonconstant objective tests descent as well as radial scaling within the specified tangent-oracle geometry.

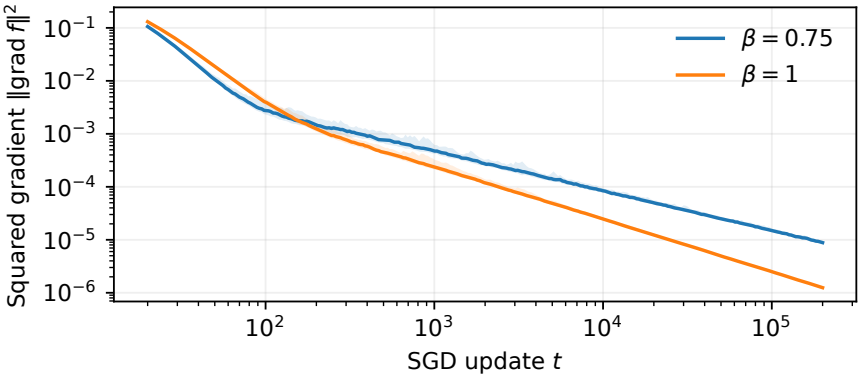

Figure 6: **Stationarity with endogenous steps and infinite fourth gradient moments.** Squared Riemannian gradients for the nonconstant alignment objective under deterministic forcing tails. Lines and bands show medians and 10th–90th percentiles over all 64 streams.

F.2 Exact forcing-order intervention

Model and data. In this section, we use the custom ResNet-18 implementation in the supplied source, not the standard torchvision model. All BatchNorm layers are affine-free with epsilon zero and running-statistics tracking disabled to ensure degree of scale-invariant property. CIFAR-10 uses its official 50,000 training and 10,000 test examples (Krizhevsky, 2009). Training draws batch-128 indices with replacement and uses reflect-padded random crops and horizontal flips.

Schedules and parameter isolation. There are 10,000 momentum-free SGD updates. The cosine path (Loshchilov & Hutter, 2017) has $\eta_0 = 0.5$ and a post-endpoint value $\eta_{10000} = 0.05$. The initial coefficient is $\lambda_0 = 0.001$, so $a_0 = 1 - \eta_0\lambda_0$. We synthesize the following across the whole training:

$$a_{t+1} = \frac{\eta_{t+1}}{\eta_t a_t B_t}, \qquad \lambda_{t+1} = \frac{1 - a_{t+1}}{\eta_{t+1}} \quad (0 \le t \le 9998). \tag{131}$$

The complete schedules are generated before training and without clipping. Signed coefficients are permitted in $[-0.08, 0.08]$ only on every convolutional weight tensor and all complementary trainable parameters receive fixed decay 0.001. Throughout the experiments, we use eight seeds and same configuration across different setting and report final test accuracy.

Forcing budgets and numerical geometry. The maximum target-forcing error is 4.44×10^{-16} and maximum reversal discrepancy is 6.66×10^{-16}. Both paths have absolute log-action 15.00002250 and signed action -0.0150015408 respectively. The native dose $\sum \eta_t\lambda_t$ is 1.13544 forward and 1.11943 reverse, with coefficient ranges $[-0.05005, 0.00498]$ and $[-0.01041, 0.06989]$. These native differences are induced by unique synthesis, not an independently matched covariate. Signed decay makes both orders reachable.

An endpoint-blind diagnostic audit of the original histories covered all 160,000 logged updates, 20 convolutional layers, and 4,800 output-filter blocks per update. For each block, let w_t denote its weights and g_t its loss gradient, excluding weight decay. We measure departure from tangency by

$$e_t^{\tan} = \frac{|\langle w_t, g_t\rangle|}{\|w_t\|\,\|g_t\|}.$$

To check the directional recurrence, define $g_t^{\perp} = g_t - \langle w_t, g_t\rangle w_t/\|w_t\|^2$ and $U_t^{\perp} = \|w_t\|^2\|g_t^{\perp}\|^2$, and compare the observed ratio $q_t^{\mathrm{obs}} = \Phi_{t+1}/\Phi_t$ with the tangent-law prediction $q_t^{\mathrm{pred}} = B_t/(1+\Phi_t^2 U_t^{\perp})$. The normalized residual is

$$e_t^{\mathrm{law}} = \frac{|q_t^{\mathrm{obs}} - q_t^{\mathrm{pred}}|}{\max\{|q_t^{\mathrm{obs}}|, |q_t^{\mathrm{pred}}|\}}.$$

Both diagnostics vanish under exact tangent updates in real arithmetic. Across all blocks and updates, their maxima are 6.97×10^{-3} and 4.27×10^{-4}, respectively. These values measure local departures from the ideal identities, rather than accumulated trajectory error. All results can be seen in Figure 7.

F.3 Extended CIFAR comparison and language model experiments

Vision model: Table 1 reports a separate expanded grid covering both datasets, all three ResNet architectures, and both training budgets. Each of the 12 settings has three schedule arms and five paired seeds, numbered 0–4, for 180 endpoints. The saved configurations specify batch size 128, cosine learning rate $0.5 \to 0.05$, initial or constant decay 0.001, and evaluation every 500 updates at the short horizon and every 782 updates at the long horizon. The sequential controller uses the same requested ramp and model-specific bounds described below.

For each setting we subtract constant-decay accuracy from narrow-ramp accuracy within seed. The earlier development sweep in Appendix F.5 and the scope follow-up in Appendix F.6 are separate runs and should not be pooled with this grid.

Language model: We train two randomly initialized GPT-2 models (Radford et al., 2019) on WikiText-103 (Merity et al., 2016) using the GPT-2 tokenizer: a smaller model with 4 layers, 4 attention heads, and hidden width 256, and GPT-2 with 12 layers, 12 heads, and hidden width 768. Both retain affine LayerNorm, biases, tied input and output embeddings, and dropout with probability 0.1. Training uses next-token cross-entropy, sequences of 256 tokens, and minibatches of 8 sequences for 58,650 updates. All strategies use momentum-free SGD with a common cosine learning-rate schedule decreasing from 0.005 toward 0.0005, without warmup. The constant-decay reference uses $\lambda = 0.001$. The controlled schedules request either $B_t = 1$ or a linear ramp from 1.002 to 0.998, starting from $\lambda_0 = 0.001$, with subsequent decay coefficients clipped to $[0, 0.01]$. Decay applies to all trainable parameters. Consequently, realized forcing can differ from the requested trajectory. Validation loss is evaluated every 500 updates and at the final update. Curves show arithmetic means over seeds 0, 1, and 2, without smoothing. These conventional GPT-2 architectures are not assumed to be exactly scale-invariant.

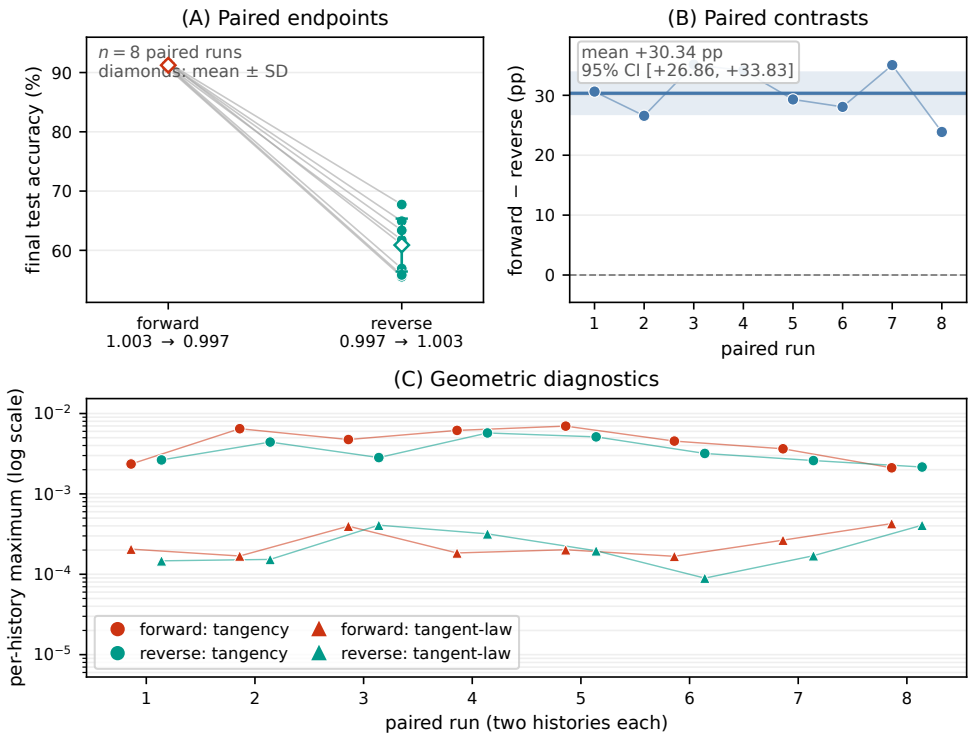

Figure 7: **Endpoint contrasts and geometry in the exact-order intervention.** Paired final accuracies, paired differences, and per-history maximum geometry residuals. The residuals measure finite-precision departures from the model's real-arithmetic scale symmetry.

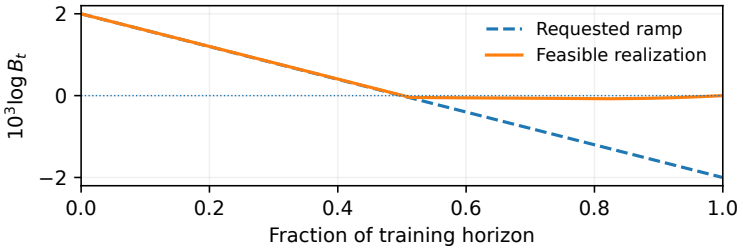

Figure 8: **Native constraints change the requested trajectory.** At 58,650 updates, the narrow-ramp request and its nonnegative-decay realization differ substantially in late forcing. Only 5.76% of the requested negative log-action survives. Curves show the requested and realized schedule arrays.

F.4 Reachability and Realization

As native constraints can clip the requested forcing, we report the realized paths to establish how closely each schedule implements its intended behavior (Appendix F.5). At the longer horizon, the narrow ramp retains only 5.76% of its requested negative log-action. The effective intervention is therefore forcing above one early in training, followed by contraction limited by the nonnegativedecay constraint, which is visible only after reconstructing the realized forcing path (Figure 8). This is the practical role of reachability: it identifies the schedule actually being compared.

F.5 Sensitivity to Forcing-Ramp Width

The practical models use torchvision ResNet-18, ResNet-34, and ResNet-50 (He et al., 2016) with standard affine BatchNorm. Models are trained from scratch on the official 50,000 training examples

and evaluated on the official 10,000 test examples with no momentum, batch size set to 128, the positive-floor cosine path is $0.5 \to 0.05$, and baseline coupled decay is 0.001.

For a requested $\widetilde{B}_t$, sequential synthesis clips the implied next decay coefficient,

$$\lambda_{t+1} = \Pi_{[0,\lambda_{\max}]}\left[\frac{1}{\eta_{t+1}}\left(1 - \frac{\eta_{t+1}}{\widetilde{B}_t \eta_t (1 - \eta_t \lambda_t)}\right)\right], \qquad \lambda_0 = 0.001. \tag{132}$$

The bounds are imposed on the native coefficient, so the realized forcing can differ from the request.

The five arms are cosine plus constant decay, a clipped $B = 1$ request, and clipped linear requests with $\delta \in \{0.002, 0.003, 0.005\}$. The upper decay cap is 0.02 in the ResNet-18/34 development records and 0.1 in the ResNet-50 records, with lower bound zero. The upper cap is inactive for these ramp requests; the lower bound is not. Consequently the differing caps do not change the ramp paths. Table 4 includes every tested setting and width. There are five paired runs per setting and arm, with the seed assignments supplied in the experiment configurations. The short horizon is 10,000 updates and the long horizon is 58,650 updates, approximately 150 epoch-equivalents of with-replacement sampling, not 150 shuffled passes.

Table 4: **Complete width sweep.** Final test accuracy in percent, mean±sample SD over five matched runs. C10/C100 denote CIFAR-10/100. Every column uses the same cosine learning-rate path. The width comparisons are development results, not multiplicity-corrected or held-out model-selection results.

Model / data	Updates	Cosine	Clipped $B = 1$	$\delta = .002$	$\delta = .003$	$\delta = .005$
R18 / C100	10k	71.86 ± 0.42	67.79 ± 1.31	72.64 ± 0.18	72.21 ± 0.51	71.52 ± 0.54
R34 / C10	10k	91.36 ± 0.68	87.54 ± 1.23	92.50 ± 0.33	92.56 ± 0.25	92.41 ± 0.16
R50 / C10	10k	90.36 ± 0.69	86.57 ± 0.93	91.06 ± 0.83	91.04 ± 0.39	90.71 ± 0.14
R50 / C100	10k	67.03 ± 1.85	61.50 ± 2.00	67.64 ± 2.55	66.64 ± 1.49	62.00 ± 1.21
R50 / C10	58.65k	92.17 ± 1.10	91.55 ± 1.09	94.65 ± 0.37	94.39 ± 0.19	93.77 ± 0.16
R50 / C100	58.65k	72.04 ± 0.79	71.64 ± 1.62	76.38 ± 0.61	75.11 ± 1.43	71.85 ± 2.11

The narrow request beats cosine and the clipped $B = 1$ across different settings. The widest request loses to cosine in three setting means, including 5.03 points on short-horizon CIFAR-100/ResNet-50. The long-horizon narrow gains are 2.49 points on CIFAR-10 and 4.34 on CIFAR-100. At 58,650 updates, the narrow ramp sets decay to zero for 61.66% of updates and retains 5.76% of the requested negative log-action. Figure 8 shows the resulting change in late forcing.

F.6 Hyperparameter selection: scope, tails, and action

The follow-up is a separate execution of the ResNet-50/CIFAR-100 58,650-update protocol, selected because that setting had the largest breadth-study gain. It comprises six arms with five paired runs each. Within each seed, model initialization, complete minibatch and augmentation streams, cosine learning rates, and deterministic execution are matched.

Stage-one scope selection. Four arms rerun the constant-decay baseline and whole-model ramp, then apply the same ramp coefficients either to all convolutional kernels or only to their complement. The convolutional group accounts for 98.91% of parameters and the complement contains all remaining trainable parameters and has 1.09% of scalars. Non-scheduled parameters retain decay 0.001.

The specified rule selects convolution-only version for stage-two if the whole-model mean gain is positive, convolution-only version retains at least half that gain, and at least four of five convolution-only differences are positive. Otherwise it selects whole-model version. All three conditions hold, so stage two uses convolutional scope. Both candidate stage-two schedule arrays were fixed before this selection; only scope was selected.

Stage-two forcing design. The benchmark-tail ($\beta = 1$) request uses a rational one-crossing prefix and then $\log B_t = -1/(t+2)$ after the midpoint. The splice ensures an in-horizon return toward one.

Its $\beta = 1$ tail is inspired by the alignment benchmark, rather than the stochastic boundary $\beta = 1/2$, and it is not the exact scalar-optimal policy. The horizon-normalized request is

$$\log B_t = \frac{\kappa}{T}\left(1 - \frac{2t}{T-1}\right), \qquad \kappa = 19.988. \tag{133}$$

Schedule-only bisection calibrates κ to the earlier 10k ramp's realized absolute log-action. Both requests are globally projected under the nonnegative-decay bounds, whereas the linear ramp retains its sequential construction.

Table 5: **Complete focused follow-up.** Mean±sample SD over five paired final-only endpoints. The last two arms use the scope selected by the specified stage-one rule.

Schedule	Scheduled parameters	Final accuracy (%)
Constant decay	All	70.43 ± 2.21
Linear ramp	All	76.48 ± 0.57
Linear ramp	Convolutional kernels	77.01 ± 0.76
Linear ramp	Complementary parameters	72.42 ± 0.40
Benchmark-tail request	Convolutional kernels	75.87 ± 0.62
Horizon-normalized projection	Convolutional kernels	76.58 ± 0.69

Table 6: **Realized forcing and native decay budgets.** All sums use operative transitions, except native dose, which uses actual parameter updates. "Negative action retained" compares realized with requested absolute log-action below one. Forcing action and native decay dose are not interchangeable.

Realized path	$\sum \lvert \log B \rvert$	$\sum \log B$	Negative action retained	$\sum \eta\lambda$
Linear ramp	30.9967	27.6162	5.76%	14.9522
Benchmark-tail request	6.1789	4.8159	98.33%	3.5586
Normalized projection	6.6023	3.3921	32.12%	2.8473

The benchmark-tail and normalized paths have maximum stored projection KKT residuals 8.46×10^{-11} and 2.22×10^{-16}, with zero primal feasibility violation. Their log-forcing root-mean-square request errors are 1.50×10^{-6} and 1.08×10^{-4}, versus 7.83×10^{-4} for the linear ramp. The normalized projection retains only 32.12% of requested negative action. It tests a feasible low-action schedule rather than an exactly action-matched change in shape. Its absolute action 6.60234 approximately matches the calibration reference 6.60260.

Relative to the selected-scope ramp, the normalized path uses 78.70% less absolute log-action and 80.96% less native dose. It retains 93.41% of the point-estimate baseline gain. The benchmark-tail path is 1.140 points below the ramp on average. Secondary baseline contrasts are $+5.442$ points for the benchmark-tail path and $+6.148$ for the normalized path. These are descriptive, multiple comparisons within the selected setting.

The focused baseline, 70.428%, is lower and more variable than the earlier breadth baseline, 72.042%. The corresponding whole-model ramp means are 76.482% and 76.378%. The suites are kept separate because their execution and provenance differ, despite the reused seed identifiers. Convolution-only actuation suffices for the observed gain in this protocol.