FROM LEARNING RATES TO ROTATIONAL SCHEDULES: FINITE-HORIZON THEORY ON THE SPHERE

Anonymous authors
Paper under double-blind review

ABSTRACT

How should step sizes be scheduled across a finite training horizon? Cosine and warmup-stable-decay schedules are widely adopted in large-scale pretraining, yet existing convergence theory for such schedules is largely Euclidean and does not capture the rotational dynamics of normalized neural networks. In this work, we formulate schedule design directly in the space of rotational motion. Working within a geodesically convex spherical setting, we derive a last-iterate convergence guarantee for arbitrary *rotational schedules*, i.e. sequences of angular step sizes, that separates the sum of the scheduled step sizes from their distribution over time. We recover the same finite-horizon schedule-dependency as in Euclidean theory, and we solve it globally: the bound-optimal schedule is unique, and approaches linear decay as the horizon grows. Finally, we connect this result to Hyperball (Wen et al., 2026), where fixed parameter norm and normalized updates make the learning rate act directly as the rotational schedule. We find that our theoretical rotational bound predicts schedule performance in practice, evaluated on language model pretraining runs optimized with Hyperball. Consistent with the bound prediction, we find that cosine and linear schedules achieve comparable performance, both ahead of warmup-stable-decay and constant schedules, thus providing practical insight into the schedule design of Hyperball-style optimizers.

1 INTRODUCTION

Recent advances in language-model pretraining make parameter magnitudes and directions explicit design choices. nGPT incorporates hyperspherical normalization into the Transformer architecture (Loshchilov et al., 2025), Muown and Magnitude-Direction Decoupling (MDD) separate row magnitudes from weight directions inside the optimizer (Lion et al., 2026; Hägele et al., 2026), and Hyperball constrains matrix norms and normalizes updates (Wen et al., 2026). These methods are now also permeating into large-scale training recipes (LLM-Core Xiaomi, 2026). Such trends make studying rotational dynamics a timely practical concern.

Rotational dynamics are intertwined with the geometric properties of normalized networks (Ioffe & Szegedy, 2015; Ba et al., 2016; van Laarhoven, 2017; Hoffer et al., 2018), which exhibit scale invariant properties. Under scale invariance, multiplying a parameter group by a positive scalar leaves the network's function unchanged. Optimization of such groups can therefore be described exactly through rotations of their parameters on the unit sphere (Arora et al., 2018; Li & Arora, 2020; Wan et al., 2021; Roburin et al., 2022; Kodryan et al., 2022). Yet, parameter norms still affect these dynamics: a larger norm reduces the angular motion of the optimization (van Laarhoven, 2017; Kosson et al., 2024). The effective rotation speed per step is therefore determined jointly by the learning rate and parameter norm.

Hyperball makes this rotational motion directly controllable by fixing parameter norms and normalizing updates; MDD and AngularMuown similarly constrain directional factors to spheres while learning magnitudes separately (Wen et al., 2026; Hägele et al., 2026; Hübler et al., 2026). Recent empirical studies further demonstrate that matching or altering effective rotational schedules governs loss trajectories far more decisively than learning rates (Liu et al., 2026; Yuan et al., 2026).

Explicit rotational control therefore makes schedule design central: as Hägele et al. (2026) highlight, the *optimal schedule on the sphere* remains an open and practically relevant question.

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

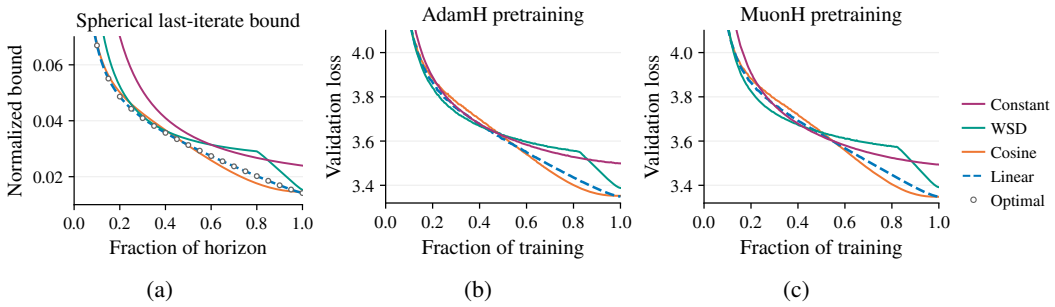


Figure 1: The exact bound-optimal schedule profile approaches linear decay. **(a)** Spherical last-iterate bound at $T = 20000$, with each amplitude optimized separately; circles mark the exact finite-horizon optimum. **(b, c)** Validation loss over 20000 updates for 124M Llama/FineWeb with AdamH and MuonH. Linear and cosine are comparable, both ahead of WSD and constant. Across both optimizers, the bound ranking matches the measured loss ordering.

Taken together, these developments motivate shifting the schedule-design from LRs to rotational displacements:

If scale-invariant optimization is intrinsically rotational,
how should rotational motion be distributed over a finite training horizon?

Scheduling of rotational motion is still lacking a comprehensive theoretical and empirical literature. On the contrary, the LR scheduling problem has been a long-standing theoretical and practical concern of the deep learning community (Loshchilov & Hutter, 2017; Defazio et al., 2023; Hu et al., 2024; Schaipp et al., 2025). We discuss the related work in detail in App. A.

Recent convex last-iterate analysis offer particularly valuable results: finite-horizon guarantees naturally select decaying profiles (Defazio et al., 2023; Schaipp et al., 2025). These analyses shed light on practical schedules such as cosine (Loshchilov & Hutter, 2017), linear decay (Defazio et al., 2023; Bergsma et al., 2025), and warmup-stable-decay (WSD) (Hu et al., 2024; Hägele et al., 2024; Schaipp et al., 2025), for vanilla SGD and for Euclidean-convex losses.

However, optimization of scale-invariant networks is intrinsically non-Euclidean: only trivially constant functions are both Euclidean-convex and scale-invariant (App. D). Furthermore, scale-invariant optimization and Hyperball-style optimizers are described by normalized tangent updates, and not by Euclidean, additive SGD steps. The existing literature is therefore insufficient to analyze the rotational scheduling problem on the sphere. In this work, instead, we formulate the problem on a geodesically convex spherical cap and derive a new progress inequality for normalized tangent updates. We then apply the weighted-tail procedure of Defazio et al. (2023), and we achieve a last-iterate bound for arbitrary positive deterministic rotational schedules that separates total coefficient mass (*how much* rotation is applied) from its temporal distribution (*when* it is applied). We find that the resulting time-distribution cost C_T matches the Euclidean one found by Schaipp et al. (2025), which however they only minimize numerically. Instead, we propose an analytic solution of the optimal-schedule problem, finding a unique bound-optimal schedule that follows an exact recursion and approaches linear decay as the horizon grows. The amplitude-optimized bound places cosine close to the exact optimum, whereas WSD and constant profiles have higher costs (Fig. 1a). To test whether the ranking transfers to practice, we use Hyperball (Wen et al., 2026), whose fixed parameter norm and normalized update make the learning rate equal the rotational coefficient under tangency. In 124M Llama/FineWeb pretraining, separately tuned linear and cosine profiles finish comparably and ahead of the tested WSD and constant profiles (Fig. 1b, c), in agreement with the bound simulation.

Our prescription is to decay rotational coefficients approximately linearly over the horizon. Under Hyperball, this translates directly into learning-rate decay; without the fixed parameter norm, evolving parameter and update norms change the induced rotational profile. In the corresponding non-Hyperball Adam and Muon runs, WSD becomes competitive with linear and cosine. Our bound is able to predict the observed loss ordering in *both* settings (Fig. 3), thanks to the scale-invariant construction absent in previous Euclidean analyses of Schaipp et al. (2025)

Our Contributions. We derive a last-iterate bound for geodesically convex losses for rotational schedules on the sphere, solve its minimization problem globally, and test the resulting ranking in pretraining:

- **Scheduling on the sphere.** We derive a new progress inequality for normalized tangent updates under geodesic convexity, then obtain a last-iterate guarantee for positive deterministic rotational schedules.
- **Optimal rotational scheduling.** The bound contains the same scheduling functional as the Euclidean last-iterate result of Schaipp et al. (2025). We solve it globally over positive schedules, obtain the unique finite-horizon recursion, and prove its linear-decay limit.
- **Rotational scheduling in pretraining.** Measured angle agreement supports the predicted connection between Hyperball's learning rate and rotation. Separately tuned 124M Llama/FineWeb runs favor linear and cosine under Hyperball. For unconstrained Adam and Muon, evaluation on measured rotational profiles recovers the inverted WSD–cosine ordering observed in practice.

2 From Learning-Rate Schedules to Rotational Schedules

We start analyzing gradient descent on a scale-invariant loss to build the geometric picture behind the paper's main construction. We consider one nonzero parameter group $\boldsymbol{x} \in \mathbb{R}^C$, $C \geq 2$, and write $L(\boldsymbol{x})$ with the other parameters held fixed. Under scale invariance, the loss depends only on direction. Yet, the parameter norm still matters: together with the learning rate, it determines the *rotational coefficient* s_t^{GD} which controls the effective step size on the sphere(van Laarhoven, 2017; Li & Arora, 2020; Roburin et al., 2022).

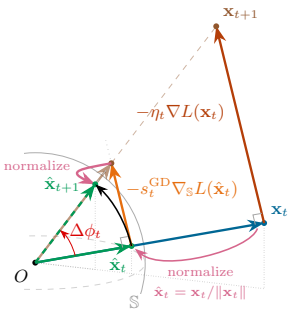


Figure 2: A gradient descent step on the sphere. The learning rate and radius determine $s_t^{\mathrm{GD}} = \eta_t / \|\boldsymbol{x}_t\|^2$.

2.1 Rotational Dynamics under Scale Invariance

We first recall scale invariance, which in neural networks can arise from normalization layers that cancel the scaling of a parameter group (van Laarhoven, 2017; Kosson et al., 2024). Although standard Transformers are not generally exactly scale invariant, prior work shows that scale-invariant models can usefully approximate their norm and rotational dynamics (Kosson et al., 2024; Defazio, 2025; Wen et al., 2026). We adopt this approximation and show in Section 4 that the resulting theoretical bounds provide useful predictions of LLM pretraining behavior.

Definition 2.1 (Scale-invariant parameter group)**.** *A parameter group $\boldsymbol{x}$ is* scale-invariant *for a differentiable loss L if $L(\kappa \boldsymbol{x}) = L(\boldsymbol{x})$ for every $\kappa > 0$. Write $\hat{\boldsymbol{x}} := \boldsymbol{x} / \|\boldsymbol{x}\|$ for its unit direction and $\mathbb{S}$ for the unit sphere in the parameter space.*

The immediate geometric consequence is that $L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}})$: the loss depends on the parameter only through its direction on the sphere. It is therefore sensible to measure the size of an update by the angle between successive directions,

$$\Delta\phi_t := \angle(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = \angle(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_{t+1}) = \arccos\big(\langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_{t+1} \rangle / (\|\hat{\boldsymbol{x}}_t\| \, \|\hat{\boldsymbol{x}}_{t+1}\|)\big).$$

To describe how updates move along the sphere, we need the spherical gradient $\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}})$: the tangent vector at $\hat{\boldsymbol{x}}$ (perpendicular to $\hat{\boldsymbol{x}}$, since $T_{\hat{\boldsymbol{x}}}\mathbb{S} = \{\boldsymbol{v} : \langle \boldsymbol{v}, \hat{\boldsymbol{x}} \rangle = 0\}$) that gives the loss derivative along each tangent direction. Since L is constant along rays, $\langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \boldsymbol{x} \rangle = 0$, so the raw gradient is already tangent and relates to the spherical gradient by $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-1} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}})$ (App. B.1). Substituting into gradient descent and factoring out the radius gives

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla_{\boldsymbol{x}} L(\boldsymbol{x}_t) = \|\boldsymbol{x}_t\| \left(\hat{\boldsymbol{x}}_t - \frac{\eta_t}{\|\boldsymbol{x}_t\|^2} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t) \right). \tag{2.1}$$

Proposition 2.2 (Exact rotational recursion for gradient descent)**.** *For the scale-invariant loss of Definition 2.1, gradient descent with $\eta_t > 0$ induces the rotational coefficient and unit-direction recursion on the sphere*

$$s_t^{\mathrm{GD}} := \frac{\eta_t}{\|\boldsymbol{x}_t\|^2}, \qquad \hat{\boldsymbol{x}}_{t+1} = \frac{\hat{\boldsymbol{x}}_t - s_t^{\mathrm{GD}} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t)}{\sqrt{1 + (s_t^{\mathrm{GD}})^2 \|\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t)\|^2}}. \tag{2.2}$$

The realized turning angle (Kosson et al., 2024) is

$$\Delta\phi_t := \angle(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = \arctan\left(s_t^{\mathrm{GD}} \|\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t)\|\right). \tag{2.3}$$

When $\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t) \neq 0$, the same update can be written as

$$\hat{\boldsymbol{x}}_{t+1} = \cos(\Delta\phi_t)\hat{\boldsymbol{x}}_t - \sin(\Delta\phi_t)\frac{\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t)}{\|\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t)\|}. \tag{2.4}$$

When the gradient is zero, $\Delta\phi_t = 0$ and $\hat{\boldsymbol{x}}_{t+1} = \hat{\boldsymbol{x}}_t$.

The negative spherical-gradient direction determines where the step goes on the sphere, while the angle determines how far it goes. The rotational coefficient s_t^{GD} controls this angle jointly with the spherical-gradient norm. Fig. 2 shows the geometric construction.

Remark. The quantity s_t^{GD} is often called the *effective learning rate* (van Laarhoven, 2017). We instead call it the *rotational coefficient*: for a scale-invariant parameter group, it multiplies the spherical gradient in the on-sphere recursion, while the learning rate is only *one* factor determining it. We call the sequence $s_{1:T}$ the *rotational coefficient schedule*. The realized *rotational schedule* is the sequence of turning angles $\Delta\phi_{1:T}$; when the distinction is clear from context, we call $s_{1:T}$ the rotational schedule for brevity.

Takeaway

Under scale invariance, the parameter rotation is not determined by the learning rate alone: their norm also matters. For gradient descent, the rotational coefficient is $s_t^{\mathrm{GD}} = \eta_t / \|\boldsymbol{x}_t\|^2$.

2.2 A Last-Iterate Bound for Rotational Schedules

The GD rotational coefficient $s_t^{\mathrm{GD}} = \eta_t / \|\boldsymbol{x}_t\|^2$ depends on the evolving radius, so a learning-rate schedule does not directly control a rotational schedule. To study how these coefficients should vary over time independently of norm dynamics, we prescribe a positive deterministic coefficient sequence $s_{1:T} = (s_1, \ldots, s_T)$, where a fresh stochastic tangent oracle supplies the update directions and magnitudes. The main guarantee of Schaipp et al. (2025) assumes convex losses and additive SGD updates. However, these assumptions are incompatible with non-trivial scale-invariant losses: a finite Euclidean-convex population loss trivializes to a constant if assumed to be scale-invariant (App. D). Furthermore, their additive update does not directly cover the normalized tangent step of (2.2). To address these limitations, we develop a new last-iterate bound for arbitrary rotational schedules on a geodesically convex spherical cap.

Let us fix a spherical cap centered at a minimizer $\hat{\boldsymbol{x}}^\star$. Using angular distance $d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) := \arccos\langle \hat{\boldsymbol{x}}, \boldsymbol{y} \rangle$, let us define

$$\mathcal{C} := \{\hat{\boldsymbol{x}} \in \mathbb{S} : d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star) < \theta_{\mathrm{cap}}\}, \quad 0 < \theta_{\mathrm{cap}} < \frac{\pi}{4}. \tag{2.5}$$

We assume that L is *geodesically convex* on $\mathcal{C}$: for any two points in the cap, the loss along the shortest path connecting them on the sphere satisfies the usual convexity inequality (Boumal, 2023, Definition 11.3). The cap geometry and formal definition are in App. B.2.

At the current iterate $\hat{\boldsymbol{x}}_t$, let a stochastic oracle supply a fresh tangent vector $\tilde{\boldsymbol{g}}_t \in T_{\hat{\boldsymbol{x}}_t}\mathbb{S}$ whose conditional mean is the spherical gradient. With the prescribed coefficient s_t, the normalized tangent-step construction of (2.2) gives the iteration

$$\hat{\boldsymbol{x}}_{t+1} = \frac{\hat{\boldsymbol{x}}_t - s_t \tilde{\boldsymbol{g}}_t}{\sqrt{1 + s_t^2 \|\tilde{\boldsymbol{g}}_t\|^2}}. \tag{2.6}$$

This is Riemannian SGD with the projection retraction (Bonnabel, 2013, Sec. 2.3).

Let $D := d_{\mathbb{S}}(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}^\star)$. For the prescribed positive coefficient schedule, define the coefficient sum M and normalized temporal allocation $\omega_{1:T}$ by

$$M := \sum_{t=1}^{T} s_t, \qquad \omega_t := \frac{s_t}{M}, \qquad s_t = M\omega_t, \qquad \omega_t > 0, \quad \sum_{t=1}^{T} \omega_t = 1. \tag{2.7}$$

Thus M sets the overall schedule scale, while ω_t records the fraction assigned to step t.

Theorem 2.3 (Last-iterate bound for rotational coefficient schedules)**.** *Let L be differentiable and geodesically convex on $\mathcal{C}$, with minimizer $\hat{\boldsymbol{x}}^\star$. Consider* (2.6) *from deterministic $\hat{\boldsymbol{x}}_1 \in \mathcal{C}$, with positive deterministic $s_1, \ldots, s_T$ fixed independently of the oracle draws. Assume all iterates through $\hat{\boldsymbol{x}}_{T+1}$ remain in $\mathcal{C}$. With $\mathcal{F}_t$ the history before draw t, suppose the tangent oracle satisfies, for a common deterministic $G > 0$,*

$$\langle \tilde{\boldsymbol{g}}_t, \hat{\boldsymbol{x}}_t \rangle = 0, \qquad \mathbb{E}[\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_t] = \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t), \tag{2.8}$$

$$\mathbb{E}\,\|\tilde{\boldsymbol{g}}_t\|^2 \le G^2 \qquad (1 \le t \le T). \tag{2.9}$$

Then, for every integer $T \ge 1$, the expected last-iterate excess loss satisfies

$$\mathbb{E}[L(\hat{\boldsymbol{x}}_T) - L(\hat{\boldsymbol{x}}^\star)] \le \frac{D^2}{2M} + \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2} \left[\omega_T + \sum_{k=1}^{T-1} \frac{\omega_k^2}{\sum_{t=k+1}^{T} \omega_t} \right]. \tag{2.10}$$

Proof sketch. We first derive a spherical analogue of the standard one-step SGD progress inequality, starting from the distance comparison of Zhang & Sra (2016, Corollary 8) for exponential-map steps. Geodesic convexity replaces Euclidean convexity, and the normalization of the stochastic tangent update can be absorbed at the cost of a factor $1 + \theta_{\mathrm{cap}}$. We then apply the weighted-tail last-iterate argument of Defazio et al. (2023) and Schaipp et al. (2025): summing the one-step inequality over successive tails and telescoping the corresponding tail averages isolates the final iterate. Collecting these tail contributions yields the schedule-dependent term in (2.10). The complete proof is in Apps. B.3 and B.4. □

The bound trades off two independent terms with competing dependence on the total rotational mass M: the first term $D^2/(2M)$ shrinks as more rotation is used, while the second grows with M and with the oracle variance. We write the schedule-dependent part of the second term as a standalone cost. Define the scale-free allocation cost for a positive unit-mass allocation $\omega_{1:T} = (\omega_1, \ldots, \omega_T)$ by

$$C_T(\omega_{1:T}) := \omega_T + \sum_{k=1}^{T-1} \frac{\omega_k^2}{\sum_{t=k+1}^{T} \omega_t}. \tag{2.11}$$

Each executed fraction ω_k contributes its square divided by the allocation remaining after that step, so placing a large fraction at a time with little future allocation is expensive. With this definition, Theorem 2.3 becomes

$$\Omega_T(M, \omega_{1:T}) := \frac{D^2}{2M} + \frac{(1+\theta_{\mathrm{cap}})G^2}{2} M C_T(\omega_{1:T}). \tag{2.12}$$

We note that the resulting allocation cost C_T is identical to the functional of Schaipp et al. (2025, Corollary 3.3), despite using different geometry, assumptions, and update rules. The bound has the form $D^2/(2M) + \mathrm{const} \cdot M\, C_T(\omega_{1:T})$, so optimizing over M at fixed allocation yields a minimum value proportional to $\sqrt{C_T(\omega_{1:T})}$; minimizing over profiles therefore reduces to minimizing C_T alone.

Proposition 2.4 (Optimal total mass for a fixed allocation)**.** *For $D, G > 0$ and a positive unit-mass allocation $\omega_{1:T}$, the unique minimizing mass and resulting bound are*

$$M^\star(\omega_{1:T}) = \frac{D}{G\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T})}}, \qquad \min_{M>0} \Omega_T(M, \omega_{1:T}) = DG\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T})}. \tag{2.13}$$

This is the amplitude calculation of Schaipp et al. (2025, Corollary 3.3); App. B.5 gives the proof. After optimizing amplitude, the remaining schedule-design problem is exactly to minimize $C_T(\omega_{1:T})$.

Takeaway

After optimizing the overall schedule scale M, schedule design reduces to minimizing a single allocation cost $C_T(\omega_{1:T})$.

3 The Bound-Optimal Rotational Schedule

After optimizing total mass, the remaining schedule-design problem is to minimize the allocation cost $C_T(\omega_{1:T})$: which temporal shape makes the best use of a given rotational budget? Finite-time convex-optimization bounds commonly make their dependence on the stepsize sequence explicit (Garrigos & Gower, 2024, Thm. 6.8). For the spherical last-iterate bound derived above, schedule design reduces to minimizing C_T. We solve this problem globally: the unique finite-horizon optimum follows an exact recursion, and its profile approaches linear decay as the training horizon grows.

3.1 Constructing the Unique Finite-Horizon Optimum

Write C_n for the allocation cost (2.11) at horizon $n \geq 1$, and let

$$C_n^\star := \inf_{\substack{\omega_t > 0 \\ \sum_{t=1}^n \omega_t = 1}} C_n(\omega_{1:n}) \tag{3.1}$$

denote its infimum over positive unit-mass allocations. For $n \geq 2$, first choose the fraction ω_1, then describe how the remaining mass is distributed. The following lemma makes this decomposition explicit.

Lemma 3.1 (First-step decomposition)**.** *For $n \geq 2$ and every positive unit-mass allocation $\omega_{1:n} = (\omega_1, \ldots, \omega_n)$, define the remaining proportions $p_t := \omega_t/(1-\omega_1)$ for $2 \leq t \leq n$. These proportions are positive and satisfy $\sum_{t=2}^n p_t = 1$. Then*

$$C_n(\omega_{1:n}) = \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1)\, C_{n-1}(p_{2:n}). \tag{3.2}$$

Proof. Separate ω_1's contribution in (2.11) and substitute $\omega_t = (1-\omega_1)p_t$. Each later ratio $\omega_k^2/\sum_{t>k}\omega_t$ has $(1-\omega_1)^2$ in its numerator and $(1-\omega_1)$ in its denominator; factoring the remaining $(1-\omega_1)$ from every term, including the terminal entry, gives $(1-\omega_1)\, C_{n-1}(p_{2:n})$. □

For fixed ω_1, taking the infimum over normalized tails gives $C_{n-1}^\star$. The remaining scalar problem is

$$C_n^\star = \inf_{0<\omega_1<1} \left\{ \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1) C_{n-1}^\star \right\}, \qquad n \geq 2. \tag{3.3}$$

The objective in (3.3) is strictly convex in ω_1, so each step has a unique closed-form minimizer; iterating from $C_1^\star = 1$ gives both the optimal cost and the optimal allocation.

Theorem 3.2 (Optimal cost and uniqueness)**.** *For every integer $n \geq 1$, $C_n^\star > 0$ is attained by a unique positive unit-mass allocation and satisfies*

$$C_1^\star = 1, \qquad C_n^\star = 2\big(\sqrt{1 + C_{n-1}^\star} - 1\big) \quad (n \geq 2). \tag{3.4}$$

Proof sketch. Lemma 3.1 reduces the optimization to (3.3). Solving this scalar problem and applying induction proves the cost recursion, attainment, and uniqueness. App. C.1 gives the full proof. □

The theorem gives the minimum cost and shows that exactly one allocation achieves it. The following corollary gives a recursion for computing each entry of that allocation.

Corollary 3.3 (Construction of the optimal allocation)**.** *For every integer $T \geq 1$, the unique optimal allocation $\omega^\star_{1:T}$ is constructed using the costs from Theorem 3.2 as*

$$\omega^\star_k = \left(1 - \sum_{j=1}^{k-1} \omega^\star_j\right)\left(1 - \frac{1}{\sqrt{1 + C^\star_{T-k}}}\right), \quad 1 \leq k < T, \qquad \omega^\star_T = 1 - \sum_{j=1}^{T-1} \omega^\star_j. \tag{3.5}$$

Proof sketch. Each normalized optimal tail solves the shorter allocation problem. Applying the minimizing first fraction found in the theorem's proof and rescaling by the remaining mass gives the construction. App. C.1 gives the full proof. □

This construction also provides an analytic solution to the constant-envelope functional optimized numerically by Schaipp et al. (2025, App. C.3) for the Euclidean case. Together with Proposition 2.4, the optimal allocation determines the unique bound-optimal coefficient schedule $s^\star_t = M^\star(\omega^\star_{1:T})\omega^\star_t$ for $D, G > 0$.

3.2 The Optimal Schedule Converges to Linear Decay

The recursion in Corollary 3.3 gives the exact optimizer at each finite horizon; we now ask what shape it approaches as the horizon grows. For a fraction $0 \leq \tau < 1$ of training completed, write $\omega^\star_{1+\lfloor \tau(T-1) \rfloor}$ for the coefficient at that position.

Proposition 3.4 (Asymptotic cost and linear allocation)**.** *As $T \to \infty$, the optimal cost decays as $C^\star_T \sim 4/T$. At every fixed training fraction $0 \leq \tau < 1$, the corresponding coefficient satisfies*

$$\omega^\star_{1+\lfloor \tau(T-1) \rfloor} \sim \frac{2(1-\tau)}{T}. \tag{3.6}$$

Proof sketch. The cost recursion gives reciprocal increments $1/C^\star_n - 1/C^\star_{n-1} = 1/(4 + C^\star_n) \to 1/4$; Stolz–Cesàro yields $C^\star_n \sim 4/n$. The allocation recursion determines each coefficient as a fraction of the remaining mass $W^\star_k = \sum_{t \geq k} \omega^\star_t$. The cost asymptotic turns the accumulated log-retained fractions into a Riemann sum whose limit gives $W^\star_k \to (1-\tau)^2$ and hence $\omega^\star_k \sim 2(1-\tau)/T$. App. C.2 gives the full proof. □

Since the optimum approaches linear decay, a natural question is whether the closed-form linear schedule itself achieves the same leading bound value. Define its positive unit-mass allocation by $\omega^{\text{lin}}_t := 2(T+1-t)/[T(T+1)]$ for $1 \leq t \leq T$. The exact calculation in App. C.3 gives

$$C_T(\omega^{\text{lin}}_{1:T}) = \frac{4}{T} + O\left(\frac{\log T}{T^2}\right). \tag{3.7}$$

After optimizing the amplitude of each schedule, (2.13) yields

$$\frac{\min_{M>0} \Omega_T(M, \omega^{\text{lin}}_{1:T})}{\min_{M>0,\, \omega_{1:T}} \Omega_T(M, \omega_{1:T})} = \sqrt{\frac{C_T(\omega^{\text{lin}}_{1:T})}{C^\star_T}} \longrightarrow 1. \tag{3.8}$$

For large horizons, the bound penalty for replacing the exact finite-horizon optimum with the closed-form linear schedule vanishes.

Fig. 6 simulates this asymptotic result at a finite horizon. Each curve evaluates the bound from Theorem 2.3 with one allocation and its own optimal amplitude from Proposition 2.4, both chosen for $T = 20000$; since the amplitude is tuned separately for each schedule, the ordering among curves is determined entirely by the allocation cost $C_T(\omega_{1:T})$. The near-coincidence of the exact optimum with linear decay shows a small bound penalty at this horizon. Cosine is also close, while the specified WSD plateau and constant profile incur larger penalties. The gray envelope reoptimizes both schedule and amplitude at every displayed index, so its gap from a fixed-horizon curve measures the cost of committing to $T = 20000$ and evaluating earlier. App. C.4 gives the full calculation and extends the comparison across horizons and amplitudes.

Takeaway

Optimizing analytically over all positive rotational coefficient schedules yields a unique exact finite-horizon solution. Its profile approaches linear decay as the training horizon increases.

4 Rotational Scheduling with Hyperball

The bound schedules a rotational coefficient. We show that Hyperball (Wen et al., 2026) approximately exposes that coefficient as a direct control by fixing each matrix radius and normalizing its update.

4.1 Hyperball as a Rotational Control

Recent spherical optimizers expose the learning rate as a control on directional motion, but its best schedule is unknown (Hägele et al., 2026; Hübler et al., 2026; Xiao et al., 2026). A prominent example is Hyperball, which fixes each matrix radius and normalizes its update. For a matrix $\mathbf{X}_t$ with $R = \|\mathbf{X}_1\|_F > 0$, write $\mathbf{V}_t := \mathbf{U}_t/\|\mathbf{U}_t\|_F$ for the normalized base-optimizer update. The update rule is

$$\mathbf{X}_{t+1} = R\frac{\mathbf{X}_t - \eta_t R\mathbf{V}_t}{\|\mathbf{X}_t - \eta_t R\mathbf{V}_t\|_F}, \qquad \|\mathbf{V}_t\|_F = 1. \tag{4.1}$$

Vectorizing and dividing by R gives $\hat{\boldsymbol{x}}_{t+1} = (\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t)/\|\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t\|$, where $\boldsymbol{v}_t$ is the vectorized direction. If $\langle \hat{\boldsymbol{x}}_t, \boldsymbol{v}_t\rangle = 0$, its coefficient in (2.6) and its turning angle are

$$s_t^{\mathrm{HB}} = \eta_t, \qquad \Delta\phi_t = \arctan s_t^{\mathrm{HB}}. \tag{4.2}$$

Under tangency, scheduling η_t therefore schedules rotation directly, without any intervening norm dynamics. If the normalized direction is also conditionally unbiased for a geodesically convex loss and the trajectory stays in the cap, Theorem 2.3 applies with $G = 1$; App. E.1 gives the full derivation.

4.2 Prescribed Schedules and Realized Rotation

We train 124M-parameter Llama models on FineWeb for 20,000 updates using AdamH and MuonH, which combine Hyperball with Adam (Kingma & Ba, 2014) and Muon (Jordan et al., 2024) directions. We compare linear decay, the asymptotic form of the bound optimum, with cosine, a WSD profile, and constant. For each schedule, we tune the peak learning rate separately. Training details are in App. E.2. The measured MuonH and AdamH angles follow $\arctan s_t^{\mathrm{HB}}$ across all four schedules (Figs. 10 and 12 in the appendix). This angle agreement shows that the applied learning rate closely controls the observed turn, as expected for approximately tangent updates.

4.3 Schedule Performance under Rotational Control

The angle agreement establishes Hyperball as a setting in which the learning-rate schedule directly prescribes a rotational coefficient schedule. Under the spherical bound of Theorem 2.3, the allocation cost C_T is lower for linear and cosine than for WSD and constant profiles (Fig. 1a); we now test whether the same ordering holds in pretraining loss.

Hyperball pretraining shows the ordering favored by the spherical bound. With the measured angle correspondence, WSD's stable phase leaves rotation nearly flat, while linear and cosine anneal it throughout the post-warmup horizon; C_T favors the latter allocations. Fig. 1 shows that cosine falls faster through much of the run, but linear catches up near the endpoint, while WSD and constant fall behind. Consistent with the bound prediction, separately tuned linear and cosine schedules finish comparably, both ahead of WSD and constant.

The appendix checks sensitivity to schedule amplitude (Fig. 11). Across the tested peak-rate grid, linear and cosine have similar optimal losses, while WSD and constant finish higher.

Without Hyperball, WSD becomes competitive with linear and cosine (Figs. 13 and 14). Thus its disadvantage under rotational control does not carry over to the unconstrained runs. This is

consistent with Hägele et al. (2026, Sec. 4.1.5, Fig. 9), who find linear decay better than the tested WSD variants on the sphere and explain that norm growth supplies implicit annealing during the stable phase of unconstrained training.

To compare the induced profiles, we evaluate the allocation functional on measured rotations from the selected Adam and Muon runs, optimizing each shape's amplitude separately (Fig. 3). The resulting bound favors cosine over WSD for the nominal on-the-sphere profiles, but favors WSD over cosine when evaluated on measured unconstrained rotational trajectories, consistent with the observed validation losses. App. E.4 gives the procedure and endpoint values.

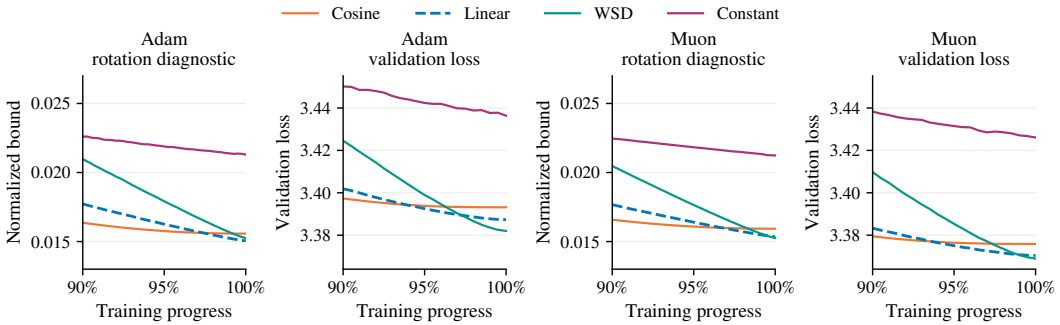


Figure 3: Measured rotational profiles and validation loss for unconstrained Adam and Muon, zoomed to the final 10% of training. Each diagnostic uses the measured post-warmup shape over steps 2,001–20,000, with its amplitude optimized for that full interval. WSD finishes below cosine in both diagnostics and both loss panels.

Takeaway

With Hyperball pretraining, measured turning angles track $\arctan \eta_t$, confirming that the learning-rate schedule approximately controls measured rotation under tangency. Separately tuned linear and cosine schedules finish comparably, both ahead of WSD and constant, consistent with the spherical bound's ranking. Without Hyperball, evaluation on measured rotational profiles recovers the observed WSD–cosine ordering.

5 Conclusion

We showed that for normalized tangent updates on a geodesically convex spherical cap, a new progress estimate leads to the same finite-horizon allocation functional as the Euclidean last-iterate analysis of Defazio et al. (2023); Schaipp et al. (2025). We solve the problem of finding the bound-optimal schedule analytically. Its unique finite-horizon optimum follows an exact recursion, while linear decay attains the same leading bound as the horizon grows. Hyperball's normalized fixed-radius update makes the learning rate act directly as the rotational coefficient, providing an empirical test of the bound's ranking. In 124M Llama/FineWeb pretraining, separately tuned linear and cosine schedules finish comparably, both ahead of WSD and constant profiles, consistent with the bound. Without Hyperball, WSD becomes competitive. Evaluating the same bound on measured rotational trajectories recovers the observed WSD–cosine ordering for Adam and Muon, supporting the use of rotational profiles to compare schedules across these settings.

AI Use Statement

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist in writing proofs, implement methods, and interpret results. We did not use generative AI tools for translation. Generating synthetic datasets, cleaning or reformatting datasets, and qualitative or thematic data analysis were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, draft parts of the paper, summarize or analyze existing literature, discover research topics or identify gaps, brainstorm, search for information, edit the paper for readability, identify relevant literature, format references, suggest the paper's structure, and propose a title or keywords.

We reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, verified mathematical proofs by hand, tested LLM-generated code manually, and revised, proofread, improved readability of LLM-generated text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs. In *International Conference on Learning Representations*, 2025. arXiv:2502.15938.

Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL `https://www.nicolasboumal.net/book`.

Aaron Defazio. Why gradients rapidly increase near the end of training. *arXiv preprint arXiv:2506.02285*, 2025.

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023.

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2024.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alexander Hägele, Alejandro Hernández-Cano, Atli Kosson, and Martin Jaggi. Improving neural network training by decoupling the magnitude and direction of weight vectors. *arXiv preprint arXiv:2606.25971*, 2026.

Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *International Conference on Learning Representations*, 2021. arXiv:2006.08217.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: Efficient and accurate normalization schemes in deep networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. arXiv:1803.01814.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Florian Hübler, Kai Lion, Antonio Orvieto, and Niao He. Muown implicitly performs angular step-size decay. *arXiv preprint arXiv:2606.23637*, 2026.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14058–14070. Curran Associates, Inc., 2022. doi: 10.52202/068431-1022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5aea56eefab60e06f35016478e21aae6-Paper-Conference.pdf.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25333–25369. PMLR, 2024.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020. arXiv:1910.07454.

Kai Lion, Florian Hübler, Bingcong Li, Antonio Orvieto, and Niao He. Muown: Row-norm control for Muon optimization. *arXiv preprint arXiv:2605.10797*, 2026.

Zihan Liu, Ruiheng Zheng, Shaobo Zhang, Changxin Tian, Kunlong Chen, Zhiqiang Zhang, and Lei Wu. Effective learning rate governs loss dynamics in language model pretraining. *arXiv preprint arXiv:2608.24814*, 2026.

LLM-Core Xiaomi. MiMo-V2.6: Scaling reinforcement learning towards self-improvement. Technical report, Xiaomi, 2026. URL https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL/blob/main/MiMo_V2_6_technical_report.pdf.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. arXiv:1608.03983.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.01131.

Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Pérez, and Mathieu Aubry. Spherical perspective on learning with normalization layers. *Neurocomputing*, 487:66–74, 2022. doi: 10.1016/j.neucom.2022.02.021. arXiv:2006.13382.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Şimşekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53267–53294. PMLR, 2025.

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using SGD and weight decay. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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.

Yihao Xiao, Jialong Sun, Zitian Gao, Zeming Wei, Chutian Wang, Ran Tao, Jiaye Teng, and Bryan Dai. Hyperball may not be a free lunch. *arXiv preprint arXiv:2607.22444*, 2026.

Jinghui Yuan, Hongtao Zhang, Jade Zou, Tianyu Li, Wenjie Zhou, Tianyu He, and Wei Chen. HyperTransfer: Understanding the equivalence between base optimizer and hyperball. *arXiv preprint arXiv:2609.07017*, 2026.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. *arXiv preprint arXiv:1602.06053v1*, 2016. URL `https://arxiv.org/abs/1602.06053v1`.

APPENDIX CONTENTS

A Related Work **13**

B Geometry and Proof of the Rotational Coefficient Scheduling Bound **14**

B.1 Scale-Invariant Gradients and the Exact Rotational Update 15

B.2 Sphere Geometry and Geodesic Convexity 17

B.3 One-Step Progress for the Normalized Update 19

B.4 From Tail Averages to the Compact Last-Iterate Bound 24

B.5 Separating Total Mass from Allocation 28

C Proofs for Schedule Optimization **29**

C.1 Exact Finite-Horizon Optimization 29

C.2 From Optimal Cost to the Linear Profile 32

C.3 Performance of Linear Decay 36

C.4 Schedule and Bound Trajectories 37

D A Simple Admissible Scale-Invariant Example **40**

E Hyperball Geometry and Experimental Details **43**

E.1 Geometric Correspondence 43

E.2 Training Protocol 43

E.3 Additional Pretraining Comparisons 43

E.4 Evaluation on Measured Rotational Profiles 46

A RELATED WORK

Finite-horizon learning-rate scheduling. Defazio et al. (2023) develop a last-iterate theory of learning-rate scheduling for stochastic convex optimization: a regret-based formulation yields linear decay with the optimal worst-case convergence rate, while a complementary general-schedule bound allows for arbitrary positive learning-rate sequences. Schaipp et al. (2025) build on the latter result and show that its predictions closely reproduce several empirical features of large-model training, including cosine–WSD behavior, cooldown dynamics, and learning-rate transfer. However, their guarantee does not apply to the scale-invariant setting of normalized neural networks. In fact, their Euclidean-convex loss assumption admits only trivial scale-invariant functions (App. D), and their additive SGD update does not cover normalized tangent steps. We therefore formulate a new last-iterate guarantee for the normalized tangent update on a geodesically convex spherical cap, a setting that admits nontrivial scale-invariant functions. On Riemannian manifolds, Bonnabel (2013) proves asymptotic convergence of stochastic gradient methods, including retraction-based updates. Zhang & Sra (2016) give non-asymptotic rates for geodesically convex optimization, and their stochastic bounds hold for averaged iterates. Our spherical bound, despite the different geometry and assumptions, recovers the same schedule dependence as the Euclidean one. Schaipp et al. (2025) solve the schedule-optimization problem only numerically, finding a linear-decay profile, but they leave global optimality open. We instead solve the functional analytically, proving uniqueness, an exact recursion, and convergence to linear decay.

Scale invariance and rotational dynamics. For scale-invariant groups, parameter norms regulate rotational speed even though they do not directly affect the loss (van Laarhoven, 2017; Arora et al., 2018; Li & Arora, 2020; Heo et al., 2021). Wan et al. (2021) formalize these coupled dynamics for SGD and momentum-SGD, identifying angular updates as the intrinsic measure of optimization

progress; Roburin et al. (2022) characterize the spherical geometry of generic first-order optimizers. More recently, Kosson et al. (2024) analyze stationary rotational equilibria, showing that direct rotational control can reproduce key effects of weight decay. These results describe how effective rotation arises. We study instead how a finite budget of rotational coefficients should be distributed *over time* under the stated spherical bound.

On-sphere optimization and angular control. Constraining optimization to the sphere has served both as an analytical tool to isolate directional dynamics under fixed effective learning rates (Kodryan et al., 2022) and, more recently, as a way to place rotational motion under explicit control. In particular, Hyperball (Wen et al., 2026) and Magnitude–Direction Decoupling (Hägele et al., 2026) fix the sphere radius and expose directional step control; related directional controls appear in Muon-style optimizers (Lion et al., 2026; Hübler et al., 2026). Empirical pretraining studies report effects of matching or transferring effective rotational trajectories on the loss (Liu et al., 2026; Xiao et al., 2026; Yuan et al., 2026). Hägele et al. (2026, Sec. 4.1.5) find linear decay better than the tested WSD variants under magnitude–direction decoupling and attribute the difference to the absence of norm-induced annealing on the sphere. Our contribution here is a finite-horizon bound for a spherical model, providing a theoretically grounded optimal schedule, which we then simulate and compare with empirical results of language model pretraining, showing that the schedule rankings of the two agree. Our bound predicts both the Hyperball and unconstrained Adam/Muon inverse WSD–cosine ordering observed in practice, corroborating the experimental observations of Hägele et al. (2026).

B Geometry and Proof of the Rotational Coefficient Scheduling Bound

We give the geometric derivation and the complete proof of Theorem 2.3. The notation and assumptions are those of Sec. 2.

For reference throughout the proofs, we restate the assumptions of Theorem 2.3 in full.

Assumption B.1 (Spherical stochastic optimization model).

(i) **Loss and initialization.** *The population loss L is differentiable and geodesically convex on $\mathcal{C}$, with minimizer $\hat{\boldsymbol{x}}^\star$. The initial point $\hat{\boldsymbol{x}}_1 \in \mathcal{C}$ is deterministic.*

(ii) **Coefficients.** *The sequence $s_1, \dots, s_T$ is positive, deterministic, and fixed independently of the oracle draws.*

(iii) **Fresh tangent oracle.** *Let $\mathcal{F}_t$ be the history before draw t, including the current point and earlier samples. Almost surely,*

$$\langle \tilde{\boldsymbol{g}}_t, \hat{\boldsymbol{x}}_t \rangle = 0, \qquad \mathbb{E}[\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_t] = \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t).$$

(iv) **Second moment.** *For a deterministic $G > 0$ and every $1 \le t \le T$,*

$$\mathbb{E}\,\|\tilde{\boldsymbol{g}}_t\|^2 \le G^2.$$

(v) **Trajectory.** *The iterates of* (2.6)*, including $\hat{\boldsymbol{x}}_{T+1}$, remain in $\mathcal{C}$ almost surely.*

Proof outline.

1. **From parameters to directions (App. B.1).** Scale invariance makes the raw gradient tangent and gives its inverse-radius scaling. Normalizing the gradient-descent update identifies the induced rotational coefficient s_t^{GD} and proves Proposition 2.2.
2. **Geometry on the cap (App. B.2).** We construct the exponential and logarithm maps, verify that shortest arcs remain in the cap, and derive the first-order inequality for a geodesically convex loss.
3. **One-step progress (App. B.3).** We prove the exponential representation of the normalized update and a deterministic distance inequality. Conditional unbiasedness and geodesic convexity then give expected one-step progress.

4. **The last iterate (App. B.4).** We sum that bound over tails and compare successive tail averages. Their telescoping identity isolates the final loss; collecting the remaining terms gives the normalized allocation cost $C_T(\omega_{1:T})$ and completes the proof of Theorem 2.3.
5. **Mass–allocation separation (App. B.5).** The bound decomposes into total mass and normalized allocation. Optimizing the mass at fixed allocation proves Proposition 2.4 and reduces the schedule-design problem to minimizing the allocation cost alone.

B.1 Scale-Invariant Gradients and the Exact Rotational Update

Let the parameter space be $\mathbb{R}^C$, $C \geq 2$, with unit sphere $\mathbb{S}$. We use the differentiable, positively scale-invariant loss of Definition 2.1. Choosing $\kappa = 1/\|\boldsymbol{x}\|$ gives $L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}})$, which identifies the loss with its restriction to unit directions.

Tangent directions and the raw gradient. At $\hat{\boldsymbol{x}} \in \mathbb{S}$, the tangent space is

$$T_{\hat{\boldsymbol{x}}}\mathbb{S} := \left\{\boldsymbol{v} \in \mathbb{R}^C : \langle \boldsymbol{v}, \hat{\boldsymbol{x}} \rangle = 0\right\}.$$

Indeed, any differentiable sphere curve γ through $\gamma(0) = \hat{\boldsymbol{x}}$ satisfies $\gamma(a) \in \mathbb{S}$, and hence $\|\gamma(a)\|^2 = 1$, for every value of its curve parameter. Differentiating this constant identity gives

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}a} \langle \gamma(a), \gamma(a) \rangle \\ &= \langle \gamma'(a), \gamma(a) \rangle + \langle \gamma(a), \gamma'(a) \rangle = 2 \langle \gamma(a), \gamma'(a) \rangle . \end{aligned}$$

Here γ' denotes differentiation with respect to the curve parameter. Evaluating at $a = 0$ yields $\langle \hat{\boldsymbol{x}}, \gamma'(0) \rangle = 0$. Thus, a sphere curve has tangent velocity. Conversely, every tangent vector is the velocity of a sphere curve, as the explicit construction below verifies.

Stretching the weight along its ray gives

$$L(\boldsymbol{x} + h\boldsymbol{x}) = L((1+h)\boldsymbol{x}) = L(\boldsymbol{x}) \qquad (h > -1).$$

The left-hand side is therefore constant as a function of h. Since $\frac{\mathrm{d}}{\mathrm{d}h}(\boldsymbol{x} + h\boldsymbol{x}) = \boldsymbol{x}$, the ordinary chain rule, following Kosson et al. (2024, App. B), gives

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}h} L(\boldsymbol{x} + h\boldsymbol{x}) \Big|_{h=0} \\ &= \langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x} + h\boldsymbol{x}), \boldsymbol{x} \rangle|_{h=0} = \langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \boldsymbol{x} \rangle . \end{aligned}$$

This zero inner product says that the raw gradient is orthogonal to the radial direction $\boldsymbol{x}$. Since $\|\boldsymbol{x}\| > 0$, division by the radius gives

$$\langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \hat{\boldsymbol{x}} \rangle = \frac{\langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \boldsymbol{x} \rangle}{\|\boldsymbol{x}\|} = 0.$$

Hence the raw gradient belongs to $T_{\hat{\boldsymbol{x}}}\mathbb{S}$.

Definition B.2 (Tangent space and spherical gradient)**.** *At a unit direction $\hat{\boldsymbol{x}}$, the tangent space is $T_{\hat{\boldsymbol{x}}}\mathbb{S} := \{\boldsymbol{v} : \langle \boldsymbol{v}, \hat{\boldsymbol{x}} \rangle = 0\}$. Its vectors are the possible instantaneous velocities of curves on the sphere through $\hat{\boldsymbol{x}}$. The spherical gradient $\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}})$ is the unique tangent vector whose inner product with each such velocity gives the corresponding directional derivative of the loss along the sphere.*

The intrinsic spherical gradient. Let f be differentiable on an open subset of the sphere. Its spherical gradient at $\hat{\boldsymbol{x}}$ is the tangent vector characterized by

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}a} f(\gamma(a)) \Big|_{a=0} = \langle \nabla_{\mathbb{S}} f(\hat{\boldsymbol{x}}), \boldsymbol{v} \rangle , \\ &\gamma(0) = \hat{\boldsymbol{x}}, \qquad \gamma'(0) = \boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}. \end{aligned} \tag{B.1}$$

For any $\boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$, expanding a tangent proposal gives

$$\|\hat{\boldsymbol{x}} + a\boldsymbol{v}\|^2 = \|\hat{\boldsymbol{x}}\|^2 + 2a \langle \hat{\boldsymbol{x}}, \boldsymbol{v} \rangle + a^2 \|\boldsymbol{v}\|^2 = 1 + a^2 \|\boldsymbol{v}\|^2 > 0.$$

Consequently the curve

$$\gamma(a) = \frac{\hat{\boldsymbol{x}} + a\boldsymbol{v}}{\sqrt{1 + a^2 \|\boldsymbol{v}\|^2}}$$

lies on the sphere and stays in the open domain of f for sufficiently small $|a|$. Differentiating the numerator and denominator gives

$$\gamma'(a) = \frac{\boldsymbol{v}}{\sqrt{1 + a^2 \|\boldsymbol{v}\|^2}} - \frac{a \|\boldsymbol{v}\|^2 (\hat{\boldsymbol{x}} + a\boldsymbol{v})}{(1 + a^2 \|\boldsymbol{v}\|^2)^{3/2}}, \qquad \gamma'(0) = \boldsymbol{v}.$$

This supplies a curve for every tangent velocity, including $\boldsymbol{v} = 0$.

Lemma B.3 (Scale-invariant gradient and spherical gradient)**.** *For a differentiable scale-invariant loss on the open domain closed under positive rescaling specified above, every $\boldsymbol{x}$ in that domain satisfies*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}) = \frac{1}{\|\boldsymbol{x}\|} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}), \qquad \hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}. \tag{B.2}$$

Proof of Lemma B.3. Fix $\boldsymbol{x}$ and any $\boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$. Choose a sphere curve in the loss domain with $\gamma(0) = \hat{\boldsymbol{x}}$ and $\gamma'(0) = \boldsymbol{v}$, as constructed above. The radius $\|\boldsymbol{x}\|$ is constant as the curve parameter varies, and $\|\boldsymbol{x}\| \gamma(0) = \boldsymbol{x}$. Scale invariance gives

$$L(\|\boldsymbol{x}\| \gamma(a)) = L(\gamma(a)).$$

The factor $\|\boldsymbol{x}\|$ is constant along the curve, so

$$\frac{\mathrm{d}}{\mathrm{d}a}\big(\|\boldsymbol{x}\| \gamma(a)\big)\Big|_{a=0} = \|\boldsymbol{x}\| \gamma'(0) = \|\boldsymbol{x}\| \boldsymbol{v}.$$

The intrinsic gradient definition and the ordinary chain rule now yield

$$\begin{aligned}
\langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}), \boldsymbol{v} \rangle &= \frac{\mathrm{d}}{\mathrm{d}a} L(\gamma(a))\Big|_{a=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}a} L(\|\boldsymbol{x}\| \gamma(a))\Big|_{a=0} \\
&= \langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \|\boldsymbol{x}\| \boldsymbol{v} \rangle \\
&= \|\boldsymbol{x}\| \langle \nabla_{\boldsymbol{x}} L(\boldsymbol{x}), \boldsymbol{v} \rangle .
\end{aligned}$$

Since this holds for every $\boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$, the tangent vector $\|\boldsymbol{x}\| \nabla_{\boldsymbol{x}} L(\boldsymbol{x})$ satisfies the characterization (B.1) of the spherical gradient, so $\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}) = \|\boldsymbol{x}\| \nabla_{\boldsymbol{x}} L(\boldsymbol{x})$. Dividing by $\|\boldsymbol{x}\| > 0$ gives (B.2). □

Positive rescaling preserves the unit direction, since $\|\kappa\boldsymbol{x}\| = \kappa \|\boldsymbol{x}\|$ and $\kappa\boldsymbol{x} / \|\kappa\boldsymbol{x}\| = \hat{\boldsymbol{x}}$. Applying the bridge at $\kappa\boldsymbol{x}$ gives

$$\nabla_{\boldsymbol{x}} L(\kappa\boldsymbol{x}) = \frac{\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}})}{\kappa \|\boldsymbol{x}\|} = \kappa^{-1} \nabla_{\boldsymbol{x}} L(\boldsymbol{x}).$$

This is the inverse-radius raw-gradient scaling used in the scale-invariant analysis of van Laarhoven (2017); Li & Arora (2020).

Completion of the proof of Proposition 2.2. The exact rotational gradient step. Choose Euclidean gradient descent with a finite step size $\eta_t > 0$. Substituting (B.2) and factoring the radius gives

$$\begin{aligned}
\boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - \eta_t \nabla_{\boldsymbol{x}} L(\boldsymbol{x}_t) \\
&= \|\boldsymbol{x}_t\| \hat{\boldsymbol{x}}_t - \frac{\eta_t}{\|\boldsymbol{x}_t\|} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t) \\
&= \|\boldsymbol{x}_t\| \left(\hat{\boldsymbol{x}}_t - \frac{\eta_t}{\|\boldsymbol{x}_t\|^2} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t) \right).
\end{aligned} \tag{B.3}$$

Define the rotational coefficient

$$s_t^{\mathrm{GD}} := \frac{\eta_t}{\|\boldsymbol{x}_t\|^2}. \tag{B.4}$$

Tangency removes the cross term in the remaining proposal:

$$\begin{aligned}\left\|\hat{\boldsymbol{x}}_t - s_t^{\mathrm{GD}}\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2 &= \|\hat{\boldsymbol{x}}_t\|^2 - 2s_t^{\mathrm{GD}}\left\langle\hat{\boldsymbol{x}}_t, \nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\rangle + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2 \\ &= 1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2 > 0.\end{aligned}$$

Taking the norm of the factored weight step therefore gives

$$\|\boldsymbol{x}_{t+1}\| = \|\boldsymbol{x}_t\|\sqrt{1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2} > 0.$$

Define $\hat{\boldsymbol{x}}_{t+1} = \boldsymbol{x}_{t+1}/\|\boldsymbol{x}_{t+1}\|$. Substitution displays the positive factor that cancels:

$$\hat{\boldsymbol{x}}_{t+1} = \frac{\|\boldsymbol{x}_t\|\left(\hat{\boldsymbol{x}}_t - s_t^{\mathrm{GD}}\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right)}{\|\boldsymbol{x}_t\|\sqrt{1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2}}.$$

Thus the exact rotational recursion is

$$\hat{\boldsymbol{x}}_{t+1} = \frac{\hat{\boldsymbol{x}}_t - s_t^{\mathrm{GD}}\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)}{\sqrt{1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2}}, \qquad s_t^{\mathrm{GD}} = \frac{\eta_t}{\|\boldsymbol{x}_t\|^2}. \tag{B.5}$$

The coefficient (2.2) is the geometric effective learning rate of van Laarhoven (2017, Sec. 3.1). One inverse-radius factor comes from the raw-gradient bridge; the other comes from factoring the weight radius before normalization.

Now let us recall the turning angle, defined $\Delta\phi_t := \angle(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = \angle(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_{t+1})$; positive normalization preserves the angle between the two vectors. Tangency makes $\hat{\boldsymbol{x}}_t$ and $-\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)/\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\|$ an orthonormal pair when the gradient is nonzero. The normalized endpoint has coordinates

$$\cos\Delta\phi_t = \frac{1}{\sqrt{1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2}}, \qquad \sin\Delta\phi_t = \frac{s_t^{\mathrm{GD}}\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|}{\sqrt{1 + (s_t^{\mathrm{GD}})^2\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|^2}}.$$

The radial coordinate is positive, so $\Delta\phi_t = \arctan(s_t^{\mathrm{GD}}\left\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\right\|) \in [0, \pi/2)$. Combining the two orthonormal components gives, for nonzero gradient,

$$\hat{\boldsymbol{x}}_{t+1} = \cos(\Delta\phi_t)\hat{\boldsymbol{x}}_t - \sin(\Delta\phi_t)\frac{\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)}{\|\nabla_{\mathbb{S}}L(\hat{\boldsymbol{x}}_t)\|}.$$

Substituting the displayed sine and cosine coordinates recovers (B.5), including the negative gradient sign. This proves (2.3) and (2.4). □

With the rotational update established, we now develop the geometric tools needed to measure progress: distance on the sphere, the exponential and logarithm maps, and the supporting inequality for geodesically convex losses.

B.2 Sphere Geometry and Geodesic Convexity

For unit directions, the spherical distance is

$$d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) := \arccos\langle\hat{\boldsymbol{x}}, \boldsymbol{y}\rangle. \tag{B.6}$$

The exponential map follows the great circle selected by a tangent vector. For $\boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$ with $\boldsymbol{v} \neq 0$, set

$$\boldsymbol{e} := \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}. \tag{B.7}$$

Tangency gives $\langle \boldsymbol{e}, \hat{\boldsymbol{x}} \rangle = 0$, while both vectors have unit norm. They therefore form an orthonormal basis of the two-dimensional plane containing the motion. Moving through angle $\|\boldsymbol{v}\|$ in that plane gives

$$\begin{aligned} \mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{v}) &= \cos(\|\boldsymbol{v}\|)\hat{\boldsymbol{x}} + \sin(\|\boldsymbol{v}\|)\boldsymbol{e} \\ &= \cos(\|\boldsymbol{v}\|)\hat{\boldsymbol{x}} + \sin(\|\boldsymbol{v}\|)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}, \\ \mathrm{Exp}_{\hat{\boldsymbol{x}}}(0) &= \hat{\boldsymbol{x}}. \end{aligned} \tag{B.8}$$

Varying the angle parametrizes a unit circle with unit speed, so the traveled arc has length $\|\boldsymbol{v}\|$.

The logarithm map reverses this construction. Let $\boldsymbol{y} \neq \hat{\boldsymbol{x}}$ be non-antipodal to $\hat{\boldsymbol{x}}$, and write $\theta = d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) \in (0, \pi)$. Since $\langle \hat{\boldsymbol{x}}, \boldsymbol{y} \rangle = \cos\theta$, the vector $\cos\theta\, \hat{\boldsymbol{x}}$ is the radial component of $\boldsymbol{y}$ along $\hat{\boldsymbol{x}}$. Subtracting it leaves the tangent component:

$$\begin{aligned} \langle \boldsymbol{y} - \cos\theta\, \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} \rangle &= \cos\theta - \cos\theta = 0, \\ \|\boldsymbol{y} - \cos\theta\, \hat{\boldsymbol{x}}\|^2 &= 1 - 2\cos^2\theta + \cos^2\theta = \sin^2\theta. \end{aligned} \tag{B.9}$$

Therefore the vector

$$\boldsymbol{e} := \frac{\boldsymbol{y} - \cos\theta\, \hat{\boldsymbol{x}}}{\sin\theta} \tag{B.10}$$

is the unit tangent direction from $\hat{\boldsymbol{x}}$ toward $\boldsymbol{y}$, and $\boldsymbol{y} = \cos\theta\, \hat{\boldsymbol{x}} + \sin\theta\, \boldsymbol{e}$. The logarithm vector points in this direction and has length θ, because Exp travels a unit-sphere arc whose length equals its angle. Hence

$$\mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) = \theta \boldsymbol{e} = \frac{\theta}{\sin\theta}(\boldsymbol{y} - \cos\theta\, \hat{\boldsymbol{x}}), \qquad \mathrm{Log}_{\hat{\boldsymbol{x}}}(\hat{\boldsymbol{x}}) = 0. \tag{B.11}$$

Substitution into (B.8) returns $\boldsymbol{y}$. These are the standard sphere maps (Boumal, 2023, Secs. 7.2 and 10.2).

For $0 \leq a \leq 1$, the point $\gamma(a) = \mathrm{Exp}_{\hat{\boldsymbol{x}}}(a\, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y}))$ travels the fraction a of the shortest arc from $\hat{\boldsymbol{x}}$ to $\boldsymbol{y}$. The tangent vector records both direction and arc length: scaling it scales the distance traveled along that same geodesic. The tangent space is a vector space through the origin; the drawing attaches its translated copy to $\hat{\boldsymbol{x}}$ so that the vector and its spherical endpoint can be compared.

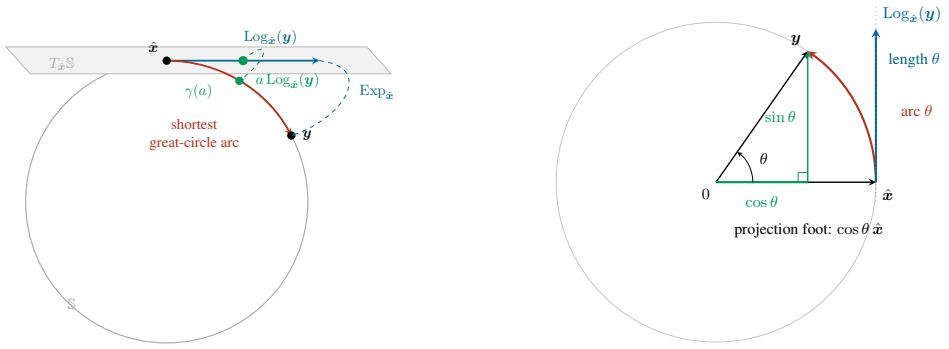


Figure 4: Sphere geometry used in the proof. **(a)** The tangent plane is drawn as $\hat{\boldsymbol{x}} + T_{\hat{\boldsymbol{x}}}\mathbb{S}$. The logarithm map assigns to $\boldsymbol{y}$ a tangent vector whose length is the shortest arc distance; the exponential map returns this vector, or any fraction of it, to the corresponding point on that arc. **(b)** In the plane spanned by $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}$, the radial projection is $\cos\theta\, \hat{\boldsymbol{x}}$, the perpendicular residual has length $\sin\theta$, and the Log vector and arc have length θ.

Recall the open spherical cap in (2.5):

$$\mathcal{C} := \{\hat{\boldsymbol{x}} \in \mathbb{S} : d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star) < \theta_{\mathrm{cap}}\}, \qquad 0 < \theta_{\mathrm{cap}} < \frac{\pi}{4}, \tag{B.12}$$

with fixed center $\hat{\boldsymbol{x}}^\star$. For any $\hat{\boldsymbol{x}}, \boldsymbol{y} \in \mathcal{C}$, the triangle inequality gives

$$\begin{aligned} d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) &\leq d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star) + d_{\mathbb{S}}(\hat{\boldsymbol{x}}^\star, \boldsymbol{y}) \\ &< 2\theta_{\mathrm{cap}} < \frac{\pi}{2}. \end{aligned} \tag{B.13}$$

Thus every pair of cap points has acute separation, used in the one-step argument below. The cap $\mathcal{C}$ is geodesically convex (Boumal, 2023, Exercise 11.19): the shortest arc between any two cap points stays in the cap.

Corollary B.4 (Shortest-arc containment of the cap)**.** *For any two points in $\mathcal{C}$, the unique shortest spherical arc joining them lies entirely in $\mathcal{C}$.*

Geodesic convexity and the supporting inequality. In Euclidean optimization, convexity compares the loss along a straight line segment with the linear interpolation of the losses at its endpoints. On our spherical cap, we make the same comparison while moving along the shortest spherical arc between the endpoints. At a fraction a of the distance along the arc, the loss must be at most $(1-a)L(\hat{\boldsymbol{x}}) + aL(\boldsymbol{y})$. This is *geodesic convexity*. The arc stays in the cap by Corollary B.4.

Definition B.5 (Geodesic convexity on the cap)**.** *For $\hat{\boldsymbol{x}}, \boldsymbol{y} \in \mathcal{C}$, let $\gamma : [0,1] \to \mathcal{C}$ be the constant-speed shortest spherical arc with $\gamma(0) = \hat{\boldsymbol{x}}$ and $\gamma(1) = \boldsymbol{y}$. Here $a \in [0,1]$ denotes the fraction of the arc traversed. The loss L is geodesically convex on $\mathcal{C}$ if, for every such pair,*

$$L(\gamma(a)) \leq (1-a)L(\hat{\boldsymbol{x}}) + aL(\boldsymbol{y}), \qquad 0 \leq a \leq 1. \tag{B.14}$$

The arc in Definition B.5 has the representation

$$\gamma(a) := \operatorname{Exp}_{\hat{\boldsymbol{x}}}(a \operatorname{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})), \qquad 0 \leq a \leq 1. \tag{B.15}$$

For differentiable L, Boumal's first-order characterization gives the following supporting inequality.

Lemma B.6 (Supporting inequality on the cap)**.** *If L is differentiable and geodesically convex on $\mathcal{C}$, then*

$$L(\boldsymbol{y}) \geq L(\hat{\boldsymbol{x}}) + \langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}), \operatorname{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \rangle \qquad (\hat{\boldsymbol{x}}, \boldsymbol{y} \in \mathcal{C}). \tag{B.16}$$

Proof of Lemma B.6. Use the curve γ from (B.15), which stays in $\mathcal{C}$ and satisfies $\gamma(0) = \hat{\boldsymbol{x}}$ and $\gamma(1) = \boldsymbol{y}$. By Boumal (2023, Theorem 11.21, Eq. (11.6)), a function that is differentiable and geodesically convex on the open cap lies above its first-order expansion along every geodesic in the cap. On the sphere, the Riemannian gradient is $\nabla_{\mathbb{S}} L$ and inner products are the usual Euclidean ones, so that the theorem reads

$$L(\gamma(a)) \geq L(\hat{\boldsymbol{x}}) + a \langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}), \gamma'(0) \rangle, \qquad 0 \leq a \leq 1.$$

Here the prime denotes differentiation with respect to a. Differentiating the spherical exponential formula (B.8) along this curve at $a = 0$ identifies its initial velocity:

$$\gamma'(0) = \operatorname{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y}). \tag{B.17}$$

Substituting this velocity and setting $a = 1$ gives (B.16). For coincident endpoints, the curve is constant, the Log vector is zero, and the inequality is an equality. □

The Log vector therefore replaces the Euclidean displacement to the comparator: its direction is the direction of travel, and its length is the distance along the shortest arc.

B.3 One-Step Progress for the Normalized Update

We now combine the geometric tools of the preceding subsection with the stochastic oracle of Assumption B.1 to derive the expected one-step progress inequality for the normalized update (2.6). The standard Euclidean SGD calculation starts from

$$\boldsymbol{x}^+ = \boldsymbol{x} - s\tilde{\boldsymbol{g}} \tag{B.18}$$

and expands squared distance to a comparator $\boldsymbol{y}$:

$$\left\|\boldsymbol{x}^+ - \boldsymbol{y}\right\|^2 = \|\boldsymbol{x} - \boldsymbol{y}\|^2 - 2s \langle \tilde{\boldsymbol{g}}, \boldsymbol{x} - \boldsymbol{y} \rangle + s^2 \left\|\tilde{\boldsymbol{g}}\right\|^2. \tag{B.19}$$

On the sphere, the analogous distance comparison lifts both endpoints into a common tangent space via the logarithm and exponential maps. For an exponential-map step, the resulting inequality is the nonnegative-curvature case of Zhang & Sra (2016, Lemma 6 and Corollary 8), in which their curvature factor equals one; Lemma B.7 gives a self-contained derivation on the sphere.

Lemma B.7 (Non-expansion of the spherical exponential map)**.** *At $\hat{\boldsymbol{x}} \in \mathbb{S}$, let $\boldsymbol{v}, \boldsymbol{w} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$ satisfy $\|\boldsymbol{v}\|, \|\boldsymbol{w}\| < \pi$. Then*

$$d_{\mathbb{S}}(\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{v}), \mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{w})) \le \|\boldsymbol{v} - \boldsymbol{w}\|. \tag{B.20}$$

Mapping two tangent vectors to the sphere by the exponential map does not increase their separation: the spherical distance between the endpoints is at most the Euclidean distance between the tangent vectors.

Proof of Lemma B.7. We compare the length of the straight tangent segment with that of its exponential image. Set

$$\boldsymbol{c}(a) = (1-a)\boldsymbol{v} + a\boldsymbol{w}, \qquad r(a) = \|\boldsymbol{c}(a)\|, \qquad \boldsymbol{e}(a) = \frac{\boldsymbol{c}(a)}{r(a)} \quad \text{when } r(a) > 0.$$

For $0 \le a \le 1$, the triangle inequality gives $r(a) \le (1-a)\|\boldsymbol{v}\| + a\|\boldsymbol{w}\| < \pi$. Primes denote derivatives with respect to a. Differentiating $\|\boldsymbol{e}\|^2 = 1$ gives $2\langle \boldsymbol{e}', \boldsymbol{e}\rangle = 0$, so $\boldsymbol{e}' \perp \boldsymbol{e}$ and the product rule on $\boldsymbol{c} = r\boldsymbol{e}$ gives

$$\begin{aligned} \boldsymbol{c}' &= r'\boldsymbol{e} + r\boldsymbol{e}', \\ \|\boldsymbol{c}'\|^2 &= \|r'\boldsymbol{e} + r\boldsymbol{e}'\|^2 \\ &= (r')^2\|\boldsymbol{e}\|^2 + 2r'r\langle \boldsymbol{e}, \boldsymbol{e}'\rangle + r^2\|\boldsymbol{e}'\|^2 \\ &= (r')^2 + r^2\|\boldsymbol{e}'\|^2. \end{aligned} \tag{B.21}$$

Since $\hat{\boldsymbol{x}}$ is fixed, the chain and product rules give

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}a}\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{c}(a)) &= \frac{\mathrm{d}}{\mathrm{d}a}\left[\cos(r(a))\,\hat{\boldsymbol{x}} + \sin(r(a))\,\boldsymbol{e}(a)\right] \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}a}\cos r\right)\hat{\boldsymbol{x}} + \left(\frac{\mathrm{d}}{\mathrm{d}a}\sin r\right)\boldsymbol{e} + \sin r\,\boldsymbol{e}' \\ &= -r'\sin r\,\hat{\boldsymbol{x}} + r'\cos r\,\boldsymbol{e} + \sin r\,\boldsymbol{e}' \\ &= r'(-\sin r\,\hat{\boldsymbol{x}} + \cos r\,\boldsymbol{e}) + \sin r\,\boldsymbol{e}'. \end{aligned}$$

Tangency gives $\langle \hat{\boldsymbol{x}}, \boldsymbol{e}\rangle = 0$; differentiating this identity at fixed $\hat{\boldsymbol{x}}$ gives $\langle \hat{\boldsymbol{x}}, \boldsymbol{e}'\rangle = 0$. Together with $\langle \boldsymbol{e}, \boldsymbol{e}'\rangle = 0$ and $\|\hat{\boldsymbol{x}}\| = \|\boldsymbol{e}\| = 1$, these identities eliminate the cross terms below. Expanding the squared speed and then using $0 \le \sin r \le r$, we obtain

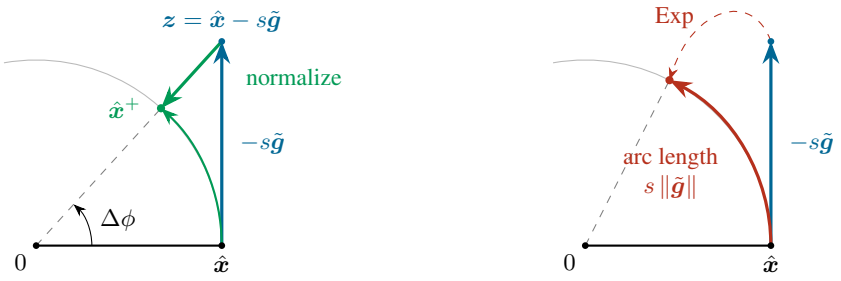
$$\begin{aligned} \left\|\frac{\mathrm{d}}{\mathrm{d}a}\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{c}(a))\right\|^2 &= \|r'(-\sin r\,\hat{\boldsymbol{x}} + \cos r\,\boldsymbol{e}) + \sin r\,\boldsymbol{e}'\|^2 \\ &= (r')^2\|-\sin r\,\hat{\boldsymbol{x}} + \cos r\,\boldsymbol{e}\|^2 + \sin^2 r\;\|\boldsymbol{e}'\|^2 \\ &\quad + 2r'\sin r\big(-\sin r\;\langle \hat{\boldsymbol{x}}, \boldsymbol{e}'\rangle + \cos r\;\langle \boldsymbol{e}, \boldsymbol{e}'\rangle\big) \\ &= (r')^2(\sin^2 r + \cos^2 r) + \sin^2 r\;\|\boldsymbol{e}'\|^2 \\ &= (r')^2 + \sin^2 r\;\|\boldsymbol{e}'\|^2 \\ &\le (r')^2 + r^2\|\boldsymbol{e}'\|^2 = \|\boldsymbol{c}'\|^2. \end{aligned} \tag{B.22}$$

As a runs from zero to one, $\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{c}(a))$ traces a curve on the sphere joining $\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{v})$ to $\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{w})$. Its length is the integral of its speed, where speed is the norm of the derivative with respect to a. Spherical distance is the length of the shortest joining curve, so it is at most this integral. Taking square roots of the preceding squared-speed inequality then bounds the image speed by $\|\boldsymbol{c}'(a)\|$. Finally, the straight segment has constant velocity $\boldsymbol{c}'(a) = -\boldsymbol{v} + \boldsymbol{w}$, giving

$$\begin{aligned} d_{\mathbb{S}}(\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{v}), \mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{w})) &\le \int_0^1 \left\|\frac{\mathrm{d}}{\mathrm{d}a}\mathrm{Exp}_{\hat{\boldsymbol{x}}}(\boldsymbol{c}(a))\right\| \mathrm{d}a \\ &\le \int_0^1 \|\boldsymbol{c}'(a)\|\,\mathrm{d}a \\ &= \int_0^1 \|\boldsymbol{w} - \boldsymbol{v}\|\,\mathrm{d}a \\ &= \|\boldsymbol{w} - \boldsymbol{v}\| \int_0^1 1\,\mathrm{d}a = \|\boldsymbol{w} - \boldsymbol{v}\|. \end{aligned} \tag{B.23}$$

□

The non-expansion lemma applies to exponential-map endpoints, whereas the algorithm normalizes a tangent proposal; Fig. 5 illustrates the difference, and the following lemma gives the exact correspondence.



(a) (b)

Figure 5: Two operations on the same tangent vector, shown at $s\,\|\tilde{\boldsymbol{g}}\| = 1.1$ to make their difference visible. (a) Normalizing the proposal $\hat{\boldsymbol{x}} - s\tilde{\boldsymbol{g}}$ projects back to the sphere, turning by $\arctan(1.1) \approx 0.833$ radians. (b) Applying Exp follows the great circle for 1.1 radians. Both endpoints lie in the same plane.

Lemma B.8 (Exponential representation of the normalized tangent update)**.** *Let $\hat{\boldsymbol{x}} \in \mathbb{S}$, $\tilde{\boldsymbol{g}} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$, and $s > 0$. Define*

$$\hat{\boldsymbol{x}}^+ := \frac{\hat{\boldsymbol{x}} - s\tilde{\boldsymbol{g}}}{\sqrt{1 + s^2\,\|\tilde{\boldsymbol{g}}\|^2}}, \qquad \chi := s\,\|\tilde{\boldsymbol{g}}\|\,,$$

and

$$h := \begin{cases} \arctan\chi / \|\tilde{\boldsymbol{g}}\|\,, & \tilde{\boldsymbol{g}} \neq 0, \\ s, & \tilde{\boldsymbol{g}} = 0. \end{cases} \tag{B.24}$$

Then

$$\hat{\boldsymbol{x}}^+ = \mathrm{Exp}_{\hat{\boldsymbol{x}}}(-h\tilde{\boldsymbol{g}}), \qquad 0 < h \leq s, \tag{B.25}$$

and

$$0 \leq (s-h)\,\|\tilde{\boldsymbol{g}}\| = \chi - \arctan\chi \leq \frac{\chi^2}{4}. \tag{B.26}$$

Proof. For $\tilde{\boldsymbol{g}} \neq 0$, the definition gives $h\,\|\tilde{\boldsymbol{g}}\| = \arctan\chi$. Using $\cos(\arctan\chi) = 1/\sqrt{1+\chi^2}$ and $\sin(\arctan\chi) = \chi/\sqrt{1+\chi^2}$ in the exponential formula yields

$$\begin{aligned} \mathrm{Exp}_{\hat{\boldsymbol{x}}}(-h\tilde{\boldsymbol{g}}) &= \cos(\arctan\chi)\hat{\boldsymbol{x}} - \sin(\arctan\chi)\frac{\tilde{\boldsymbol{g}}}{\|\tilde{\boldsymbol{g}}\|} \\ &= \frac{\hat{\boldsymbol{x}} - (\chi/\|\tilde{\boldsymbol{g}}\|)\tilde{\boldsymbol{g}}}{\sqrt{1+\chi^2}} = \frac{\hat{\boldsymbol{x}} - s\tilde{\boldsymbol{g}}}{\sqrt{1 + s^2\,\|\tilde{\boldsymbol{g}}\|^2}} = \hat{\boldsymbol{x}}^+. \end{aligned}$$

Since $0 < \arctan\chi \leq \chi$ for $\chi > 0$, division by $\|\tilde{\boldsymbol{g}}\| > 0$ gives $0 < h \leq s$. If $\tilde{\boldsymbol{g}} = 0$, then $h = s$, $\chi = 0$, and both endpoints equal $\hat{\boldsymbol{x}}$.

In either case, $(s-h)\,\|\tilde{\boldsymbol{g}}\| = \chi - \arctan\chi$. For every $\chi \geq 0$, integration gives

$$0 \leq \chi - \arctan\chi = \int_0^\chi \frac{a^2}{1+a^2}\,\mathrm{d}a \leq \int_0^\chi \frac{a}{2}\,\mathrm{d}a = \frac{\chi^2}{4}.$$

The integrand inequality follows from $a/2 - a^2/(1+a^2) = a(a-1)^2/[2(1+a^2)] \geq 0$ for $a \geq 0$. □

The next lemma combines Lemma B.7 with Lemma B.8 to give a distance inequality in the prescribed coefficient s.

Lemma B.9 (Distance comparison for the normalized tangent update)**.** *Let $\hat{\boldsymbol{x}}, \boldsymbol{y} \in \mathcal{C}$, $\tilde{\boldsymbol{g}} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$, and $s > 0$, and set $\hat{\boldsymbol{x}}^+ = (\hat{\boldsymbol{x}} - s\tilde{\boldsymbol{g}})/\sqrt{1 + s^2 \|\tilde{\boldsymbol{g}}\|^2}$. Then*

$$\begin{aligned} d_{\mathbb{S}}(\hat{\boldsymbol{x}}^+, \boldsymbol{y})^2 \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})^2 + 2s \langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \rangle \\ + (1 + \theta_{\mathrm{cap}}) s^2 \|\tilde{\boldsymbol{g}}\|^2 . \end{aligned} \tag{B.27}$$

The left side is the squared distance to the comparator after one normalized step. The right side has the same three-term structure as the Euclidean expansion above: the squared distance before the step, a signed inner-product term that will become the descent direction after taking conditional expectations, and a quadratic cost proportional to $s^2 \|\tilde{\boldsymbol{g}}\|^2$.

Proof. Set $\chi = s\|\tilde{\boldsymbol{g}}\|$ and use h from Lemma B.8. Both $-h\tilde{\boldsymbol{g}}$ and $\mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})$ are tangent at $\hat{\boldsymbol{x}}$, with lengths

$$h\|\tilde{\boldsymbol{g}}\| = \arctan\chi < \pi/2, \qquad \|\mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\| = d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) < 2\theta_{\mathrm{cap}} < \pi/2.$$

Their exponential-map endpoints are $\hat{\boldsymbol{x}}^+$ and $\boldsymbol{y}$, respectively. Thus Lemma B.7 applies. Squaring its inequality and expanding gives

$$\begin{aligned} d_{\mathbb{S}}(\hat{\boldsymbol{x}}^+, \boldsymbol{y})^2 &\le \|-h\tilde{\boldsymbol{g}} - \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\|^2 \\ &= \|\mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\|^2 + 2h\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle + h^2\|\tilde{\boldsymbol{g}}\|^2 \\ &= d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})^2 + 2h\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle + h^2\|\tilde{\boldsymbol{g}}\|^2 . \end{aligned}$$

The inner product can have either sign, so $h \le s$ alone does not justify replacing its coefficient by s. Add and subtract the term with coefficient s:

$$\begin{aligned} 2h\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle = \underbrace{2s\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle}_{\text{term with coefficient } s} \\ + \underbrace{2(h-s)\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle}_{\text{error from replacing } h \text{ by } s} . \end{aligned} \tag{B.28}$$

Since $s - h \ge 0$, the magnitude of the error is bounded by Cauchy–Schwarz:

$$\begin{aligned} 2(h-s)\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle &\le 2(s-h)\,|\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle| \\ &\le 2(s-h)\|\tilde{\boldsymbol{g}}\|\,\|\mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\| \\ &= 2d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})(s-h)\|\tilde{\boldsymbol{g}}\| . \end{aligned}$$

The estimate in Lemma B.8 therefore gives

$$\begin{aligned} 2d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})(s-h)\|\tilde{\boldsymbol{g}}\| &= 2d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})(\chi - \arctan\chi) \\ &\le \frac{d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})}{2}\chi^2 . \end{aligned} \tag{B.29}$$

Also, $0 < h \le s$ implies

$$h^2\|\tilde{\boldsymbol{g}}\|^2 \le s^2\|\tilde{\boldsymbol{g}}\|^2 = \chi^2 .$$

After the decomposition (B.28), the expanded distance inequality has two terms besides the desired $2s\langle \tilde{\boldsymbol{g}}, \mathrm{Log}_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\rangle$: the error and $h^2\|\tilde{\boldsymbol{g}}\|^2$. Adding their bounds and using $d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) < 2\theta_{\mathrm{cap}}$ gives

$$\frac{d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})}{2}\chi^2 + \chi^2 = \left(1 + \frac{d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \boldsymbol{y})}{2}\right)\chi^2 \le (1 + \theta_{\mathrm{cap}})s^2\|\tilde{\boldsymbol{g}}\|^2 .$$

Substitution into the squared-distance bound at the start of the proof gives (B.27). □

For the stochastic update, h depends on the sampled gradient norm, so conditional unbiasedness does not directly simplify $\mathbb{E}[h\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_t]$. Lemma B.9 absorbs this difference into the error term, producing a distance inequality whose inner-product term has the deterministic coefficient s_t.

Proposition B.10 (Expected one-step progress)**.** *Under Assumption B.1, let $\boldsymbol{y} \in \mathcal{C}$ be $\mathcal{F}_t$-measurable. Then*

$$s_t\mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] \le \frac{1}{2}\mathbb{E}\big[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2\big] + \frac{1+\theta_{\mathrm{cap}}}{2} s_t^2 \mathbb{E}\left\|\tilde{\boldsymbol{g}}_t\right\|^2. \tag{B.30}$$

Proof. Step 1: deterministic distance bound. Apply Lemma B.9 with $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_t$, $\tilde{\boldsymbol{g}} = \tilde{\boldsymbol{g}}_t$, and $s = s_t$. The normalized endpoint is exactly $\hat{\boldsymbol{x}}_{t+1}$, so, on every sample path,

$$d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 + 2s_t \left\langle \tilde{\boldsymbol{g}}_t, \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle + (1+\theta_{\mathrm{cap}}) s_t^2 \left\|\tilde{\boldsymbol{g}}_t\right\|^2.$$

Step 2: conditional expectation and unbiasedness. The points $\hat{\boldsymbol{x}}_t, \boldsymbol{y}$ and the Log vector are $\mathcal{F}_t$-measurable, while s_t is deterministic. The Log vector has norm at most $2\theta_{\mathrm{cap}}$, so Cauchy–Schwarz and the second-moment bound make the inner product integrable. Taking conditional expectation term by term gives

$$\mathbb{E}[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \mid \mathcal{F}_t] \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 + \mathbb{E}[2s_t \left\langle \tilde{\boldsymbol{g}}_t, \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle \mid \mathcal{F}_t] + (1+\theta_{\mathrm{cap}}) s_t^2 \mathbb{E}[\left\|\tilde{\boldsymbol{g}}_t\right\|^2 \mid \mathcal{F}_t].$$

Pulling the $\mathcal{F}_t$-measurable Log vector out of the inner product and applying conditional unbiasedness gives

$$\begin{aligned}\mathbb{E}[2s_t \left\langle \tilde{\boldsymbol{g}}_t, \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle \mid \mathcal{F}_t] &= 2s_t \left\langle \mathbb{E}[\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_t], \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle \\ &= 2s_t \left\langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t), \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle.\end{aligned} \tag{B.31}$$

Insertion into the conditional distance bound gives

$$\mathbb{E}[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \mid \mathcal{F}_t] \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 + 2s_t \left\langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t), \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle + (1+\theta_{\mathrm{cap}}) s_t^2 \mathbb{E}[\left\|\tilde{\boldsymbol{g}}_t\right\|^2 \mid \mathcal{F}_t].$$

Step 3: geodesic convexity. By geodesic convexity, Lemma B.6 gives

$$\left\langle \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t), \mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\boldsymbol{y}) \right\rangle \le L(\boldsymbol{y}) - L(\hat{\boldsymbol{x}}_t).$$

Substituting and multiplying through by $2s_t > 0$ gives

$$\mathbb{E}[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \mid \mathcal{F}_t] \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - 2s_t[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] + (1+\theta_{\mathrm{cap}}) s_t^2 \mathbb{E}[\left\|\tilde{\boldsymbol{g}}_t\right\|^2 \mid \mathcal{F}_t].$$

Rearranging:

$$2s_t[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - \mathbb{E}[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \mid \mathcal{F}_t] + (1+\theta_{\mathrm{cap}}) s_t^2 \mathbb{E}[\left\|\tilde{\boldsymbol{g}}_t\right\|^2 \mid \mathcal{F}_t].$$

Step 4: total expectation. Taking total expectation, with s_t deterministic, gives

$$2s_t\mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] \le \mathbb{E}d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - \mathbb{E}\big[\mathbb{E}[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \mid \mathcal{F}_t]\big] + (1+\theta_{\mathrm{cap}}) s_t^2 \mathbb{E}\Big[\mathbb{E}[\left\|\tilde{\boldsymbol{g}}_t\right\|^2 \mid \mathcal{F}_t]\Big].$$

The tower property collapses each nested conditional expectation. Dividing by two gives

$$s_t\mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] \le \frac{1}{2}\mathbb{E}\big[d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2\big] + \frac{1+\theta_{\mathrm{cap}}}{2} s_t^2 \mathbb{E}\left\|\tilde{\boldsymbol{g}}_t\right\|^2,$$

which is (B.30). □

B.4 From Tail Averages to the Compact Last-Iterate Bound

We follow the last-iterate argument of Defazio et al. (2023, Lemma 5 and Theorem 10), also used by Schaipp et al. (2025, Lemma D.1 and Theorem E.2), with Proposition B.10 supplying the spherical one-step inequality.

Proof idea. Proposition B.10 bounds expected progress at each step. Summing those bounds and telescoping the squared distances gives a bound on the weighted-average expected loss, but Theorem 2.3 needs the expected loss at the last iterate alone. The average could be small because early iterates are good, even if the final one is not. We therefore remove the earlier entries one at a time, renormalizing the remaining weights, until the average contains only the last iterate. Removing an entry with expected loss below the remaining average raises that average, so we must allow the upper bound to increase accordingly.

To bound this increase when removing $\hat{\boldsymbol{x}}_k$, we sum Proposition B.10 over $t = k, \dots, T$, choosing $\boldsymbol{y} = \hat{\boldsymbol{x}}_k$ throughout. This compares the weighted-average expected loss with the expected loss of the entry being removed. The resulting bound limits how much removing that entry can raise the average. The last iterate's expected loss equals the original average plus the successive changes in the average. Its upper bound is therefore the bound on the original average plus the sum of the upper bounds on those changes.

Recall $D = d_{\mathbb{S}}(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}^\star)$. Write $s_t = M\omega_t$ as in (2.7), and define the remaining allocation mass

$$W_k := \sum_{t=k}^{T} \omega_t. \tag{B.32}$$

Thus $W_1 = 1$, every W_k is positive, and $W_k - W_{k+1} = \omega_k$.

Proof of Theorem 2.3. Step 1: sum one-step progress on an arbitrary tail. Let $\boldsymbol{y}$ be a cap-valued comparator with integrable loss that is $\mathcal{F}_k$-measurable. It is then known before every draw on the tail $t \geq k$. Summing (B.30) gives

$$\begin{aligned} M \sum_{t=k}^{T} \omega_t \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] \leq \frac{1}{2} \sum_{t=k}^{T} \left(\mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \right) \\ + \frac{(1 + \theta_{\mathrm{cap}}) M^2}{2} \sum_{t=k}^{T} \omega_t^2 \mathbb{E} \left\| \tilde{\boldsymbol{g}}_t \right\|^2 . \end{aligned} \tag{B.33}$$

Consecutive distance terms cancel exactly:

$$\sum_{t=k}^{T} \left(\mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \boldsymbol{y})^2 - \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y})^2 \right) = \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_k, \boldsymbol{y})^2 - \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{T+1}, \boldsymbol{y})^2.$$

Using the unconditional bound $\mathbb{E} \left\| \tilde{\boldsymbol{g}}_t \right\|^2 \leq G^2$, we obtain

$$\begin{aligned} M \sum_{t=k}^{T} \omega_t \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\boldsymbol{y})] &\leq \frac{1}{2} \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_k, \boldsymbol{y})^2 - \frac{1}{2} \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{T+1}, \boldsymbol{y})^2 \\ &\quad + \frac{(1 + \theta_{\mathrm{cap}}) G^2 M^2}{2} \sum_{t=k}^{T} \omega_t^2 \\ &\leq \frac{1}{2} \mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_k, \boldsymbol{y})^2 + \frac{(1 + \theta_{\mathrm{cap}}) G^2 M^2}{2} \sum_{t=k}^{T} \omega_t^2. \end{aligned} \tag{B.34}$$

Starting at $\hat{\boldsymbol{x}}_1$, the first $T - 1$ updates take us to $\hat{\boldsymbol{x}}_T$, which is the iterate we care about. But Proposition B.10 bounds the loss at the point before an update. So, to include the loss at $\hat{\boldsymbol{x}}_T$, we apply it once more, using s_T, and introduce the next point $\hat{\boldsymbol{x}}_{T+1}$. This extra update is only part of the proof; it does not change $\hat{\boldsymbol{x}}_T$. Once we sum the inequalities, the term involving $\hat{\boldsymbol{x}}_{T+1}$ has a minus sign, so we can drop it when taking an upper bound. The coefficient s_T remains in the bound,

which means we can choose it to improve the bound without changing the iterate whose loss we are estimating. The left side of (B.34) is still a weighted sum; division by its tail mass will turn it into an average.

This step is the geodesic counterpart of Lemma E.1 of Schaipp et al. (2025).

Step 2: compare the full average to the optimum. Let b_t denote the expected excess loss at step t, and let A_k denote its weighted average over the tail beginning at k:

$$b_t := \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}^\star)], \qquad A_k := \frac{\sum_{t=k}^T \omega_t b_t}{W_k}. \tag{B.35}$$

Taking $k = 1$ and $\boldsymbol{y} = \hat{\boldsymbol{x}}^\star$ in (B.34), then dividing by $M > 0$, gives

$$A_1 = \sum_{t=1}^T \omega_t b_t \le \frac{D^2}{2M} + \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2} \sum_{t=1}^T \omega_t^2. \tag{B.36}$$

This controls the full average. The remaining steps account for the change in that average when early entries are removed.

Step 3: express the endpoint through changes of tail average. The final tail has $A_T = b_T$. Removing its first entry from tail k removes $\omega_k b_k$ from the weighted sum and ω_k from its mass, so

$$W_{k+1} A_{k+1} = W_k A_k - \omega_k b_k.$$

For $k < T$, dividing and subtracting A_k yields

$$\begin{aligned} A_{k+1} - A_k &= \frac{W_k A_k - \omega_k b_k - W_{k+1} A_k}{W_{k+1}} \\ &= \frac{(W_k - W_{k+1})A_k - \omega_k b_k}{W_{k+1}} = \frac{\omega_k}{W_{k+1}}(A_k - b_k). \end{aligned} \tag{B.37}$$

Here $A_{k+1} - A_k$ is the change in the weighted average when its first entry is removed, while $A_k - b_k$ compares the current tail average with its first entry. The factor ω_k / W_{k+1} is the removed fraction divided by the remaining allocation mass. Our target is $b_T = \mathbb{E}[L(\hat{\boldsymbol{x}}_T) - L(\hat{\boldsymbol{x}}^\star)]$. The last tail contains only this entry, so $A_T = b_T$. We therefore express the target as the full average plus the accumulated changes from removing earlier entries. Telescoping gives the weighted identity of Defazio et al. (2023, Lemma 5), reproduced by Schaipp et al. (2025, Lemma D.1):

$$\begin{aligned} b_T = A_T &= A_1 + \sum_{k=1}^{T-1} (A_{k+1} - A_k) \\ &= A_1 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_{k+1}} (A_k - b_k). \end{aligned} \tag{B.38}$$

Step 4: compare each tail to its own starting iterate. The identity (B.38) requires an upper bound on $A_k - b_k$. To obtain it, fix a tail starting at k and choose $\boldsymbol{y} = \hat{\boldsymbol{x}}_k$ in (B.34). This comparator has integrable loss: Lemma B.6 at $\hat{\boldsymbol{x}}_k$ with $\boldsymbol{y} = \hat{\boldsymbol{x}}^\star$, followed by Cauchy–Schwarz and the cap radius, gives $0 \le L(\hat{\boldsymbol{x}}_k) - L(\hat{\boldsymbol{x}}^\star) \le \theta_{\mathrm{cap}} \|\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_k)\|$; conditional unbiasedness and Jensen's inequality then bound $\mathbb{E}[L(\hat{\boldsymbol{x}}_k) - L(\hat{\boldsymbol{x}}^\star)] \le \theta_{\mathrm{cap}} G$. The same comparator is used for every step $t = k, \dots, T$. Substitution gives

$$\begin{aligned} M \sum_{t=k}^T \omega_t \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}_k)] &\le \frac{1}{2}\mathbb{E} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_k, \hat{\boldsymbol{x}}_k)^2 + \frac{(1+\theta_{\mathrm{cap}})G^2 M^2}{2} \sum_{t=k}^T \omega_t^2 \\ &= \frac{(1+\theta_{\mathrm{cap}})G^2 M^2}{2} \sum_{t=k}^T \omega_t^2. \end{aligned}$$

The initial distance is zero because the comparator is the tail's starting iterate. Dividing by the positive quantity $M W_k$ yields

$$\frac{\sum_{t=k}^T \omega_t \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}_k)]}{W_k} \le \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2W_k} \sum_{t=k}^T \omega_t^2.$$

To identify the left side, cancel the common optimum loss:

$$\begin{aligned}\mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}_k)] &= \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}^\star)] - \mathbb{E}[L(\hat{\boldsymbol{x}}_k) - L(\hat{\boldsymbol{x}}^\star)] \\ &= b_t - b_k.\end{aligned}$$

Since $\sum_{t=k}^{T} \omega_t = W_k$, subtracting b_k from the tail average gives

$$\begin{aligned}A_k - b_k &= \frac{\sum_{t=k}^{T} \omega_t b_t}{W_k} - \frac{W_k b_k}{W_k} \\ &= \frac{\sum_{t=k}^{T} \omega_t (b_t - b_k)}{W_k} \\ &= \frac{\sum_{t=k}^{T} \omega_t \mathbb{E}[L(\hat{\boldsymbol{x}}_t) - L(\hat{\boldsymbol{x}}_k)]}{W_k} \le \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2W_k} \sum_{t=k}^{T} \omega_t^2.\end{aligned} \tag{B.39}$$

This is the comparison with the starting iterate used by Defazio et al. (2023, App. F, derivation preceding Corollary 11), here obtained from Proposition B.10.

Step 5: insert the full-average and tail-start bounds. Substitute (B.36) and (B.39) into (B.38), retaining each positive multiplier:

$$\begin{aligned}\mathbb{E}[L(\hat{\boldsymbol{x}}_T) - L(\hat{\boldsymbol{x}}^\star)] &= A_1 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_{k+1}} (A_k - b_k) \\ &\le \frac{D^2}{2M} + \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2} \sum_{t=1}^{T} \omega_t^2 \\ &\quad + \sum_{k=1}^{T-1} \frac{\omega_k}{W_{k+1}} \left(\frac{(1+\theta_{\mathrm{cap}})G^2 M}{2W_k} \sum_{t=k}^{T} \omega_t^2 \right) \\ &= \frac{D^2}{2M} + \frac{(1+\theta_{\mathrm{cap}})G^2 M}{2} \left[\sum_{t=1}^{T} \omega_t^2 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_{k+1}} \frac{\sum_{t=k}^{T} \omega_t^2}{W_k} \right].\end{aligned} \tag{B.40}$$

This step is the geodesic counterpart of Theorem E.2 of Schaipp et al. (2025).

Step 6: express the bound using $C_T(\omega_{1:T})$. We now show that the bracket in Step 5 equals $C_T(\omega_{1:T})$. Each squared coefficient appears in several sums. Writing these sums out makes their occurrences explicit (the displayed expansion is for $T \ge 3$):

$$\begin{aligned}&\sum_{t=1}^{T} \omega_t^2 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_k W_{k+1}} \sum_{t=k}^{T} \omega_t^2 \\ &= (\omega_1^2 + \omega_2^2 + \cdots + \omega_T^2) \\ &\quad + \frac{\omega_1}{W_1 W_2} (\omega_1^2 + \omega_2^2 + \cdots + \omega_T^2) \\ &\quad + \frac{\omega_2}{W_2 W_3} (\omega_2^2 + \cdots + \omega_T^2) \\ &\quad + \cdots + \frac{\omega_{T-1}}{W_{T-1} W_T} (\omega_{T-1}^2 + \omega_T^2).\end{aligned}$$

For example, ω_1^2 occurs in the first two lines of the expanded expression, whereas ω_2^2 also occurs in the next line. Collecting their multipliers and continuing in the same way gives

$$\begin{aligned}&= \omega_1^2 \left(1 + \frac{\omega_1}{W_1 W_2}\right) \\ &\quad + \omega_2^2 \left(1 + \frac{\omega_1}{W_1 W_2} + \frac{\omega_2}{W_2 W_3}\right) \\ &\quad + \cdots + \omega_T^2 \left(1 + \frac{\omega_1}{W_1 W_2} + \cdots + \frac{\omega_{T-1}}{W_{T-1} W_T}\right).\end{aligned}$$

For every $T \geq 2$ and $t < T$, the coefficient ω_t^2 occurs in the initial sum and in the sums starting at $k = 1, \ldots, t$. Its multiplier is therefore

$$1 + \sum_{k=1}^{t} \frac{\omega_k}{W_k W_{k+1}}.$$

The final coefficient ω_T^2 occurs in every sum, but the outer sum stops at $T - 1$. Its multiplier is

$$1 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_k W_{k+1}}.$$

To simplify these multipliers, recall that

$$W_k = \sum_{j=k}^{T} \omega_j = \omega_k + \sum_{j=k+1}^{T} \omega_j = \omega_k + W_{k+1}.$$

Hence

$$\begin{aligned} \frac{\omega_k}{W_k W_{k+1}} &= \frac{W_k - W_{k+1}}{W_k W_{k+1}} \\ &= \frac{W_k}{W_k W_{k+1}} - \frac{W_{k+1}}{W_k W_{k+1}} \\ &= \frac{1}{W_{k+1}} - \frac{1}{W_k}. \end{aligned}$$

For $t < T$, substitution gives

$$\begin{aligned} 1 + \sum_{k=1}^{t} \frac{\omega_k}{W_k W_{k+1}} &= 1 + \sum_{k=1}^{t} \left(\frac{1}{W_{k+1}} - \frac{1}{W_k} \right) \\ &= 1 + \frac{1}{W_{t+1}} - \frac{1}{W_1} \\ &= \frac{1}{W_{t+1}}. \end{aligned}$$

The intermediate reciprocal terms cancel, and the last equality uses $W_1 = 1$. For the final coefficient, the same substitution with upper limit $T - 1$ gives

$$\begin{aligned} 1 + \sum_{k=1}^{T-1} \frac{\omega_k}{W_k W_{k+1}} &= 1 + \sum_{k=1}^{T-1} \left(\frac{1}{W_{k+1}} - \frac{1}{W_k} \right) \\ &= 1 + \frac{1}{W_T} - \frac{1}{W_1} \\ &= \frac{1}{W_T}. \end{aligned}$$

Substituting the multipliers back into the collected expression shows that the bracket in Step 5 equals

$$\begin{aligned} \sum_{t=1}^{T-1} \frac{\omega_t^2}{W_{t+1}} + \frac{\omega_T^2}{W_T} &= \sum_{t=1}^{T-1} \frac{\omega_t^2}{\sum_{j=t+1}^{T} \omega_j} + \omega_T \\ &= C_T(\omega_{1:T}), \end{aligned}$$

where $W_T = \omega_T$. For $T = 1$, the bracket is simply $\omega_1^2 = 1 = C_1(\omega_{1:1})$. Thus Step 5 becomes

$$\mathbb{E}[L(\hat{\boldsymbol{x}}_T) - L(\hat{\boldsymbol{x}}^\star)] \leq \frac{D^2}{2M} + \frac{(1 + \theta_{\text{cap}}) G^2 M}{2} C_T(\omega_{1:T}),$$

which is the claimed bound (2.10).

Finally, using $s_t = M\omega_t$, the same cost can be written in the original coefficients:

$$\begin{aligned} MC_T(\omega_{1:T}) &= M\omega_T + M\sum_{k=1}^{T-1}\frac{\omega_k^2}{\sum_{t=k+1}^{T}\omega_t} \\ &= s_T + \sum_{k=1}^{T-1}\frac{M^2\omega_k^2}{M\sum_{t=k+1}^{T}\omega_t} \\ &= s_T + \sum_{k=1}^{T-1}\frac{s_k^2}{\sum_{t=k+1}^{T}s_t}. \end{aligned} \tag{B.41}$$

□

The scalar last-iterate conversion above is the general-schedule argument of Defazio et al. (2023, Theorem 10) and Schaipp et al. (2025, Theorem 3.1 and App. E). The bound (2.10) depends on total mass M and normalized allocation $\omega_{1:T}$ separately and the next subsection optimizes M at fixed $\omega_{1:T}$.

B.5 Separating Total Mass from Allocation

Changing the schedule changes the visited points, so moments measured on one trajectory do not by themselves bound those on another one; as remarked by Schaipp et al. (2025, Remark 3.2), we therefore need to use a common oracle envelope G.

Every positive coefficient sequence has the unique decomposition

$$M = \sum_{t=1}^{T} s_t, \qquad \omega_t = \frac{s_t}{M}, \qquad s_t = M\omega_t, \qquad \sum_{t=1}^{T}\omega_t = 1.$$

Proof of Proposition 2.4. Fix a positive unit-mass allocation $\omega_{1:T}$ and $D, G > 0$. The preceding identities transform the bound into

$$\Omega_T(M, \omega_{1:T}) = \frac{D^2}{2M} + \frac{(1+\theta_{\text{cap}})G^2}{2} MC_T(\omega_{1:T}).$$

Here only $M > 0$ varies; the allocation and its cost are fixed. Differentiating gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}M}\Omega_T(M, \omega_{1:T}) &= -\frac{D^2}{2M^2} + \frac{(1+\theta_{\text{cap}})G^2}{2} C_T(\omega_{1:T}), \\ \frac{\mathrm{d}^2}{\mathrm{d}M^2}\Omega_T(M, \omega_{1:T}) &= \frac{D^2}{M^3} > 0. \end{aligned} \tag{B.42}$$

Setting the first derivative to zero yields

$$M^2 = \frac{D^2}{(1+\theta_{\text{cap}})G^2 C_T(\omega_{1:T})}, \qquad M^\star(\omega_{1:T}) = \frac{D}{G\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}}.$$

Strict convexity makes this the unique global minimum. The first objective term diverges as $M \downarrow 0$, and the second diverges as $M \to \infty$, so neither boundary supplies a competing minimum. Substitution into the two terms separately gives

$$\begin{aligned} \frac{D^2}{2M^\star} &= \frac{D^2 G\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}}{2D} = \frac{DG}{2}\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}, \\ \frac{(1+\theta_{\text{cap}})G^2}{2} M^\star C_T(\omega_{1:T}) &= \frac{(1+\theta_{\text{cap}})G^2 DC_T(\omega_{1:T})}{2G\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}} = \frac{DG}{2}\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}. \end{aligned} \tag{B.43}$$

Adding these equal contributions proves the minimized bound is $DG\sqrt{(1+\theta_{\text{cap}})C_T(\omega_{1:T})}$. □

C Proofs for Schedule Optimization

App. B.5 reduced the schedule-design problem to minimizing the allocation cost $C_T(\omega_{1:T})$ over positive unit-mass sequences. Sec. 3 stated the resulting optimality and convergence results; we now supply their full proofs, together with the exact cost of the prescribed linear schedule used in the asymptotic comparison.

C.1 Exact Finite-Horizon Optimization

The main text reduced the allocation problem to the scalar optimization (3.3) via Lemma 3.1; we first prove the cost and uniqueness theorem by induction, then prove the corollary that constructs the full allocation.

Proof of Theorem 3.2. We first prove by induction on n that $C_n^\star > 0$ and that exactly one positive unit-mass allocation attains $C_n^\star$. The induction step also yields the cost recursion.

Base case: $n = 1$. The only positive unit-mass allocation is $\omega_1 = 1$. Its cost is $C_1(\omega_1) = 1$, so $C_1^\star = 1 > 0$ is attained uniquely.

Induction hypothesis. Fix $n \geq 2$ and assume the claim holds for length $n-1$. Using indices $2, \ldots, n$ for this shorter allocation, the hypothesis is

$$C_{n-1}^\star > 0, \qquad \exists!\, p_{2:n} \text{ such that } \quad p_t > 0 \ (2 \leq t \leq n), \quad \sum_{t=2}^{n} p_t = 1, \quad C_{n-1}(p_{2:n}) = C_{n-1}^\star.$$

Induction step: from length $n-1$ to length n.

Step 1: apply the first-step decomposition lemma. Fix $0 < \omega_1 < 1$. This leaves a budget of $1-\omega_1$ for entries $2, \ldots, n$. Recall the normalized tail proportions p_t from Lemma 3.1: divide each remaining entry by that remaining budget,

$$p_t = \frac{\omega_t}{1-\omega_1} \quad (2 \leq t \leq n), \qquad \sum_{t=2}^{n} p_t = \frac{\sum_{t=2}^{n} \omega_t}{1-\omega_1} = 1.$$

Applying Lemma 3.1 to this allocation gives

$$C_n(\omega_{1:n}) = \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1) C_{n-1}(p_{2:n}).$$

By the induction hypothesis, minimizing over $p_{2:n}$ gives $C_{n-1}^\star$, attained uniquely. Hence

$$C_n(\omega_{1:n}) \geq \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1) C_{n-1}^\star.$$

Equality holds when the normalized tail is the unique length-$(n-1)$ optimizer.

Step 2: find the optimal first fraction and cost. Rewrite the scalar objective (3.3) as

$$\frac{\omega_1^2}{1-\omega_1} + (1-\omega_1) C_{n-1}^\star = \frac{1}{1-\omega_1} - 2 + (1 + C_{n-1}^\star)(1-\omega_1).$$

Its first and second derivatives are

$$\frac{\mathrm{d}}{\mathrm{d}\omega_1}\left[\frac{1}{1-\omega_1} - 2 + (1 + C_{n-1}^\star)(1-\omega_1)\right] = \frac{1}{(1-\omega_1)^2} - (1 + C_{n-1}^\star),$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\omega_1^2}\left[\frac{1}{1-\omega_1} - 2 + (1 + C_{n-1}^\star)(1-\omega_1)\right] = \frac{2}{(1-\omega_1)^3} > 0.$$

The objective is strictly convex. Setting its first derivative to zero gives

$$(1-\omega_1)^2 = \frac{1}{1 + C_{n-1}^\star}, \qquad \omega_1 = 1 - \frac{1}{\sqrt{1 + C_{n-1}^\star}}.$$

Since $C^\star_{n-1} > 0$, we have $0 < 1/\sqrt{1 + C^\star_{n-1}} < 1$. Thus the stationary point lies in the admissible interval $0 < \omega_1 < 1$. Strict convexity on this interval then makes it the unique global minimizer of the scalar objective.

With $1 - \omega_1 = 1/\sqrt{1 + C^\star_{n-1}}$,

$$\begin{aligned} C^\star_n &= \left[\frac{1}{1-\omega_1} - 2 + (1 + C^\star_{n-1})(1-\omega_1)\right]\Bigg|_{\omega_1 = 1 - 1/\sqrt{1+C^\star_{n-1}}} \\ &= \sqrt{1 + C^\star_{n-1}} - 2 + \frac{1 + C^\star_{n-1}}{\sqrt{1 + C^\star_{n-1}}} \\ &= 2\big(\sqrt{1 + C^\star_{n-1}} - 1\big) > 0. \end{aligned}$$

Step 3: complete the induction step.

(a) Attainment. Choose the length-$(n-1)$ allocation $p_{2:n}$ satisfying

$$p_t > 0, \qquad \sum_{t=2}^{n} p_t = 1, \qquad C_{n-1}(p_{2:n}) = C^\star_{n-1}.$$

Scale these proportions by the remaining budget:

$$\omega_1 = 1 - \frac{1}{\sqrt{1 + C^\star_{n-1}}}, \qquad \omega_t = (1-\omega_1)p_t \quad (2 \le t \le n).$$

Since $0 < \omega_1 < 1$, all entries are positive. Their sum is

$$\sum_{t=1}^{n} \omega_t = \omega_1 + (1-\omega_1)\sum_{t=2}^{n} p_t = \omega_1 + (1-\omega_1) = 1.$$

For this allocation, Lemma 3.1 gives

$$\begin{aligned} C_n(\omega_{1:n}) &= \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1)C_{n-1}(p_{2:n}) \\ &= \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1)C^\star_{n-1} = C^\star_n. \end{aligned}$$

The second equality uses the tail's optimality; the last uses the scalar minimum from Step 2.

(b) Show that every attaining allocation is the one just constructed. Take any positive unit-mass allocation $\omega_{1:n}$, and normalize its tail as in Step 1. The two minimizations give

$$C_n(\omega_{1:n}) \ge \frac{\omega_1^2}{1-\omega_1} + (1-\omega_1)C^\star_{n-1} \ge C^\star_n.$$

If this allocation attains $C^\star_n$, both inequalities must be equalities. Equality in the second inequality fixes its first entry: by the strict convexity proved in Step 2, it must have $\omega_1 = 1 - 1/\sqrt{1 + C^\star_{n-1}}$.

Equality in the first inequality requires $C_{n-1}(p_{2:n}) = C^\star_{n-1}$, since $1 - \omega_1 > 0$. Here we use the uniqueness part of the induction hypothesis: there is exactly one such normalized tail. With both ω_1 and $p_{2:n}$ fixed, the relations $\omega_t = (1-\omega_1)p_t$ fix every remaining entry. Hence every attaining allocation equals the one constructed in (a).

(c) Conclude the induction. Attainment, uniqueness, and $C^\star_n > 0$ for length n are established; the base case completes the induction, and the cost recursion (3.4) follows from the displayed calculation in Step 2. □

Proof of Corollary 3.3. The scalar minimization in Step 2 of the proof of Theorem 3.2 shows that, for a unit-mass allocation of length $n \ge 2$, the optimal first fraction is $1 - 1/\sqrt{1 + C^\star_{n-1}}$. For $T = 1$, the only allocation is $\omega^\star_1 = 1$, as stated. We now construct the allocation for $T \ge 2$.

Step 1: apply the minimizing fraction to each remaining tail. Part (b) of Step 3 in the proof of Theorem 3.2 shows that an allocation attains the minimum only if its normalized tail attains the

minimum for the shorter problem. Applying this result to the length-T optimizer shows that the proportions $p_{2:T}$ from Lemma 3.1, now evaluated at that optimizer, satisfy

$$(p_2, \dots, p_T) = \left(\frac{\omega_2^\star}{1-\omega_1^\star}, \dots, \frac{\omega_T^\star}{1-\omega_1^\star} \right)$$

and form the optimal unit-mass allocation of length $T-1$. For $T \geq 3$, the first entry of this shorter allocation is $p_2 = \omega_2^\star/(1-\omega_1^\star)$. Applying the same result again normalizes each subsequent entry as

$$\frac{p_t}{1-p_2} = \frac{\omega_t^\star/(1-\omega_1^\star)}{1-\omega_2^\star/(1-\omega_1^\star)} = \frac{\omega_t^\star}{1-\omega_1^\star-\omega_2^\star}, \qquad 3 \leq t \leq T.$$

Consequently,

$$C_{T-2}\left(\frac{\omega_3^\star}{1-\omega_1^\star-\omega_2^\star}, \dots, \frac{\omega_T^\star}{1-\omega_1^\star-\omega_2^\star} \right) = C_{T-2}^\star.$$

Repeating this normalization gives, for every $1 \leq k \leq T$,

$$C_{T-k+1}\left(\frac{\omega_k^\star}{1-\sum_{j=1}^{k-1}\omega_j^\star}, \dots, \frac{\omega_T^\star}{1-\sum_{j=1}^{k-1}\omega_j^\star} \right) = C_{T-k+1}^\star.$$

Write the remaining mass before step k as

$$W_k^\star := \sum_{t=k}^{T} \omega_t^\star = 1 - \sum_{j=1}^{k-1} \omega_j^\star,$$

so $W_1^\star = 1$. For $k < T$, the normalized remaining allocation is $\left(\omega_k^\star/W_k^\star, \dots, \omega_T^\star/W_k^\star\right)$. It has $T-k+1$ entries, and its first entry is $\omega_k^\star/W_k^\star$. Step 2 of the theorem's proof gives the minimizing first entry of any optimal unit-mass allocation of length n as $1 - 1/\sqrt{1+C_{n-1}^\star}$. Applying that formula here means substituting $n = T-k+1$ and identifying the first entry with $\omega_k^\star/W_k^\star$:

$$\frac{\omega_k^\star}{W_k^\star} = 1 - \frac{1}{\sqrt{1+C_{(T-k+1)-1}^\star}} = 1 - \frac{1}{\sqrt{1+C_{T-k}^\star}}.$$

The fraction left after this entry is $1 - \omega_k^\star/W_k^\star = W_{k+1}^\star/W_k^\star$, because $W_{k+1}^\star = W_k^\star - \omega_k^\star$. The assigned and retained fractions are therefore

$$\frac{\omega_k^\star}{W_k^\star} = 1 - \frac{1}{\sqrt{1+C_{T-k}^\star}}, \qquad \frac{W_{k+1}^\star}{W_k^\star} = 1 - \frac{\omega_k^\star}{W_k^\star} = \frac{1}{\sqrt{1+C_{T-k}^\star}}. \tag{C.1}$$

Step 2: reconstruct the allocation and verify its total mass. Start with the full unit budget, $W_1^\star = 1$. Suppose $W_k^\star$, the mass available before step $k < T$, is known. The second identity in (C.1) gives the mass left after that step:

$$W_{k+1}^\star = \frac{W_k^\star}{\sqrt{1+C_{T-k}^\star}}.$$

Therefore

$$\omega_k^\star = W_k^\star - W_{k+1}^\star = W_k^\star \left(1 - \frac{1}{\sqrt{1+C_{T-k}^\star}} \right).$$

These two formulas determine the next remaining mass and the current coefficient from $W_k^\star$. Apply them for $k = 1, \dots, T-1$; the last entry receives the remaining mass, $\omega_T^\star = W_T^\star$. Since $C_{T-k}^\star > 0$, each step satisfies $0 < W_{k+1}^\star < W_k^\star$, so every coefficient is positive. Their sum is

$$\sum_{k=1}^{T} \omega_k^\star = \sum_{k=1}^{T-1} (W_k^\star - W_{k+1}^\star) + W_T^\star = W_1^\star = 1.$$

Thus the construction yields a feasible allocation, and each of its tails uses the unique minimizing split. Repeated application of Lemma 3.1, ending at the one-entry tail, therefore gives the minimum cost at every tail. The construction is thus the unique optimizer of Theorem 3.2. Since its total mass is one, substituting $W_k^\star = 1 - \sum_{j=1}^{k-1}\omega_j^\star$ into (C.1) gives the coefficient recursion (3.5); the terminal identity $\omega_T^\star = W_T^\star = 1 - \sum_{j=1}^{T-1}\omega_j^\star$ gives its final entry. □

For $D, G > 0$, combining this allocation with the unique optimal mass in Proposition 2.4 gives the full schedule

$$s_t^\star = M^\star(\omega_{1:T}^\star)\omega_t^\star.$$

Indeed, the amplitude-optimized bound is $DG\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T})}$, which strictly increases with $C_T(\omega_{1:T})$. It therefore has the same unique minimizing allocation.

C.2 From Optimal Cost to the Linear Profile

We first prove the optimal-cost asymptotics, then use the allocation recursion to establish convergence to linear decay.

Proof of Proposition 3.4. Part 1: asymptotic optimal cost. We begin with (3.4) from Theorem 3.2 (Optimal cost and uniqueness). Rearranging and squaring gives, for $n \geq 2$,

$$\sqrt{1 + C_{n-1}^\star} = 1 + \frac{C_n^\star}{2}, \qquad 1 + C_{n-1}^\star = 1 + C_n^\star + \frac{(C_n^\star)^2}{4}.$$

Hence the decrease between successive costs satisfies the exact identity

$$C_{n-1}^\star = C_n^\star + \frac{(C_n^\star)^2}{4}. \tag{C.2}$$

The costs are positive by Theorem 3.2, so this identity shows that they strictly decrease. They therefore converge to a limit $\ell \geq 0$. Both $C_{n-1}^\star$ and $C_n^\star$ converge to ℓ, so taking limits in (C.2) gives

$$\ell = \ell + \frac{\ell^2}{4} \quad \Longrightarrow \quad \ell^2 = 0 \quad \Longrightarrow \quad \ell = 0.$$

Thus $C_n^\star \to 0$.

We now prove the sharper claim $C_n^\star \sim 4/n$. Because the costs are positive, this claim is equivalent to

$$C_n^\star \sim \frac{4}{n} \quad \Longleftrightarrow \quad \frac{1/C_n^\star}{n} \longrightarrow \frac{1}{4}.$$

The expression on the right is a ratio of two sequences. We evaluate it with the Stolz–Cesàro theorem, which states that if b_n is strictly increasing and unbounded and

$$\frac{a_n - a_{n-1}}{b_n - b_{n-1}} \longrightarrow L,$$

then $a_n/b_n \to L$.

Apply this theorem with

$$a_n = \frac{1}{C_n^\star}, \qquad b_n = n.$$

The denominator $b_n = n$ is strictly increasing and unbounded, and $b_n - b_{n-1} = 1$. It remains to calculate the numerator increment. Using (C.2),

$$\begin{aligned} a_n - a_{n-1} &= \frac{1}{C_n^\star} - \frac{1}{C_{n-1}^\star} \\ &= \frac{C_{n-1}^\star - C_n^\star}{C_n^\star C_{n-1}^\star} = \frac{C_n^\star}{4C_{n-1}^\star} = \frac{1}{4 + C_n^\star}. \end{aligned}$$

The last equality uses $C_{n-1}^\star = C_n^\star(1 + C_n^\star/4)$. Since $C_n^\star \to 0$, continuity gives

$$\frac{a_n - a_{n-1}}{b_n - b_{n-1}} = \frac{1}{4 + C_n^\star} \longrightarrow \frac{1}{4}.$$

All hypotheses of Stolz–Cesàro are therefore satisfied, and it yields

$$\frac{a_n}{b_n} = \frac{1/C_n^\star}{n} = \frac{1}{nC_n^\star} \longrightarrow \frac{1}{4}.$$

This limit is positive, so taking reciprocals gives

$$nC_n^\star \longrightarrow 4. \tag{C.3}$$

In particular, setting $n = T$ proves the optimal-cost claim of the proposition.

Part 2: convergence to linear decay. Fix $0 < \tau < 1$. For each $T \geq 2$, write $k = 1 + \lfloor \tau(T-1) \rfloor$ for the index of the coefficient at that position. We must prove

$$T\omega_k^\star \longrightarrow 2(1-\tau).$$

The indices in the tail are $k, k+1, \ldots, T$, so this tail contains exactly $T - k + 1$ entries. Moreover,

$$\frac{k}{T} = \frac{1 + \lfloor \tau(T-1) \rfloor}{T} \longrightarrow \tau, \qquad \frac{T-k+1}{T} \longrightarrow 1 - \tau > 0,$$

and hence $T - k + 1 \to \infty$. This allows us to apply the cost asymptotic from Part 1 to the optimal cost at this shorter horizon.

Recall that

$$W_k^\star = \sum_{t=k}^{T} \omega_t^\star$$

is the mass available before entry k. The coefficient separates as

$$\omega_k^\star = W_k^\star \frac{\omega_k^\star}{W_k^\star}.$$

Here $\omega_k^\star / W_k^\star$ is the fraction of the available mass assigned to entry k. We first evaluate this fraction using Part 1.

The normalized tail beginning at k has $T - k + 1$ entries. Substituting $n = T - k + 1$ into (3.4) gives

$$C_{T-k+1}^\star = 2\big(\sqrt{1 + C_{T-k}^\star} - 1\big), \qquad \sqrt{1 + C_{T-k}^\star} = 1 + \frac{C_{T-k+1}^\star}{2}.$$

Using this identity in (C.1), proved for Corollary 3.3 (Construction of the optimal allocation), yields

$$\begin{aligned} \frac{\omega_k^\star}{W_k^\star} &= 1 - \frac{1}{\sqrt{1 + C_{T-k}^\star}} = 1 - \frac{1}{1 + C_{T-k+1}^\star/2} \\ &= \frac{C_{T-k+1}^\star}{2 + C_{T-k+1}^\star}. \end{aligned}$$

Thus

$$\omega_k^\star = W_k^\star \frac{C_{T-k+1}^\star}{2 + C_{T-k+1}^\star}.$$

To use the asymptotic $(T - k + 1)C_{T-k+1}^\star \to 4$, multiply and divide by the remaining length $T - k + 1$:

$$T\frac{\omega_k^\star}{W_k^\star} = \frac{T}{T-k+1} \frac{(T-k+1)C_{T-k+1}^\star}{2 + C_{T-k+1}^\star}.$$

By (C.3),

$$(T-k+1)C_{T-k+1}^\star \longrightarrow 4, \qquad C_{T-k+1}^\star \longrightarrow 0.$$

Together with $T/(T-k+1) \to 1/(1-\tau)$, these limits give

$$T\frac{\omega_k^\star}{W_k^\star} \longrightarrow \frac{2}{1-\tau}.$$

We have that

$$T\omega_k^\star = W_k^\star \left(T\frac{\omega_k^\star}{W_k^\star} \right),$$

and we have proved that

$$T\frac{\omega_k^\star}{W_k^\star} \longrightarrow \frac{2}{1-\tau}.$$

Thus the only quantity in the product whose limit remains unknown is $W_k^\star$. The target $T\omega_k^\star \to 2(1-\tau)$ will follow once we prove

$$W_k^\star \longrightarrow (1-\tau)^2.$$

To determine $W_k^\star$, we accumulate the fraction of mass retained after each of the preceding entries. The second identity in (C.1) gives, for each $1 \le i < k$,

$$\frac{W_{i+1}^\star}{W_i^\star} = (1 + C_{T-i}^\star)^{-1/2}.$$

The total fraction retained after entries $1, \dots, k-1$ is the product of their successive retained fractions. On the left, all intermediate masses cancel:

$$\frac{W_2^\star}{W_1^\star}\frac{W_3^\star}{W_2^\star}\cdots\frac{W_k^\star}{W_{k-1}^\star} = \frac{W_k^\star}{W_1^\star}.$$

Therefore,

$$\begin{aligned}
\frac{W_k^\star}{W_1^\star} &= \prod_{i=1}^{k-1}\frac{W_{i+1}^\star}{W_i^\star} = \prod_{i=1}^{k-1}(1 + C_{T-i}^\star)^{-1/2}, \\
W_k^\star &= \prod_{m=T-k+1}^{T-1}(1 + C_m^\star)^{-1/2}, \qquad \text{(C.4)} \\
\log W_k^\star &= -\frac{1}{2}\sum_{m=T-k+1}^{T-1}\log(1 + C_m^\star).
\end{aligned}$$

The second line uses $W_1^\star = 1$ and reindexes with $m = T - i$; reversing the order lists m from $T-k+1$ through $T-1$.

We evaluate the sum in (C.4) by recognizing it as a Riemann sum for $\int_{1-\tau}^{1} 4/x\,\mathrm{d}x$. To make this explicit, rewrite each summand as

$$\log(1 + C_m^\star) = \frac{1}{T}\left[\frac{T}{m}\,(mC_m^\star)\,\frac{\log(1 + C_m^\star)}{C_m^\star}\right].$$

Part 1 gives $mC_m^\star \to 4$. It also gives $C_m^\star \to 0$, so the standard limit $\log(1+x)/x \to 1$ as $x \to 0$ implies

$$\frac{\log(1 + C_m^\star)}{C_m^\star} \longrightarrow 1.$$

To replace the actual rectangle heights by $4/(m/T)$ in the entire sum, we must show that the largest replacement error over the summation range tends to zero. Pointwise control of one fixed summand would not suffice because the number and indices of the summands depend on T.

Since $(T-k+1)/T \to 1-\tau > 0$, for all sufficiently large T, every summation index satisfies

$$m \ge T - k + 1 \ge \frac{1-\tau}{2}\,T.$$

The smallest index therefore tends to infinity. By the definition of ordinary sequence convergence, the two limits above then hold simultaneously for every index in the range:

$$\sup_{T-k+1 \le m \le T-1} |mC_m^\star - 4| \longrightarrow 0,$$

$$\sup_{T-k+1 \le m \le T-1} \left|\frac{\log(1 + C_m^\star)}{C_m^\star} - 1\right| \longrightarrow 0.$$

Moreover,

$$\frac{T}{m} \le \frac{2}{1-\tau}$$

throughout the range, so multiplication by T/m cannot enlarge these vanishing errors without bound. Since the convergent sequence $mC_m^\star$ is also bounded, the product estimate

$$\left|(mC_m^\star)\frac{\log(1+C_m^\star)}{C_m^\star}-4\right| \le |mC_m^\star|\left|\frac{\log(1+C_m^\star)}{C_m^\star}-1\right| + |mC_m^\star-4|$$

shows that

$$\sup_{T-k+1\le m\le T-1}\left|\frac{T}{m}(mC_m^\star)\frac{\log(1+C_m^\star)}{C_m^\star}-\frac{4}{m/T}\right|\longrightarrow 0.$$

This controls the replacement error in the complete sum:

$$\left|\frac{1}{T}\sum_{m=T-k+1}^{T-1}\left[\frac{T}{m}(mC_m^\star)\frac{\log(1+C_m^\star)}{C_m^\star}-\frac{4}{m/T}\right]\right| \le \frac{k-1}{T}\sup_{T-k+1\le m\le T-1}\left|\frac{T}{m}(mC_m^\star)\frac{\log(1+C_m^\star)}{C_m^\star}-\frac{4}{m/T}\right|\longrightarrow 0.$$

There are $k-1$ summands, and $(k-1)/T\to\tau$, so this prefactor remains bounded.

To recall why this is a Riemann sum, the points m/T are spaced by $1/T$. Thus

$$\frac{1}{T}\sum_{m=T-k+1}^{T-1}\frac{4}{m/T}$$

adds the areas of rectangles of width $1/T$ and height $4/(m/T)$. For each T, the rectangles cover the interval

$$\left[\frac{T-k+1}{T},1\right].$$

Because $(T-k+1)/T\to 1-\tau>0$, for all sufficiently large T this interval is contained in the fixed interval $[(1-\tau)/2,1]$, where $4/x$ is continuous. The Riemann-sum theorem therefore gives

$$\frac{1}{T}\sum_{m=T-k+1}^{T-1}\frac{4}{m/T}-\int_{(T-k+1)/T}^{1}\frac{4}{x}\,\mathrm{d}x\longrightarrow 0.$$

The lower endpoint of this integral converges to $1-\tau$. Continuity of $4/x$ on the fixed interval above then gives

$$\int_{(T-k+1)/T}^{1}\frac{4}{x}\,\mathrm{d}x\longrightarrow\int_{1-\tau}^{1}\frac{4}{x}\,\mathrm{d}x.$$

Finally, the replacement-error bound established above shows that using the actual rectangle heights instead of $4/(m/T)$ does not change this limit. Therefore

$$\log W_k^\star=-\frac{1}{2T}\sum_{m=T-k+1}^{T-1}\left[\frac{T}{m}\,(mC_m^\star)\,\frac{\log(1+C_m^\star)}{C_m^\star}\right] \longrightarrow -\frac{1}{2}\int_{1-\tau}^{1}\frac{4}{x}\,\mathrm{d}x=2\log(1-\tau).$$

Exponentiating gives

$$\log W_k^\star\longrightarrow 2\log(1-\tau),\qquad W_k^\star\longrightarrow(1-\tau)^2. \tag{C.5}$$

Combining the assigned-fraction limit established above with (C.5) now gives

$$T\omega_k^\star=W_k^\star\left(T\frac{\omega_k^\star}{W_k^\star}\right)\longrightarrow(1-\tau)^2\frac{2}{1-\tau}=2(1-\tau). \tag{C.6}$$

For $\tau=0$, the selected coefficient is $\omega_1^\star$. Using $W_1^\star=1$ in the exact assigned-fraction identity and applying (C.3) gives

$$T\omega_1^\star=\frac{TC_T^\star}{2+C_T^\star}\longrightarrow\frac{4}{2}=2.$$

This proves (3.6) for every fixed $0 \le \tau < 1$.

Dividing the coefficient asymptotic by its value at $\tau = 0$ also gives

$$\frac{\omega^\star_{1+\lfloor \tau(T-1)\rfloor}}{\omega^\star_1} \longrightarrow 1-\tau \qquad \text{for each fixed } 0 \le \tau < 1.$$

□

C.3 Performance of Linear Decay

The exact optimizer approaches linear decay; we now verify that the closed-form linear allocation has the same leading cost $4/T$. Start from the linear shape $1-\tau$, where τ is a fraction of the horizon. At step $t = 1, \dots, T$, set $\tau = (t-1)/T$. The linear weights are then

$$q_t = 1 - \frac{t-1}{T} = \frac{T+1-t}{T}.$$

These weights decrease from one to $1/T$, reaching zero at the virtual index $T+1$. Their total mass is

$$\sum_{t=1}^{T} q_t = \frac{1}{T}\sum_{t=1}^{T}(T+1-t) = \frac{1}{T}\frac{T(T+1)}{2} = \frac{T+1}{2}.$$

Dividing each weight by this sum preserves its linear shape and gives the unit-mass allocation

$$\omega_t^{\mathrm{lin}} = \frac{q_t}{\sum_{j=1}^{T} q_j} = \frac{(T+1-t)/T}{(T+1)/2} = \frac{2(T+1-t)}{T(T+1)}, \qquad 1 \le t \le T. \tag{C.7}$$

Every entry is positive (the smallest is $\omega_T^{\mathrm{lin}} = 2/[T(T+1)]$), so every following-tail denominator in the allocation cost is nonzero.

Since $(T+1-k)/(T-k) = 1 + 1/(T-k)$, the cost sum will reduce to a harmonic series. The tail mass after step k is

$$\sum_{t=k+1}^{T} \omega_t^{\mathrm{lin}} = \frac{2}{T(T+1)} \sum_{t=k+1}^{T} (T+1-t) = \frac{(T-k)(T-k+1)}{T(T+1)}.$$

Therefore, the contribution of step k to the cost is

$$\frac{(\omega_k^{\mathrm{lin}})^2}{\sum_{t=k+1}^{T} \omega_t^{\mathrm{lin}}} = \frac{4(T+1-k)}{T(T+1)(T-k)}.$$

Adding these contributions to the terminal term $\omega_T^{\mathrm{lin}} = 2/[T(T+1)]$ gives

$$\begin{aligned} C_T(\omega_{1:T}^{\mathrm{lin}}) &= \frac{2}{T(T+1)} + \frac{4}{T(T+1)} \sum_{k=1}^{T-1} \frac{T+1-k}{T-k} \\ &= \frac{2}{T(T+1)} + \frac{4}{T(T+1)} \left(T - 1 + \sum_{k=1}^{T-1} \frac{1}{k} \right) \\ &= \frac{4}{T} + \frac{4\sum_{k=1}^{T-1} k^{-1} - 6}{T(T+1)}. \end{aligned} \tag{C.8}$$

For $T = 1$, the sums are empty, and the cost is one, as required.

For $T \ge 2$, integral comparison gives $\sum_{k=1}^{T-1} k^{-1} \le 1 + \log(T-1)$. The correction in (C.8) is therefore $O(\log T/T^2)$. Multiplying the cost by T makes that correction vanish, so $TC_T(\omega_{1:T}^{\mathrm{lin}}) \to 4$. Comparing with the exact optimal cost,

$$\frac{C_T(\omega_{1:T}^{\mathrm{lin}})}{C_T^\star} = \frac{TC_T(\omega_{1:T}^{\mathrm{lin}})}{TC_T^\star} \longrightarrow \frac{4}{4} = 1.$$

Optimizing each schedule's amplitude separately via Proposition 2.4,

$$\frac{\min_{M>0}\Omega_T(M,\omega_{1:T}^{\mathrm{lin}})}{\min_{M>0,\,\omega_{1:T}}\Omega_T(M,\omega_{1:T})} = \frac{DG\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T}^{\mathrm{lin}})}}{DG\sqrt{(1+\theta_{\mathrm{cap}})C_T^\star}} = \sqrt{\frac{C_T(\omega_{1:T}^{\mathrm{lin}})}{C_T^\star}} \longrightarrow 1. \tag{C.9}$$

The remaining cost ratio tends to one, so prescribed linear decay, with its own optimal amplitude, attains the same leading bound as the exact finite-horizon optimizer.

Both bounds have leading value

$$DG\sqrt{(1+\theta_{\mathrm{cap}})\frac{4}{T}} = \frac{2DG\sqrt{1+\theta_{\mathrm{cap}}}}{\sqrt{T}},$$

optimizing the general-schedule bound inherited from Defazio et al. (2023, Theorem 10) and Schaipp et al. (2025, Theorem 3.1) under the spherical oracle and cap-containment hypotheses of Theorem 2.3.

C.4 Schedule and Bound Trajectories

Fig. 6 compares the principal schedule families at one horizon.

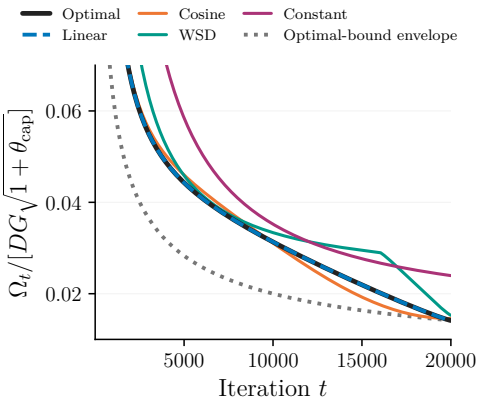


Figure 6: Normalized last-iterate bounds at $T = 20{,}000$. Each schedule uses its final-horizon optimal amplitude; the dotted envelope reoptimizes allocation and amplitude at every iteration.

For a horizon T, normalize each of the following positive sequences to have unit sum:

$$\omega_t^{\mathrm{const}} \propto 1, \qquad \omega_t^{\mathrm{lin}} \propto 1 - \frac{t-1}{T}, \qquad \omega_t^{\cos} \propto \frac{1+\cos(\pi(t-1)/T)}{2}, \qquad 1 \le t \le T. \tag{C.10}$$

For the horizons below, set $T_0 = 4T/5$. Following Schaipp et al. (2025, Eq. (3)), the WSD shape has no warmup, stays constant through most of the horizon, and then uses a linear cooldown:

$$\omega_t^{\mathrm{WSD}} \propto \begin{cases} 1, & 1 \le t < T_0, \\ 1 - \dfrac{t-T_0}{T+1-T_0}, & T_0 \le t \le T. \end{cases} \tag{C.11}$$

The proportionality constant in each line is fixed by $\sum_{t=1}^{T}\omega_t = 1$. The linear, cosine, and WSD sequences reach zero at the virtual index $T+1$, whereas the constant sequence remains constant. The exact allocation is $\omega_{1:T}^\star$ from Corollary 3.3.

Each allocation receives the optimal mass from Proposition 2.4. Hence the rescaled coefficient plotted in the left panels is simply

$$\frac{G\sqrt{1+\theta_{\mathrm{cap}}}}{D}s_t = \frac{\omega_t}{\sqrt{C_T(\omega_{1:T})}}.$$

At an intermediate index t, retain this final-horizon amplitude and apply Theorem 2.3 to the prefix. Direct substitution gives the ordinate in the bound plots:

$$\frac{\Omega_t}{DG\sqrt{1+\theta_{\mathrm{cap}}}} = \frac{\sqrt{C_T(\omega_{1:T})}}{2\sum_{j=1}^{t}\omega_j} + \frac{1}{2\sqrt{C_T(\omega_{1:T})}}\left[\omega_t + \sum_{k=1}^{t-1}\frac{\omega_k^2}{\sum_{j=k+1}^{t}\omega_j}\right]. \tag{C.12}$$

Each curve therefore uses one allocation and amplitude chosen for its final horizon T.

For comparison, suppose instead that each displayed index t were chosen as the final horizon and both allocation and amplitude were optimized again. Theorem 3.2 and Proposition 2.4 give the gray optimal-bound envelope directly:

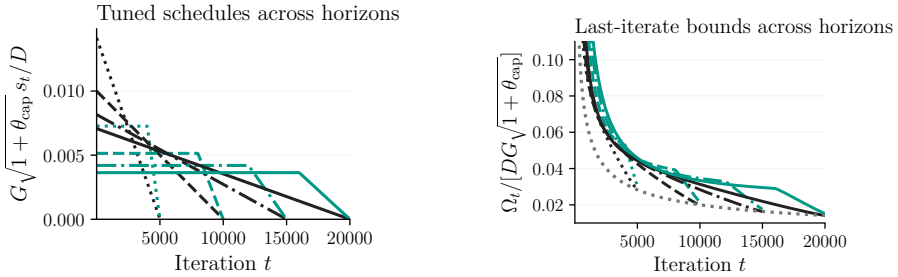
$$\frac{1}{DG\sqrt{1+\theta_{\mathrm{cap}}}} \min_{\substack{M>0,\ \omega_j>0 \\ \sum_{j=1}^{t}\omega_j=1}} \Omega_t(M,\omega_{1:t}) = \sqrt{C_t^\star}. \tag{C.13}$$

Each point on this envelope uses a different allocation and amplitude.

For the sensitivity comparison, hold $\omega_{1:T}$ fixed and vary its initial coefficient s_1, so $s_t = s_1\omega_t/\omega_1$. The final bound becomes

$$\begin{aligned}\frac{\Omega_T}{DG\sqrt{1+\theta_{\mathrm{cap}}}} &= \frac{D\omega_1}{2G\sqrt{1+\theta_{\mathrm{cap}}}\,s_1} + \frac{G\sqrt{1+\theta_{\mathrm{cap}}}\,s_1}{2D\omega_1}C_T(\omega_{1:T}),\\ s_1^\star &= \frac{D\omega_1}{G\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T})}}.\end{aligned} \tag{C.14}$$

Thus Fig. 8 compares the preferred initial coefficient of each allocation on one common scale. Multiplying any allocation's own optimum $s_1^\star$ by a factor $r > 0$ multiplies its minimum bound by $\left(r + r^{-1}\right)/2$, so the relative sensitivity is universal.



(a) (b)

Figure 7: Exact optimum (black) and WSD (teal) trajectories at horizons $T \in \{5{,}000, 10{,}000, 15{,}000, 20{,}000\}$, shown against iteration. WSD uses no warmup and a final 20% linear cooldown. Dotted, dashed, dash-dotted, and solid lines indicate $T = 5{,}000, 10{,}000, 15{,}000, 20{,}000$, respectively. **(a)** Normalized rotational coefficients. **(b)** Prefix bounds from (C.12), with the gray dotted optimal-bound envelope from (C.13). Each exact optimum meets the envelope at its final horizon.

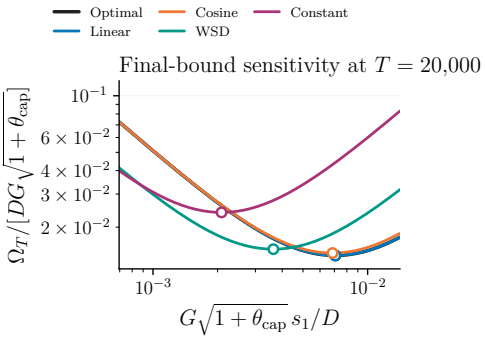


Figure 8: Final bound versus the common horizontal coordinate $G\sqrt{1+\theta_{\text{cap}}}\, s_1/D$ at $T = 20{,}000$. Open circles mark the minima from (C.14). The minima occur at different absolute coefficients, while relative sensitivity around each minimum is $\left(r + r^{-1}\right)/2$.

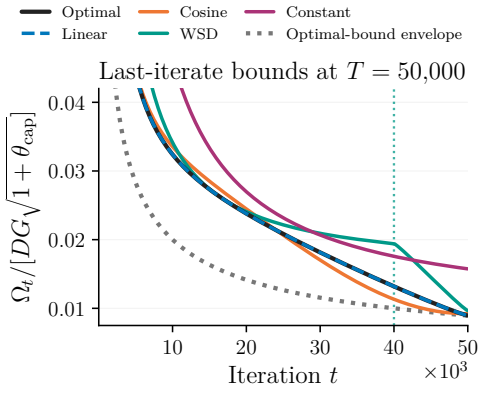


Figure 9: Last-iterate bounds at the longer horizon $T = 50{,}000$. The dotted gray curve is the optimal-bound envelope, and the vertical dotted line marks the start of WSD's final 20% linear cooldown at $t = 40{,}000$.

D A Simple Admissible Scale-Invariant Example

Full-domain Euclidean convexity. Suppose $L : \mathbb{R}^C \to \mathbb{R}$ is finite and convex, and $L(\kappa \boldsymbol{x}) = L(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^C$ and $\kappa > 0$. Fix $\boldsymbol{x}$ and $0 < a < 1$. Since $\boldsymbol{x} = (1-a)0 + a(\boldsymbol{x}/a)$, convexity and scale invariance give

$$L(\boldsymbol{x}) \le (1-a)L(0) + aL(\boldsymbol{x}/a) = (1-a)L(0) + aL(\boldsymbol{x}).$$

Subtracting $aL(\boldsymbol{x})$ and dividing by $1 - a > 0$ yields $L(\boldsymbol{x}) \le L(0)$. The same argument gives $L(-\boldsymbol{x}) \le L(0)$. Convexity at the midpoint $0 = (\boldsymbol{x} + (-\boldsymbol{x}))/2$ then gives

$$L(0) \le \frac{L(\boldsymbol{x}) + L(-\boldsymbol{x})}{2} \le L(0).$$

Both values in the average are at most $L(0)$, so equality forces $L(\boldsymbol{x}) = L(-\boldsymbol{x}) = L(0)$. Since $\boldsymbol{x}$ was arbitrary, L is constant.

We now give a nonconstant scale-invariant loss on the domain away from the origin. Its convexity is geodesic on a spherical cap, and the assumptions of Theorem 2.3 hold. The loss measures how far a direction points away from the cap center, so its gradient provides a direct way to check trajectory containment. Scaling any positive deterministic unit-mass allocation by the optimal mass of Proposition 2.4 will keep its iterates inside $\mathcal{C}$.

Proposition D.1 (A simple admissible scale-invariant model)**.** *Fix an integer $T \ge 1$, $\Lambda > 0$, and the cap $\mathcal{C}$ of* (2.5)*, centered at $\hat{\boldsymbol{x}}^\star \in \mathbb{S}$. For $\boldsymbol{x} \ne 0$, define*

$$L(\boldsymbol{x}) := \Lambda \left(1 - \left\langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \hat{\boldsymbol{x}}^\star \right\rangle\right). \tag{D.1}$$

Choose a deterministic initial point satisfying

$$\hat{\boldsymbol{x}}_1 \in \mathcal{C} \setminus \{\hat{\boldsymbol{x}}^\star\}, \qquad D = d_{\mathbb{S}}(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}^\star) \in (0, \theta_{\mathrm{cap}}), \tag{D.2}$$

and use the exact-gradient tangent oracle and envelope

$$\tilde{\boldsymbol{g}}_t := \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t), \qquad G := \Lambda \theta_{\mathrm{cap}}. \tag{D.3}$$

The loss is smooth and positively scale-invariant on $\mathbb{R}^C \setminus \{0\}$. When evaluated only at the unit directions $\hat{\boldsymbol{x}} \in \mathcal{C}$, it is strictly geodesically convex with unique minimizer $\hat{\boldsymbol{x}}^\star$, and its spherical gradient has norm at most G throughout the cap.

For every fixed deterministic positive unit-mass allocation $\omega_{1:T}$, the schedule

$$s_t = M^\star(\omega_{1:T})\omega_t, \qquad 1 \le t \le T, \tag{D.4}$$

with $M^\star$ from Proposition 2.4, keeps all iterates of (2.6)*, including $\hat{\boldsymbol{x}}_{T+1}$, inside $\mathcal{C}$. The loss, oracle, initialization, and schedule therefore satisfy Assumption B.1, so Theorem 2.3 applies, in particular for $\omega^\star_{1:T}$ from Corollary 3.3.*

Proof. Step 1: scale invariance and geodesic convexity. The map $\boldsymbol{x} \mapsto \boldsymbol{x}/\|\boldsymbol{x}\|$ is smooth away from zero. For $\kappa > 0$, $\|\kappa \boldsymbol{x}\| = \kappa \|\boldsymbol{x}\|$, so $\kappa\boldsymbol{x}/\|\kappa\boldsymbol{x}\| = \boldsymbol{x}/\|\boldsymbol{x}\|$ and (D.1) gives $L(\kappa\boldsymbol{x}) = L(\boldsymbol{x})$. On the sphere, $\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^\star\|^2 = 2 - 2\langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle$, since both vectors have unit norm. Hence

$$L(\hat{\boldsymbol{x}}) = \Lambda\big(1 - \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle\big) = \frac{\Lambda}{2} \|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^\star\|^2. \tag{D.5}$$

This is zero exactly at $\hat{\boldsymbol{x}}^\star$, the unique minimizer on the cap.

We still need to show convexity along spherical arcs. For this we compute the spherical gradient and then its change in a tangent direction: the latter gives the Hessian used in Boumal's convexity criterion. By (D.5), the affine function $\boldsymbol{x} \mapsto \Lambda(1 - \langle \boldsymbol{x}, \hat{\boldsymbol{x}}^\star \rangle)$ agrees with L at every unit vector. This function is a smooth ambient extension of the loss on the sphere, and its ordinary gradient is $-\Lambda\hat{\boldsymbol{x}}^\star$. For the sphere's induced Euclidean metric, Boumal (2023, Proposition 3.61) obtains the spherical gradient by projecting that ambient gradient onto the tangent space. At $\hat{\boldsymbol{x}}$, a vector's component parallel to the unit normal $\hat{\boldsymbol{x}}$ is its inner product with $\hat{\boldsymbol{x}}$ times $\hat{\boldsymbol{x}}$. Removing that component gives

$$\begin{aligned} \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}) &= -\Lambda\hat{\boldsymbol{x}}^\star - \langle -\Lambda\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{x}} \rangle \hat{\boldsymbol{x}} \\ &= -\Lambda\big(\hat{\boldsymbol{x}}^\star - \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle \hat{\boldsymbol{x}}\big). \end{aligned} \tag{D.6}$$

Next we compute the Hessian, which records the second derivative of the loss along a spherical geodesic. For this calculation, the formula $\boldsymbol{x} \mapsto -\Lambda(\hat{\boldsymbol{x}}^\star - \langle \boldsymbol{x}, \hat{\boldsymbol{x}}^\star \rangle \boldsymbol{x})$ extends the spherical gradient (D.6) to a smooth vector field around the sphere. Differentiating it at $\hat{\boldsymbol{x}}$ in a tangent direction $\boldsymbol{v} \in T_{\hat{\boldsymbol{x}}}\mathbb{S}$ gives, by the product rule,

$$\Lambda \langle \boldsymbol{v}, \hat{\boldsymbol{x}}^\star \rangle \hat{\boldsymbol{x}} + \Lambda \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle \boldsymbol{v}.$$

By Boumal (2023, Corollary 5.16), the Riemannian Hessian is the tangent projection of this derivative. Its first term is radial, and its second is tangent because $\langle \boldsymbol{v}, \hat{\boldsymbol{x}} \rangle = 0$. Therefore

$$\operatorname{Hess} L(\hat{\boldsymbol{x}})[\boldsymbol{v}] = \Lambda \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle \boldsymbol{v}. \tag{D.7}$$

The scalar $\langle \boldsymbol{v}, \operatorname{Hess} L(\hat{\boldsymbol{x}})[\boldsymbol{v}] \rangle$ is the second derivative of L at the start of the spherical geodesic with initial point $\hat{\boldsymbol{x}}$ and velocity $\boldsymbol{v}$. Using (D.7), it is

$$\begin{aligned} \langle \boldsymbol{v}, \operatorname{Hess} L(\hat{\boldsymbol{x}})[\boldsymbol{v}] \rangle &= \Lambda \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle \|\boldsymbol{v}\|^2 \\ &= \Lambda \cos\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star)\big) \|\boldsymbol{v}\|^2 . \end{aligned} \tag{D.8}$$

For $\hat{\boldsymbol{x}} \in \mathcal{C}$, we have $d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star) < \theta_{\mathrm{cap}} < \pi/4 < \pi/2$. Since $\Lambda > 0$, (D.8) is strictly positive for every $\boldsymbol{v} \neq 0$. The cap is open by (2.5) and geodesically convex by Corollary B.4. The loss is smooth, hence twice differentiable, on this domain. The positive-definite-Hessian criterion of Boumal (2023, Theorem 11.23, part 3) proves strict geodesic convexity of L on $\mathcal{C}$.

Step 2: tangent oracle and uniform bound. We now check the three oracle conditions of Theorem 2.3: tangency, conditional unbiasedness, and a common second-moment bound. From (D.3) and (D.6),

$$\langle \tilde{\boldsymbol{g}}_t, \hat{\boldsymbol{x}}_t \rangle = -\Lambda\big(\langle \hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{x}}_t \rangle - \langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star \rangle \|\hat{\boldsymbol{x}}_t\|^2\big) = 0,$$

because $\|\hat{\boldsymbol{x}}_t\| = 1$. Thus the oracle is tangent. The history $\mathcal{F}_t$ already contains $\hat{\boldsymbol{x}}_t$ before draw t. The oracle is exactly the gradient at that point, so it is already determined by this history; taking its conditional expectation does not change it:

$$\mathbb{E}[\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_t] = \tilde{\boldsymbol{g}}_t = \nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}}_t).$$

For $\hat{\boldsymbol{x}} \in \mathcal{C}$, (D.6) and the unit norms of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}^\star$ give

$$\begin{aligned} \|\nabla_{\mathbb{S}} L(\hat{\boldsymbol{x}})\|^2 &= \Lambda^2 \left(1 - \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star \rangle^2\right) \\ &= \Lambda^2 \sin^2\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star)\big) \\ &\leq \Lambda^2 d_{\mathbb{S}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^\star)^2 < \Lambda^2 \theta_{\mathrm{cap}}^2 = G^2. \end{aligned} \tag{D.9}$$

The last line uses $\sin u \leq u$ for $u \geq 0$ and the cap radius. This gives one bound G for every point in the cap, and Step 4 will show that every oracle evaluation occurs there.

Step 3: optimal-mass coefficients are sufficiently small. The oracle bound alone does not ensure that an update stays inside $\mathcal{C}$. We will show in Step 4 that the negative gradient points toward $\hat{\boldsymbol{x}}^\star$ and that $\Lambda s_t < 1$ prevents a step from passing it. We first prove that this coefficient bound holds for every allocation scaled by (D.4). Fix a deterministic positive unit-mass allocation. For $t < T$, the nonnegative summands of (2.11) and $0 < \sum_{j=t+1}^{T} \omega_j \leq 1$ give

$$C_T(\omega_{1:T}) \geq \frac{\omega_t^2}{\sum_{j=t+1}^{T} \omega_j} \geq \omega_t^2.$$

For the terminal entry, $0 < \omega_T \leq 1$ gives

$$C_T(\omega_{1:T}) \geq \omega_T \geq \omega_T^2.$$

For $T = 1$, the terminal case has $\omega_1 = C_1 = 1$. Taking positive square roots in both cases yields

$$\frac{\omega_t}{\sqrt{C_T(\omega_{1:T})}} \leq 1, \qquad 1 \leq t \leq T. \tag{D.10}$$

Substitute $s_t = M^\star(\omega_{1:T})\omega_t$ from (D.4), the mass formula of Proposition 2.4, and $G = \Lambda\theta_{\mathrm{cap}}$ from (D.3):

$$\begin{aligned}\Lambda s_t &= \frac{\Lambda D\omega_t}{G\sqrt{(1+\theta_{\mathrm{cap}})C_T(\omega_{1:T})}} \\ &= \frac{D}{\theta_{\mathrm{cap}}\sqrt{1+\theta_{\mathrm{cap}}}}\,\frac{\omega_t}{\sqrt{C_T(\omega_{1:T})}} \\ &\le \frac{D}{\theta_{\mathrm{cap}}\sqrt{1+\theta_{\mathrm{cap}}}} < 1.\end{aligned} \tag{D.11}$$

The first inequality uses (D.10); the last uses $D < \theta_{\mathrm{cap}}$ and $\sqrt{1+\theta_{\mathrm{cap}}} > 1$. Since D, G, and the allocation are fixed and positive, every s_t is positive and deterministic, independently of oracle draws.

Step 4: cap preservation. The remaining requirement of Theorem 2.3 is that its entire trajectory stay in $\mathcal{C}$, including $\hat{\boldsymbol{x}}_{T+1}$. We prove that one update from any point in the cap stays there, then apply induction. Suppose $\hat{\boldsymbol{x}}_t \in \mathcal{C}$. At $\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}^\star$, (D.6) gives $\tilde{\boldsymbol{g}}_t = 0$, so the update (2.6) leaves the point fixed. Otherwise $0 < d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star) < \theta_{\mathrm{cap}}$. We first identify the direction of the negative gradient. The sphere Log identity (B.11), with $\boldsymbol{y} = \hat{\boldsymbol{x}}^\star$ and $\cos\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big) = \langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star\rangle$, reads

$$\mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\hat{\boldsymbol{x}}^\star) = \frac{d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)}{\sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)}\big(\hat{\boldsymbol{x}}^\star - \langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star\rangle\,\hat{\boldsymbol{x}}_t\big).$$

The sine is positive because the distance lies in $(0, \theta_{\mathrm{cap}})$. Rearranging this display and using the gradient formula (D.6) gives both equalities below:

$$\begin{aligned}-\tilde{\boldsymbol{g}}_t &= \Lambda\big(\hat{\boldsymbol{x}}^\star - \langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star\rangle\,\hat{\boldsymbol{x}}_t\big) \\ &= \Lambda\frac{\sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)}{d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)}\,\mathrm{Log}_{\hat{\boldsymbol{x}}_t}(\hat{\boldsymbol{x}}^\star).\end{aligned} \tag{D.12}$$

The Log vector points along the shortest arc from $\hat{\boldsymbol{x}}_t$ to $\hat{\boldsymbol{x}}^\star$. The factor multiplying it in (D.12) is positive, so the update starts in that direction. We now check how far it turns: a turn larger than the remaining distance to $\hat{\boldsymbol{x}}^\star$ could pass the center.

We can read the turning angle directly from the normalized update (2.6). The equality in (D.9) gives $\|\tilde{\boldsymbol{g}}_t\|^2 = \Lambda^2 \sin^2\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)$. Here $\Lambda > 0$ and the distance lies in $(0, \theta_{\mathrm{cap}})$, so its sine is positive and $\|\tilde{\boldsymbol{g}}_t\| = \Lambda \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big) > 0$. Step 2 showed that $\tilde{\boldsymbol{g}}_t$ is perpendicular to $\hat{\boldsymbol{x}}_t$. Thus $\hat{\boldsymbol{x}}_t$ and $-\tilde{\boldsymbol{g}}_t/\|\tilde{\boldsymbol{g}}_t\|$ are perpendicular unit directions. Writing the update in those directions gives

$$\begin{aligned}\hat{\boldsymbol{x}}_{t+1} = &\frac{1}{\sqrt{1 + s_t^2\,\|\tilde{\boldsymbol{g}}_t\|^2}}\,\hat{\boldsymbol{x}}_t \\ &+ \frac{s_t\,\|\tilde{\boldsymbol{g}}_t\|}{\sqrt{1 + s_t^2\,\|\tilde{\boldsymbol{g}}_t\|^2}}\left(-\frac{\tilde{\boldsymbol{g}}_t}{\|\tilde{\boldsymbol{g}}_t\|}\right).\end{aligned}$$

The two coefficients are positive and their squares sum to one, so they are the cosine and sine of the turn from $\hat{\boldsymbol{x}}_t$ toward the negative gradient. The tangent of that angle is the sine component divided by the cosine component; their common denominator cancels, leaving $\tan\Delta\phi_t = s_t\,\|\tilde{\boldsymbol{g}}_t\|$. Hence $0 < \Delta\phi_t < \pi/2$ and the turn is the arctangent of this ratio, as in Lemma B.8. Substituting the norm above gives

$$\Delta\phi_t = \arctan\big(\Lambda s_t \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)\big) > 0.$$

First, apply $\arctan u \le u$ to the positive argument in the preceding display. This bounds the turn by $\Delta\phi_t \le \Lambda s_t \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)$.

Next, (D.11) gives $\Lambda s_t < 1$. The sine factor is positive, so multiplying this inequality by it gives $\Lambda s_t \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big) < \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big)$.

Finally, $\sin u \le u$ for nonnegative u bounds this sine by the distance $d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)$. Collecting the three bounds,

$$\begin{aligned}0 < \Delta\phi_t &\le \Lambda s_t \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big) \\ &< \sin\big(d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)\big) \le d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star).\end{aligned} \tag{D.13}$$

The current distance $d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star)$ is the length of the shortest arc to the center. The update travels $\Delta\phi_t$ along that arc, and (D.13) says it travels less than the arc's remaining length. Therefore $\hat{\boldsymbol{x}}_{t+1}$ lies between $\hat{\boldsymbol{x}}_t$ and $\hat{\boldsymbol{x}}^\star$; its distance to the center is the old distance minus the length traveled:

$$\begin{aligned} d_{\mathbb{S}}(\hat{\boldsymbol{x}}_{t+1}, \hat{\boldsymbol{x}}^\star) &= d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star) - \Delta\phi_t \\ &< d_{\mathbb{S}}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}^\star) < \theta_{\mathrm{cap}}. \end{aligned} \tag{D.14}$$

Together with the fixed-point case, this shows that a point in $\mathcal{C}$ remains there after one update. Induction from $\hat{\boldsymbol{x}}_1 \in \mathcal{C}$ through $t = T$ includes $\hat{\boldsymbol{x}}_{T+1}$. Consequently (D.9) holds at every oracle evaluation, and $\mathbb{E}\,\|\tilde{\boldsymbol{g}}_t\|^2 \le G^2$. The established loss, oracle, coefficient, and trajectory properties verify Assumption B.1. They hold for every fixed positive unit-mass allocation, including the unique bound-optimal $\omega^\star_{1:T}$ of Corollary 3.3. □

E Hyperball Geometry and Experimental Details

E.1 Geometric Correspondence

Vectorization preserves both the Frobenius norm and inner product. Set $\hat{\boldsymbol{x}}_t = \mathrm{vec}(\mathbf{X}_t)/R$ and $\boldsymbol{v}_t = \mathrm{vec}(\mathbf{V}_t)$. Hyperball keeps $\|\mathbf{X}_t\|_F = R$, so $\|\hat{\boldsymbol{x}}_t\| = 1$; normalization of the base update gives $\|\boldsymbol{v}_t\| = 1$. Dividing (4.1) by R therefore yields

$$\hat{\boldsymbol{x}}_{t+1} = \frac{\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t}{\|\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t\|}.$$

The denominator changes the length of the trial vector, but not its direction. Thus $\Delta\phi_t$ is the angle between $\hat{\boldsymbol{x}}_t$ and $\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t$. Write $c_t = \langle \hat{\boldsymbol{x}}_t, \boldsymbol{v}_t \rangle$. Their inner product and the squared length of the trial vector are

$$\langle \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t \rangle = 1 - \eta_t c_t, \qquad \|\hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{v}_t\|^2 = 1 - 2\eta_t c_t + \eta_t^2.$$

For a nonzero trial vector, the cosine formula gives the realized angle:

$$\Delta\phi_t = \arccos\left(\frac{1 - \eta_t c_t}{\sqrt{1 - 2\eta_t c_t + \eta_t^2}} \right). \tag{E.1}$$

For a tangent direction, $c_t = 0$. The trial vector then has one unit component along $\hat{\boldsymbol{x}}_t$ and an orthogonal component of length η_t. These are the adjacent and opposite sides of a right triangle whose angle is $\Delta\phi_t$, so $\tan \Delta\phi_t = \eta_t/1 = \eta_t$. Since $\eta_t \ge 0$, the angle lies in $[0, \pi/2)$, where tangent has a unique inverse. Therefore $\Delta\phi_t = \arctan \eta_t$, as claimed in (4.2).

Tangency also makes the displayed update identical to (2.6) when its coefficient is $s_t^{\mathrm{HB}} = \eta_t$ and its oracle is $\tilde{\boldsymbol{g}}_t = \boldsymbol{v}_t$. Because $\|\boldsymbol{v}_t\| = 1$, this oracle satisfies the theorem's second-moment bound with $G = 1$.

E.2 Training Protocol

We use the released pretraining codebase of Semenov et al. (2025) for a 124M-parameter Llama model on FineWeb, with sequence length 512, batch size 256, and 20,000 updates. Their 124M, large-batch benchmark motivates $(\beta_1, \beta_2) = (0.8, 0.999)$ for AdamH and for the AdamW auxiliary parameter groups. MuonH uses momentum 0.95, Nesterov momentum, and five Newton–Schulz steps. Weight decay is zero in all our runs.

We compare linear, cosine, warmup-stable-decay, and constant schedules on the constrained attention and MLP matrices. All runs use 2,000 warmup updates; WSD cools over the final 20% of the remaining updates. Auxiliary parameters keep an AdamW cosine schedule with peak rate 10^{-3}. For each optimizer and schedule, we select the peak matrix rate by final full-validation loss. The additional Adam and Muon runs use the same schedule comparison without Hyperball.

E.3 Additional Pretraining Comparisons

Figs. 10 and 12 compare measured turning angles with the tangent-update prediction for MuonH and AdamH. The remaining figures show schedule-amplitude sensitivity under Hyperball, followed by loss trajectories and peak-rate sensitivity for Adam and Muon without it.

Schedule	AdamH	MuonH	Adam	Muon
Linear	0.01414	0.01000	0.00100	0.01000
Cosine	0.01414	0.01000	0.00141	0.01000
WSD	0.00707	0.00500	0.00100	0.01000
Constant	0.00354	0.00177	0.00141	0.00707

Table 1: Best tested peak matrix learning rates, selected by final full-validation loss.

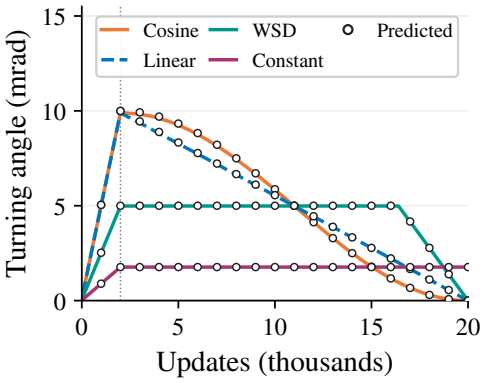


Figure 10: MuonH turning angles. Open circles show the predicted angle $\arctan s_t^{\mathrm{HB}}$ under tangency (4.2).

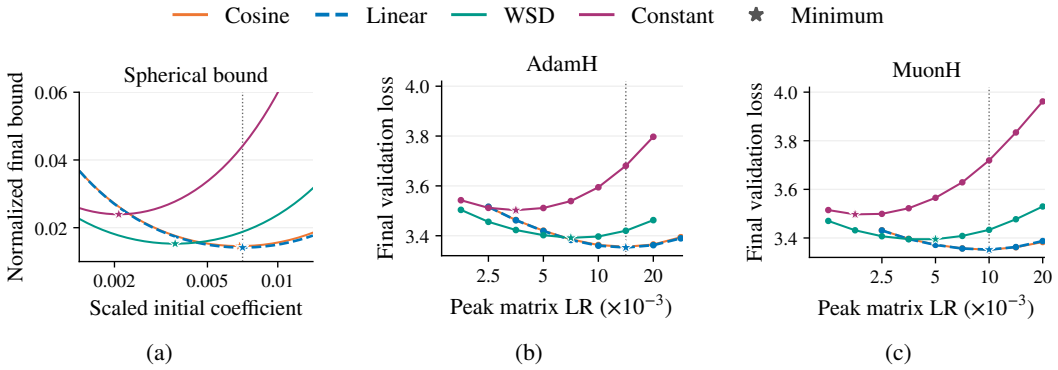


Figure 11: Sensitivity to schedule amplitude. **(a)** Spherical bound at $T = 20000$ against $G\sqrt{1+\theta_{\mathrm{cap}}}\, s_1/D$. **(b,c)** Final full-validation loss against peak matrix learning rate for AdamH and MuonH. Dotted lines mark the linear bound optimum in (a) and the best-tested linear rates in (b,c); stars mark each schedule's minimum.

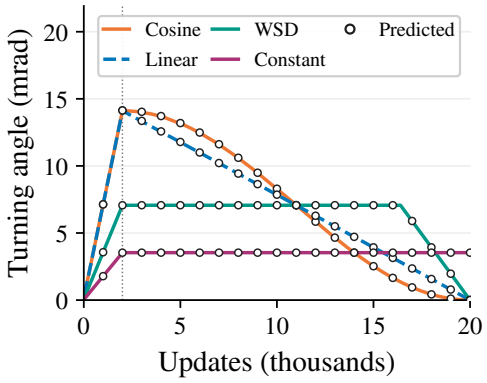


Figure 12: AdamH turning angles and the prediction $\arctan \eta_t$ under tangency.

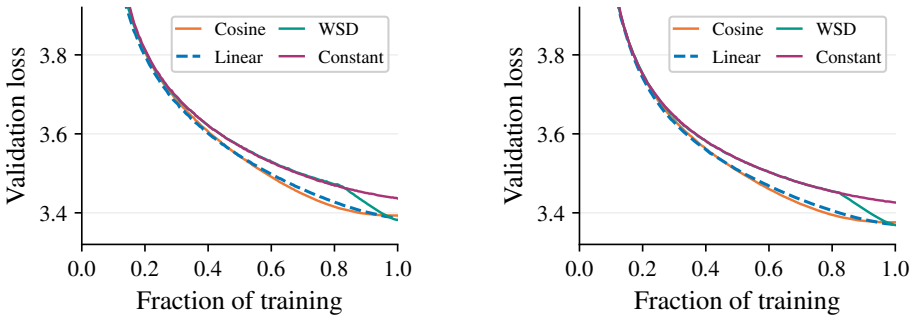


Figure 13: Periodic validation trajectories of the selected **(a)** Adam and **(b)** Muon runs without Hyperball or weight decay. Each schedule is selected by its final full-validation loss, as in Fig. 1.

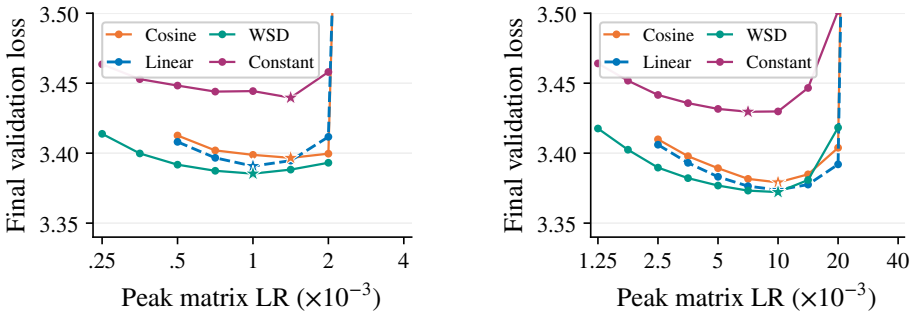


Figure 14: Full-validation peak-rate sensitivity for **(a)** Adam and **(b)** Muon without Hyperball or weight decay.

E.4 Evaluation on Measured Rotational Profiles

We use the same loss-selected Adam and Muon runs as in Fig. 13. Each configuration uses one training seed (0) and data seed 1337. Turning angles are recorded every 50 updates as medians across 60 attention and MLP matrices. We linearly interpolate these angles to each update, retain steps 2,001–20,000, and define

$$q_j = \tan(\Delta\phi_{2000+j}), \qquad \omega_j = \frac{q_j}{\sum_{i=1}^{18000} q_i}, \qquad 1 \le j \le 18000.$$

For a unit tangent direction, q_j is the coefficient producing the measured turn. For the theorem's general oracle, however, $\tan \Delta\phi_t = s_t \|\tilde{\boldsymbol{g}}_t\|$. The median-angle representation therefore does not identify the theorem's stochastic-gradient coefficients or establish its oracle assumptions.

For each fixed shape we use the optimal mass from (2.13) with common $D, G, \theta_{\mathrm{cap}}$. In dimensionless units the coefficients are $\omega_j / \sqrt{C_{18000}(\omega)}$, and the final bound divided by $DG\sqrt{1+\theta_{\mathrm{cap}}}$ is $\sqrt{C_{18000}(\omega)}$. We evaluate prefixes with this same final-horizon amplitude.

Schedule	Adam	Muon
Linear	0.01505	0.01535
Cosine	0.01557	0.01593
WSD	0.01525	0.01526
Constant	0.02130	0.02123

Table 2: Amplitude-optimized normalized endpoints for measured post-warmup rotational shapes.

WSD has a lower value than cosine for both optimizers. For Adam, linear retains a lower diagnostic value than WSD, although the plotted validation loss favors WSD. The comparison is retrospective: changing an unconstrained optimizer's learning rate can change its rotational shape, so optimizing amplitude here does not prescribe a retuned learning rate. Geodesic convexity, cap containment, and conditional unbiasedness are not established for these runs; the functional is not a certified bound on their validation losses. The single-seed results support the observed ordering without estimating its variability across seeds. Periodic validation uses 64 batches, whereas run selection uses final full-validation loss.