TAMING SOFTMAX AT THE LIMITS OF MUON: A UNIFIED APPROACH THROUGH HILBERT GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon controls weight updates through the RMS movement of a linear layer's output. This implicitly treats the linear output as a proxy for the full module, including the nonlinearity that follows. This proxy can be misleading for softmax, where the RMS shift in activations does not faithfully measure the change in the output distribution. We propose Hilbert distance as a natural metric for bounding changes in these output distributions. Hilbert distance bounds changes in cross-entropy loss and limits probability redistribution. For softmax, it takes the form of the variation seminorm of the preactivation perturbation, allowing probability constraints to be expressed as parameter constraints. Formulating the steepest descent problem with a Hilbert distance constraint brings two recently proposed methods under a unified mathematical framework. For the LM head, solving a row-wise approximation to the resulting diameter-constrained problem leads us to RowNorm (Pethick et al., 2025). For scaled dot-product attention, the same principle yields the first-order query–key steepest descent problem underlying Compositional Muon (Keigwin et al., 2026). These connections motivate using Hilbert distance as a guiding principle for optimizer design in softmax modules.

1 INTRODUCTION

An important consideration for stable training is the extent to which a single update changes the network's intermediate representations. Each intermediate representation is both the output of one layer and the input to the next. This dependence is especially important in deep sequential models, where changing one representation alters the inputs processed by every layer that follows. Without proper control, these representation shifts can accumulate across layers and destabilize training. Muon formalizes this idea for linear layers by controlling the worst-case RMS movement of the layer's output over inputs of bounded size (Jordan et al., 2024; Bernstein, 2025a). This requirement leads to a suitably scaled spectral norm constraint on the weight update.

This spectral constraint is not arbitrary: it follows from choosing RMS to measure the functional change produced by the update. More broadly, it illustrates how the geometry of an optimizer depends on the notion of change assigned to a module. The Modula project and Scion generalize this perspective by advocating update geometries that reflect the function of each module (Bernstein, 2025b; Pethick et al., 2025).

Muon's standard update geometry, however, is derived for the linear map and does not explicitly adapt to architectural operations such as normalization and nonlinear activations. As a result, Muon does not yet provide a single optimization rule for every parameter group in the model. In standard Muon training recipes, the LM head, embeddings, and scalar or vector parameters are typically optimized with AdamW (Jordan et al., 2024). Large-scale Muon training can also require additional stabilization. For example, Kimi K2 introduced MuonClip to address excessive attention-score growth by rescaling query and key weights after an update (Team et al., 2026).

We focus on how optimizer design changes when a linear map is followed by softmax. The linear layer's output is an intermediate state, and its RMS displacement does not faithfully measure the change in the probability distribution produced by softmax. We instead directly constrain the change in the output probability distribution. We propose using Hilbert's projective metric, or Hilbert distance, for this purpose.

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.

Our contributions are threefold:

- We formulate the steepest descent problem using Hilbert distance to constrain changes in softmax output distributions. This single formulation brings two seemingly independent optimization methods under a unified mathematical framework.
- For the LM head, we show that this probability constraint becomes a row diameter constraint. Its exact dual scales quadratically with vocabulary size. A tighter common-ball constraint instead leads to RowNorm in closed form (Pethick et al., 2025).
- For standard attention, we show that this probability constraint produces a spectral constraint on the first-order change in the QK product, recovering the steepest descent problem posed in Compositional Muon (Keigwin et al., 2026).

Beyond unifying these recent methods, we provide a guiding principle for optimizer design for any neural network module that outputs a probability distribution.

We first expose the mismatch between RMS and softmax (Section 2). Next, we introduce Hilbert geometry and establish its suitability for measuring changes in softmax output distributions (Section 3). We subsequently formulate the corresponding steepest descent problems for the LM head and attention. For the LM head, a tighter approximation leads to RowNorm in closed form. For attention, our formulation recovers the QK steepest descent problem studied by Compositional Muon (Sections 4 and 5).

2 Preliminaries on steepest descent

Steepest descent chooses an update that gives the greatest first-order decrease in the loss while remaining small under a chosen measure of size. For a minibatch gradient $G = \nabla_W \mathcal{L}$, an update ΔW changes the loss to first order by $\langle G, \Delta W \rangle_F$. Let $\mathcal{D}_W(\Delta W)$ measure the size of this update at the current weights W. We consider

$$\min_{\Delta W} \langle G, \Delta W \rangle_F \qquad \text{subject to} \qquad \mathcal{D}_W(\Delta W) \leq \rho, \tag{1}$$

where $\rho > 0$ sets the permitted update size. We leave $\mathcal{D}_W$ unspecified for now, allowing its choice to reflect the module being optimized.

The importance of the underlying choice of metric cannot be overstated. Choosing the Frobenius norm measures the update directly in parameter space and gives the SGD direction (Robbins & Monro, 1951). Muon takes a functional perspective instead: for a linear layer, it controls the worst-case change in the output produced by the update (Bernstein, 2025a). Measuring this change with RMS leads to a spectral norm constraint on the weight update, whose steepest descent solution gives the Muon direction. The Modula project generalizes this perspective by advocating update geometries that reflect the function of each module (Bernstein, 2025b). Around this principle, Scion develops a systematic optimization framework, formulating updates as Linear Minimization Oracles under $\|\cdot\|_{p\to q}$ operator norm constraints (Pethick et al., 2025). Here, the norms $\|\cdot\|_p$ and $\|\cdot\|_q$ describe the input and output geometries of the module. The shared ethos of these works is to build optimizers whose geometry reflects the function of each module.

This functional perspective becomes more delicate when a linear layer is followed by a nonlinearity. The linear output is then only an intermediate state in the computation of the full module. Controlling the RMS movement of this intermediate state does not account for the geometry introduced by the subsequent nonlinearity. For a softmax module, the linear output is the preactivation vector, which softmax maps to a probability distribution on the simplex. We therefore ask whether RMS displacement of the preactivation faithfully measures changes in the output distribution.

2.1 Geometric mismatch between Muon and softmax

We argue that Muon is inadequate for modules involving softmax. Muon's spectral geometry can allocate updates poorly for such modules. The distinction lies between the nature of RMS and softmax: RMS measures how much the preactivation vector moves on average, while softmax measures how much one coordinate *gains* relative to *others*.

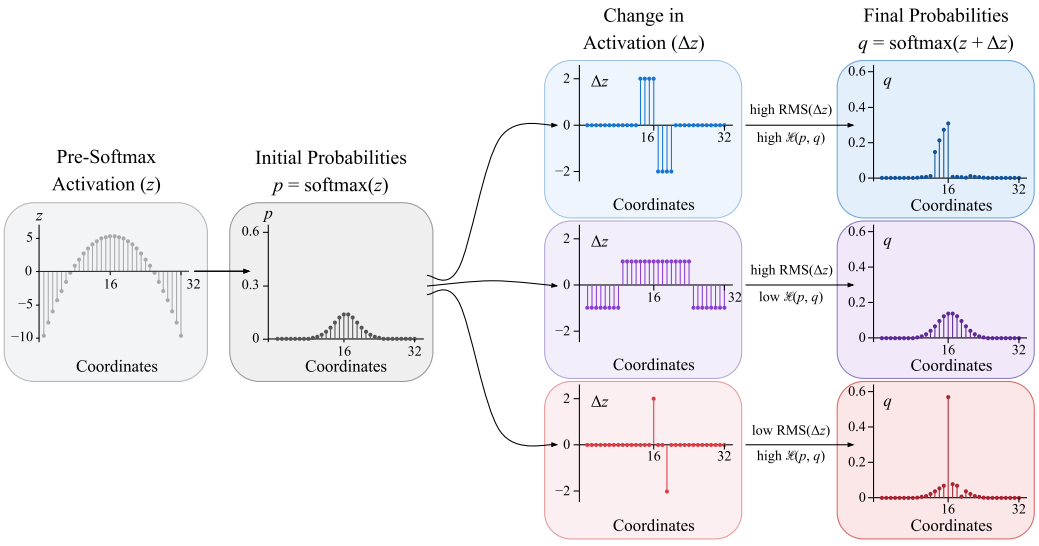


Figure 1: **A common RMS bound permits very different changes in softmax probabilities.** All cases share the same centered $z \in \mathbb{R}^{32}$ and initial distribution $p = \text{softmax}(z)$. The centered perturbations produce $q = \text{softmax}(z + \Delta z)$ and all satisfy $\text{RMS}(\Delta z) \leq 1$. The pairs $(\text{RMS}(\Delta z), \mathcal{H}(p, q))$ are $(1, 4)$ for blue, $(1, 2)$ for purple, and $(1/2, 4)$ for red, where $\mathcal{H}$ denotes Hilbert distance. Blue substantially reshapes the distribution, while purple barely changes it. Red matches blue's Hilbert distance at half its RMS.

A simple example. Consider an update where all the rows are equal, $\Delta W = \mathbf{1}b^\top$. In this case, $\Delta z = \Delta W x = (b^\top x)\mathbf{1}$, a scaled ones vector. Such a perturbation to the preactivation vector does not alter the probability distribution, despite incurring an arbitrarily large RMS norm. This particular case is not problematic in practice because softmax modules receive row-centered gradients, causing Muon to consider only centered directions; see Section 6. It nevertheless illustrates a broader geometric mismatch between RMS bounds and softmax that can produce practical pathologies.

Figure 1 shows that this discrepancy persists under centered perturbations. Unlike the preceding example, all three cases share the same initial distribution and use centered perturbations satisfying a common RMS bound. The blue and purple updates have equal RMS, yet the blue perturbation substantially reshapes the distribution while the purple perturbation barely changes it. The red perturbation has half their RMS but still causes a substantial probability redistribution.

This mismatch compels direct control of the output distribution. Designing an effective update rule for softmax modules requires choosing a suitable metric on the output distribution and formulating the induced steepest descent problem. This raises a fundamental question:

What is the appropriate metric for perturbations of the softmax probabilities?

3 Hilbert Geometry of Softmax

We turn to Hilbert's projective metric, or simply Hilbert distance (Birkhoff, 1957; De La Harpe, 1993). For strictly positive vectors $x, y \in \mathbb{R}^n_{>0}$, it is defined by $\mathcal{H}(x, y) = \log \frac{\max_i x_i/y_i}{\min_j x_j/y_j}$. We apply this metric to strictly positive probability distributions on the simplex $\mathcal{P} = \{x \in \mathbb{R}^n_{>0} : \sum_i x_i = 1\}$ (Nielsen & Sun, 2018; 2023). For $p, q \in \mathcal{P}$, Hilbert distance can equivalently be written as

$$\mathcal{H}(p, q) = \max_{i,j} \left| \log \frac{q_i/q_j}{p_i/p_j} \right|. \tag{2}$$

Thus, Hilbert distance measures the largest change in pairwise odds.

Hilbert's projective metric has several useful geometric properties on positive cones and probability simplices (Lemmens & Nussbaum, 2012; Papadopoulos & Troyanov, 2014). Bounding the

Hilbert distance directly controls the change in cross-entropy loss and limits probability redistribution in total variation. For softmax, the distance also admits a closed-form expression in terms of the preactivation update: it coincides with the range between the largest and smallest coordinates (the variation seminorm). Consequently, uniform shifts leave the distance invariant, mirroring the translation invariance of the softmax itself. Motivated by these geometric properties, we adopt the Hilbert distance to quantify changes in output distributions and formulate the resulting steepest descent problem.

Hilbert distance bounds pairwise odds ratios. Throughout, let $z \in \mathbb{R}^n$ denote the preactivation supplied to softmax, and let $p = \mathrm{softmax}(z)$ be the resulting output distribution. A perturbation Δz produces the updated distribution $q = \mathrm{softmax}(z + \Delta z)$. By Equation 2, bounding Hilbert distance uniformly bounds the odds ratio for every pair of outcomes:

$$\mathcal{H}(p, q) \le \varepsilon \quad \Longleftrightarrow \quad e^{-\varepsilon} \le \frac{q_i/q_j}{p_i/p_j} \le e^{\varepsilon} \qquad \forall i, j.$$

In particular, for softmax, Hilbert distance has the closed form

$$\mathcal{H}\big(\mathrm{softmax}(z), \mathrm{softmax}(z + \Delta z)\big) = \max_{i,j} |\Delta z_i - \Delta z_j| = \max_i \Delta z_i - \min_i \Delta z_i = \|\Delta z\|_{\mathrm{var}}. \quad (3)$$

The variation seminorm $\|\Delta z\|_{\mathrm{var}}$, also called the oscillation $\mathrm{osc}(\Delta z)$, is the difference between the largest and smallest entries of Δz. It vanishes precisely when Δz is constant, matching the translation invariance of softmax.

Hilbert distance bounds probability redistribution. Bounding pairwise odds ratios also limits the redistribution of probability mass. We measure this redistribution using total variation,

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_i |p_i - q_i| = \max_{A \subseteq \{1, \dots, n\}} \left| \sum_{i \in A} (p_i - q_i) \right|.$$

Thus, $\mathrm{TV}(p, q) \le \delta$ guarantees that no set of outcomes can gain or lose more than δ probability mass. Hilbert distance gives a sharp bound on total variation (Cohen & Fausti, 2024, Theorem 5.1),

$$\mathrm{TV}(p, q) \le \tanh\left(\frac{\mathcal{H}(p, q)}{4}\right).$$

The bound is sharp even when the preactivation perturbation is fixed. For every $\Delta z \in \mathbb{R}^n$,

$$\sup_{z \in \mathbb{R}^n} \mathrm{TV}\big(\mathrm{softmax}(z), \mathrm{softmax}(z + \Delta z)\big) = \tanh\left(\frac{\|\Delta z\|_{\mathrm{var}}}{4}\right).$$

The variation seminorm therefore determines the largest probability redistribution that a given preactivation update can produce across all starting distributions. Appendix A.2 proves this sharpness statement.

Hilbert distance bounds change in cross-entropy (CE) loss. Let p and q be the prediction distributions before and after an update. For a target outcome i, the CE loss is $\ell_i(p) = -\log p_i$, so its change is $\Delta\ell_i = \ell_i(q) - \ell_i(p) = \log \frac{p_i}{q_i}$. Hilbert distance is the range of these changes, $\mathcal{H}(p, q) = \max_i \Delta\ell_i - \min_i \Delta\ell_i$. It therefore measures how differently the update affects the loss across possible targets. Moreover, $\sum_i p_i = \sum_i q_i = 1$ implies $\min_i \frac{p_i}{q_i} \le 1 \le \max_i \frac{p_i}{q_i}$. Taking logarithms gives $\min_i \Delta\ell_i \le 0 \le \max_i \Delta\ell_i$. Consequently, $\mathcal{H}(p, q) \le \varepsilon$ guarantees that the change in CE loss is at most ε,

$$|\Delta\ell_i| \le \max_j \Delta\ell_j - \min_j \Delta\ell_j \le \mathcal{H}(p, q) \quad \forall i.$$

Steepest descent problem. Together, these results show that Hilbert distance controls pairwise odds, probability redistribution, and changes in cross-entropy loss. Having established these properties, we now formulate the steepest descent problem using Hilbert distance to measure the size of an update. We return to Equation 1 and express the update constraint in terms of this distance. Let W denote the module's parameters and let $p_W(x)$ be the output distribution produced at input x. An update ΔW changes this distribution from $p_W(x)$ to $p_{W+\Delta W}(x)$. Requiring these distributions to remain within Hilbert distance ε for every input in a prescribed set $\mathcal{X}$ gives

$$\min_{\Delta W} \langle \nabla_W \mathcal{L}, \Delta W \rangle_F \qquad \text{subject to} \qquad \sup_{x \in \mathcal{X}} \mathcal{H}\big(p_W(x), p_{W+\Delta W}(x)\big) \le \varepsilon. \quad (4)$$

4 THE LM HEAD

For the LM head, bounding Hilbert distance over the hidden state space is equivalent to bounding the diameter of the update rows. The resulting steepest descent problem has an exact dual whose size grows quadratically with the vocabulary. A tighter common-ball approximation avoids this coupling and leads to RowNorm in closed form.

We first specialize Equation 4 to the LM head. For a hidden state $h \in \mathbb{R}^d$, the LM head produces the logit vector $z = Wh \in \mathbb{R}^V$ and the token distribution $p = \mathrm{softmax}(z)$. We consider an untied head without bias, trained with CE and no other loss term depending on W. Throughout this section, we compare the head before and after an update at the same hidden state.

Hilbert distance induces a row diameter constraint. Let ΔW be a weight update with rows $\Delta w_i^\top$. The update changes the logit for token i by $\Delta w_i^\top h$. For two tokens i and j, their logit difference therefore changes by $(\Delta w_i - \Delta w_j)^\top h$. Equation 3 gives

$$\mathcal{H}\big(p_W(h), p_{W+\Delta W}(h)\big) = \|\Delta W h\|_{\mathrm{var}} = \max_{i,j} \left|(\Delta w_i - \Delta w_j)^\top h\right|.$$

For a fixed pair of rows, the largest value of this inner product over $\|h\|_2 \le R$ is $R\,\|\Delta w_i - \Delta w_j\|_2$. Taking the largest value over all row pairs gives

$$\sup_{\|h\|_2 \le R} \mathcal{H}\big(p_W(h), p_{W+\Delta W}(h)\big) = R \max_{i,j} \|\Delta w_i - \Delta w_j\|_2 = RD(\Delta W). \tag{5}$$

The maximum pairwise distance is precisely the diameter $D(\Delta W)$ of the row cloud $\{\Delta w_1, \ldots, \Delta w_V\}$ (Rudin, 1976, Definition 3.9). For this input ball, the general constraint therefore specializes to $\mathcal{D}_W(\Delta W) = RD(\Delta W)$. Choosing $R = \sqrt{d}$ gives the RMS unit ball. Thus, requiring a Hilbert distance of at most ε throughout this ball is equivalent to imposing $D(\Delta W) \le \rho_{\mathrm{head}}$, where $\rho_{\mathrm{head}} = \varepsilon/R$.

This single constraint on the update rows controls both the predicted distribution and the cross-entropy loss. By Section 3, no group of tokens can gain or lose more than $\tanh(\varepsilon/4)$ probability mass, and the CE loss changes by at most ε for any target token.

Steepest descent under the row diameter constraint. We seek the update with the greatest first-order loss decrease under the row diameter bound. Let $G = \nabla_W \mathcal{L}$ have rows $g_i^\top$. Using the row diameter constraint in Equation 5, we obtain

$$\min_{\Delta W} \langle G, \Delta W \rangle_F \qquad \text{subject to} \qquad D(\Delta W) \le \rho_{\mathrm{head}}. \tag{6}$$

Since the loss depends on W only through the softmax output, its gradient satisfies $\mathbf{1}^\top G = 0$. Adding the same vector to every update row produces a common shift of the logits and therefore leaves both the output distribution and the loss unchanged. Because pairwise row differences are also unchanged, the diameter constraint places no restriction on the common translation of the update rows.

Appendix A.3 gives an exact dual characterization of this problem. The dual is a vector-valued optimal transport problem (Ciosmak, 2021). In our setting, this transport problem introduces a d-dimensional flow variable for each of the $\binom{V}{2}$ token pairs. This quadratic growth with vocabulary size makes solving the flow problem computationally expensive. We proceed on a quest to find a cheaper approximation.

A common-ball approximation leads to RowNorm. We ask for a tighter constraint, requiring all update rows to lie in a common closed Euclidean ball $\mathcal{B}(\zeta, \rho_{\mathrm{head}}/2)$. With this restriction, Equation 6 becomes

$$\min_{\Delta W, \zeta} \langle G, \Delta W \rangle_F \qquad \text{subject to} \qquad \Delta w_i \in \mathcal{B}(\zeta, \rho_{\mathrm{head}}/2) \quad \forall i. \tag{7}$$

Since the gradient rows sum to zero, translating every update row by the same vector leaves the objective unchanged. Using this freedom, we invoke a change of variable $\widehat{\Delta W} = \Delta W - \mathbf{1}\zeta^\top$,

which expresses each row relative to the ball's center. Each row $\widehat{\Delta w}_i$ then lies in $\mathcal{B}(0, \rho_{\text{head}}/2)$. Equation 7 therefore has the same minimum as

$$\min_{\widehat{\Delta W}} \left\langle G, \widehat{\Delta W} \right\rangle_F \qquad \text{subject to} \qquad \widehat{\Delta w}_i \in \mathcal{B}(0, \rho_{\text{head}}/2) \quad \forall i. \tag{8}$$

Conversely, any $\widehat{\Delta W}$ satisfying this constraint gives a feasible common-ball update by choosing $\Delta W = \widehat{\Delta W}$ and $\zeta = 0$.

A natural curiosity would be the suboptimality gap introduced by the tighter ball constraint. Let $\Delta_\star$ and Δ_{ball} denote the greatest first-order decreases under the diameter and common-ball constraints, respectively. These are the negatives of the minimum objective values in Equations 6 and 7. For $G \neq 0$ and $\rho_{\text{head}} > 0$, let $k = \operatorname{rank}(G)$. Jung's theorem (Jung, 1901) bounds the radius needed to enclose a set of rows with a given diameter. Applying it to the two optimization problems bounds the relative suboptimality gap by

$$0 \leq \frac{\Delta_\star - \Delta_{\text{ball}}}{\Delta_\star} \leq 1 - \sqrt{\frac{k+1}{2k}}. \tag{9}$$

The bound is sharp. Appendix A.4 provides its proof and elaborates on the equivalence brought about by the change of variable.

To find an update with this guarantee, we solve Equation 8 one row at a time. Each row has its own constraint and contributes independently to the objective. For a feasible row $\widehat{\Delta w}_i$, the Cauchy–Schwarz inequality and the radius constraint give

$$\left\langle g_i, \widehat{\Delta w}_i \right\rangle \geq -\|g_i\|_2 \left\|\widehat{\Delta w}_i\right\|_2 \geq -\frac{\rho_{\text{head}}}{2} \|g_i\|_2 .$$

For $g_i \neq 0$, equality is attained by choosing $\widehat{\Delta w}_i$ opposite to g_i with norm $\rho_{\text{head}}/2$. For a zero gradient row, we choose $\widehat{\Delta w}_i = 0$. Let RN normalize each nonzero row to unit Euclidean norm and leave zero rows unchanged. Choosing $\Delta W = \widehat{\Delta W}$ and $\zeta = 0$, we obtain one minimizing update for the common-ball problem,

$$\Delta W_{\text{RN}} = -\frac{\rho_{\text{head}}}{2} \mathsf{RN}(G). \tag{10}$$

This update is feasible for the original diameter problem and attains the common-ball optimum. Its first-order decrease is Δ_{ball}, so Equation 9 applies directly to RowNorm.

A practical version subtracts the mean row from ΔW_{RN} to obtain projected RowNorm. This common translation preserves the objective, the row diameter, and the output probabilities. Our Hilbert distance argument therefore independently recovers the direction derived by Lau & Su (2026) from the symmetries of the LM head. Appendix A.5 gives an implementation with momentum and numerical stabilization.

5 The Attention Module

We extend our principle from the LM head to attention. Attention presents a more intricate setting: its softmax probabilities arise from bilinear scores that jointly depend on the query and key weights. Bounding the Hilbert distance between attention distributions leads us to the QK constraint underlying Compositional Muon.

Attention distributions. Attention uses softmax to distribute probability across the positions visible to each query. We study how this distribution changes when the query and key weights are updated, holding the input hidden states $h_1, \ldots, h_T \in \mathbb{R}^d$ fixed. For a single head with query and key dimension d_h, the weights $W_Q, W_K \in \mathbb{R}^{d \times d_h}$ determine the attention scores through

$$M = W_Q W_K^\top \in \mathbb{R}^{d \times d}, \qquad s_{ij} = \frac{h_i^\top M h_j}{\sqrt{d_h}}.$$

We consider scaled dot-product attention without rotary position transforms or query–key normalization. Under causal masking, let p_{ij} denote the probability assigned by query i to position $j \in \mathcal{J}_i = \{1, \ldots, i\}$. We measure Hilbert distance on this visible set, where all probabilities are positive.

Hilbert distance bounds changes in score differences. Let ΔM denote the change in the QK product, so that the updated product is $M + \Delta M$. At position i, write $p_i = (p_{ij})_{j \in \mathcal{J}_i}$ for the original attention distribution and p_i^+ for the distribution after this change. Holding the hidden states fixed, the score s_{ij} changes by

$$\Delta s_{ij} = \frac{h_i^\top (M + \Delta M) h_j - h_i^\top M h_j}{\sqrt{d_h}} = \frac{h_i^\top \Delta M h_j}{\sqrt{d_h}}.$$

Within row i, the score difference between positions j and k therefore changes by $h_i^\top \Delta M (h_j - h_k)/\sqrt{d_h}$. Applying Equation 3 within each attention row gives

$$\mathcal{H}(p_i, p_i^+) = \max_{j,k \in \mathcal{J}_i} |\Delta s_{ij} - \Delta s_{ik}| = \frac{1}{\sqrt{d_h}} \max_{j,k \in \mathcal{J}_i} \left| h_i^\top \Delta M (h_j - h_k) \right|. \tag{11}$$

Intuitively, the probability constraint, specific to the attention mechanism, asks how much an update can change a query's preference between two keys.

The induced spectral norm constraint. Assume that every hidden state satisfies $\|h_\ell\|_2 \le R$, with $R = \sqrt{d}$ corresponding to the RMS unit ball. The hidden state h_i has norm at most R, while the difference $h_j - h_k$ has norm at most $2R$. Consequently, $\left| h_i^\top \Delta M (h_j - h_k) \right| \le \|h_i\|_2 \|\Delta M\|_{\mathrm{op}} \|h_j - h_k\|_2 \le 2R^2 \|\Delta M\|_{\mathrm{op}}$.

For causal sequences with $T \ge 3$, the worst-case Hilbert distance over this input set is

$$\sup_{\substack{\|h_\ell\|_2 \le R \\ 1 \le \ell \le T}} \max_{1 \le i \le T} \mathcal{H}(p_i, p_i^+) = \frac{2R^2}{\sqrt{d_h}} \|\Delta M\|_{\mathrm{op}}. \tag{12}$$

Requiring a Hilbert distance of at most ε for every attention row throughout this input set is therefore equivalent to

$$\|\Delta M\|_{\mathrm{op}} \le \rho_{\mathrm{QK}}, \qquad \rho_{\mathrm{QK}} = \frac{\varepsilon \sqrt{d_h}}{2R^2}. \tag{13}$$

Why does the same Hilbert distance principle give a row diameter for the LM head and a spectral norm for attention? In the LM head, each comparison involves a single hidden state, whereas attention involves a query input and a difference between key inputs. Under the stated norm bounds, these two vectors can be aligned independently to attain the spectral norm of ΔM. The bound is attained when h_i is R times a unit leading left singular vector of ΔM, and $h_j - h_k$ is $2R$ times the corresponding unit right singular vector.

Steepest descent over query and key updates. Let ΔW_Q and ΔW_K denote updates to the query and key weights W_Q and W_K, respectively. Expanding the updated product gives

$$\Delta M = (W_Q + \Delta W_Q)(W_K + \Delta W_K)^\top - W_Q W_K^\top = \underbrace{\Delta W_Q W_K^\top + W_Q \Delta W_K^\top}_{\Delta M_{\mathrm{lin}}} + \Delta W_Q \Delta W_K^\top.$$

The first two terms give the first-order change $\Delta M_{\mathrm{lin}}(\Delta W_Q, \Delta W_K)$, while $\Delta W_Q \Delta W_K^\top$ is second order in the updates and is therefore omitted. Applying Equation 13 to this first-order change gives

$$\left\| \Delta W_Q W_K^\top + W_Q \Delta W_K^\top \right\|_{\mathrm{op}} \le \rho_{\mathrm{QK}}.$$

We assume that these weights affect the loss only through the attention probabilities. Write $G_Q = \nabla_{W_Q} \mathcal{L}$ and $G_K = \nabla_{W_K} \mathcal{L}$ for their gradients. Constructing our steepest descent problem under this Hilbert distance bound, we obtain

$$\begin{aligned} \min_{\Delta W_Q, \Delta W_K} \quad & \langle G_Q, \Delta W_Q \rangle_F + \langle G_K, \Delta W_K \rangle_F \\ \text{subject to} \quad & \left\| \Delta W_Q W_K^\top + W_Q \Delta W_K^\top \right\|_{\mathrm{op}} \le \rho_{\mathrm{QK}}. \end{aligned} \tag{14}$$

Equation 14 is precisely the first-order QK steepest descent problem posed by Keigwin et al. (2026, Section 1.2). Their derivation controls changes in the composed bilinear map through the spectral norm of its first-order perturbation. Our starting point is a bound on the Hilbert distance between

attention distributions. Translating this probability bound through softmax and the bilinear scores recovers the same constraint on the first-order QK product change. Our Hilbert distance derivation therefore provides a natural prelude to Compositional Muon, motivating its QK constraint through the stability of the attention distributions.

The LM head and attention affect optimization through both forward probability distributions and the gradients propagated to the hidden states. Our main analysis bounds changes in the former and derives the corresponding steepest descent problems. Holding the inputs and incoming gradients fixed, Appendix B bounds the resulting perturbations of the latter.

6 Discussion

Can Muon be adapted to softmax simply by removing the common shifts that softmax ignores? Because softmax is invariant to common shifts, one might try adapting Muon by centering the update rows before measuring their spectral norm. We show that the resulting steepest descent problem has the same minimum and that the Muon update is already row centered. Translation symmetry alone therefore does not alter Muon, motivating the stronger relative control provided by Hilbert distance.

Centering leaves Muon unchanged. Consider a linear module with n outputs and a parameter matrix whose influence on the loss is entirely through the softmax output distribution. Let $P = I_n - \mathbf{1}\mathbf{1}^\top/n$ denote the orthogonal projection that subtracts the mean row. Since gradient rows sum to zero, $PG = G$. Since P is an orthogonal projection, $\|P\Delta W\|_{\mathrm{op}} \le \|P\|_{\mathrm{op}} \|\Delta W\|_{\mathrm{op}} \le \|\Delta W\|_{\mathrm{op}}$. Thus, every update satisfying $\|\Delta W\|_{\mathrm{op}} \le \rho$ also satisfies $\|P\Delta W\|_{\mathrm{op}} \le \rho$. Conversely, take any update ΔW satisfying $\|P\Delta W\|_{\mathrm{op}} \le \rho$. The centered update $P\Delta W$ has spectral norm at most ρ. It also preserves the objective, since $\langle G, P\Delta W\rangle_F = \langle PG, \Delta W\rangle_F = \langle G, \Delta W\rangle_F$. The additional freedom therefore leaves the minimum unchanged,

$$\min_{\|P\Delta W\|_{\mathrm{op}} \le \rho} \langle G, \Delta W\rangle_F = \min_{\|\Delta W\|_{\mathrm{op}} \le \rho} \langle G, \Delta W\rangle_F .$$

The Muon update is already centered. Indeed, if $G = A\Sigma B^\top$ is the compact SVD of a nonzero gradient, then every left singular vector satisfies $\mathbf{1}^\top a_j = (\mathbf{1}^\top G b_j)/\sigma_j = 0$, giving

$$\Delta W_{\mathrm{Muon}} = -\rho AB^\top, \qquad P\Delta W_{\mathrm{Muon}} = -\rho(PA)B^\top = \Delta W_{\mathrm{Muon}}.$$

Accounting for translation symmetry alone therefore recovers the standard Muon solution. Our Hilbert distance approach goes further by bounding the largest change in the difference between any two preactivation coordinates. For a centered preactivation perturbation $\Delta z \in \mathbb{R}^n$, with $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z + \Delta z)$, we have $2\,\|\Delta z\|_{\mathrm{RMS}} \le \mathcal{H}(p, q) \le \sqrt{2n}\ \|\Delta z\|_{\mathrm{RMS}}$. Since the upper bound is sharp, a fixed bound on centered RMS permits Hilbert distance to grow with the output width. Appendix A.1 proves these bounds and discusses their sharpness.

Limitations. The main objective of this work is to provide a unifying mathematical foundation for softmax modules. Through Hilbert distance, we independently recover RowNorm for the LM head and the first-order QK constraint underlying Compositional Muon. These mathematical connections naturally raise the question of what this perspective offers in practice. We leave that question to future experimental work and make no empirical claim of improved training speed or final model quality. The work of Lau & Su (2026) on RowNorm and that of Keigwin et al. (2026) on Compositional Muon report improvements in language-model pretraining and include ablations of their optimizer choices. These experiments provide existing empirical evidence for the practical relevance of the methods connected by our framework.

A second limitation arises from the relative nature of Hilbert distance. Because it controls pairwise odds, it can be sensitive to large relative changes in low probability outcomes even when little probability mass moves. For example, suppose an update raises the probability of one outcome from 10^{-9} to 10^{-6} while leaving the rest of the distribution nearly unchanged. The probability of this outcome increases by a factor of 1000, producing a large Hilbert distance despite the small movement of probability mass, $\mathrm{TV}(p, q) = 9.99 \times 10^{-7}, \mathcal{H}(p, q) \approx 6.91$. A small Hilbert distance constraint can therefore exclude updates that move very little probability mass. Future work could investigate metrics that preserve the useful guarantees developed here while being less sensitive to low probability outcomes.

7 Related work

From Adam to module specific optimizers. Adam has long served as the default optimizer across deep learning (Kingma & Ba, 2014). Muon has recently challenged this convention (Jordan et al., 2024). For a linear layer, Muon controls the worst-case change in its output as measured by RMS, resulting in a bound on the spectral norm of the weight update. The resulting steepest descent problem forms the mathematical foundation of the Muon update (Bernstein, 2025a). The Modula project generalizes this perspective and advocates constructing optimizers whose geometry reflects the function of each module (Bernstein, 2025b; Large et al., 2024). Modular duality and Scion provide systematic ways to construct updates from such geometries (Bernstein & Newhouse, 2024; Pethick et al., 2025). We extend this principle to modules involving softmax.

Geometry on probability distributions. A related line of work connects parameter updates to the probability distributions they induce. Natural gradient equips a family of probability distributions with the Fisher geometry and chooses updates according to the resulting notion of steepest descent (Amari, 1998; Martens, 2020). TRPO makes this connection explicit by constraining the expected KL divergence between the action distributions induced by the original and updated policies (Schulman et al., 2015). Hilbert's projective metric offers a different geometry for positive vectors. Beginning with the contraction theory of positive operators, this geometry has subsequently appeared in matrix scaling, clustering, nonlinear embeddings, and the study of probability simplices (Birkhoff, 1957; De La Harpe, 1993; Cuturi, 2013; Nielsen & Sun, 2018; 2023). More recently, Cohen & Fausti (2024) established a sharp relationship between Hilbert distance and total variation. We draw on this literature to study the parameter geometry induced by Hilbert distance on the probability distributions produced by softmax.

Optimizers for the LM head and attention. The LM head and attention have each motivated specialized optimization methods. For the LM head, Scion (Pethick et al., 2025) includes RowNorm and SCALE (Glentis et al., 2026) proposes a momentum-based variant; Lau & Su (2026) derive projected RowNorm from its permutation and shared-shift symmetries. Attention has motivated a different collection of interventions. Query and key normalization controls the scale of attention scores, while spectral reparameterization addresses attention entropy collapse (Henry et al., 2020; Zhai et al., 2023). MuonClip monitors the maximum attention logits and conditionally rescales the query and key projection weights after a Muon update (Team et al., 2026). Compositional Muon goes further by constraining the first-order change in the product of the query and key matrices (Keigwin et al., 2026). Although row normalization for the LM head and product control for attention appear to be independent design choices, we show that they share common ground in bounding changes in the probability distributions produced by softmax.

8 Conclusion

We develop a unified framework for optimizing composed softmax modules by bounding changes in their output probabilities and deriving the corresponding parameter constraints. This addresses a limitation of Muon's RMS control, which can conceal substantial probability redistribution when applied to the intermediate linear output. We propose using Hilbert distance to constrain these changes in the output distribution directly. Hilbert distance respects the common-shift invariance of softmax and controls both the redistribution of probability mass and changes in cross-entropy loss. Its representation as the variation seminorm of the preactivation perturbation lets us translate a common constraint on the output distribution into a parameter geometry for each module. For the LM head, this gives a row diameter constraint. A tighter ball approximation leads to RowNorm, with centering giving its projected form. For attention, the same principle recovers the first-order QK steepest descent problem underlying Compositional Muon. A single probability constraint thus connects the separately developed RowNorm update and Compositional Muon's QK optimization problem. We offer this framework as a contribution to the broader Modula perspective, demonstrating how the output geometry of composed softmax modules can guide optimizer design. We hope these connections spark new ideas for optimizers and architectures built around composed modules. Hilbert distance provides a starting point for softmax, and we encourage researchers to ask which output geometries could play that role for other compositions.

AI USE STATEMENT

The idea of bounding changes in softmax output probabilities originated with the authors. The choice of Hilbert distance was found by studying many existing distances in the literature. AI-assisted mathematical exploration also informed the selection of Hilbert distance. AI tools were extensively used to survey literature, identify related work, check mathematical arguments, and verify correctness. In particular, the authors initially obtained a weaker approximation bound for RowNorm relative to the diameter-constrained optimum. AI tools subsequently identified Jung's theorem and its applicability in tightening this bound.

The initial manuscript was written by the authors. AI tools were subsequently used to improve wording, technical exposition, and presentation, and to provide critical feedback on the manuscript. The authors reviewed all AI assisted editorial changes before incorporating them. We, the authors, take responsibility for the final content of the paper, including its mathematical claims, proofs, and references.

ETHICS STATEMENT

This work develops a unified framework for optimizing modules which involve softmax. The research presented here does not involve human participants, personal data, or the collection or release of datasets. The framework is general purpose, and all ethical considerations in downstream applications depend on the training data, models, and deployment contexts in which it is used.

REPRODUCIBILITY STATEMENT

The assumptions underlying our theoretical results are stated alongside the derivations in the main text. Appendix A provides the sharpness argument for the total variation bound, the exact dual formulation of the diameter-constrained problem, and the proof of the sharp RowNorm approximation guarantee. Appendix B derives the forward and backward perturbation bounds for the LM head and attention. Appendix A.5 specifies the practical RowNorm update, including centering, momentum, and numerical stabilization. To reproduce Figure 1, Appendix C gives deterministic formulas for the initial logits and all three updates, together with their RMS norms, Hilbert distances, and total variation values.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Jeremy Bernstein. Deriving muon, 2025a. URL https://jeremybernste.in/writing/deriving-muon.

Jeremy Bernstein. The modula docs, 2025b. URL https://docs.modula.systems/.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Krzysztof J Ciosmak. Optimal transport of vector measures. *Calculus of Variations and Partial Differential Equations*, 60(6):230, 2021.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the hilbert metric on probability measures, 2024. URL https://arxiv.org/abs/2309.02413.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Pierre De La Harpe. On hilbert's metric for simplices. *Geometric group theory*, 1:97–119, 1993.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. Memory-efficient llm pretraining via minimalist optimizer design, 2026. URL https://arxiv.org/abs/2506.16659.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

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/.

Heinrich Jung. Ueber die kleinste kugel, die eine räumliche figur einschliesst. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1901(123):241–257, 1901.

Ben Keigwin, Li Yang, Yunzhe Zhang, Dhruv Pai, and Alec DeWulf. Towards compositional steepest descent. 2026. URL https://tilderesearch.com/blog/compositional-muon.

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

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm, 2024. URL https://arxiv.org/abs/2405.14813.

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, lm heads, swiglu mlps, and moe routers, 2026. URL https://arxiv.org/abs/2605.18106.

Bas Lemmens and Roger D Nussbaum. *Nonlinear Perron-Frobenius Theory*. Cambridge University Press, 2012.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Frank Nielsen and Ke Sun. *Clustering in Hilbert's Projective Geometry: The Case Studies of the Probability Simplex and the Elliptope of Correlation Matrices*, pp. 297–331. Springer International Publishing, November 2018. ISBN 9783030025205. doi: 10.1007/978-3-030-02520-5_11. URL http://dx.doi.org/10.1007/978-3-030-02520-5_11.

Frank Nielsen and Ke Sun. Non-linear embeddings in hilbert simplex geometry. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 254–266. PMLR, 2023.

Athanase Papadopoulos and Marc Troyanov. *Handbook of Hilbert geometry*. European Mathematical Society, 2014.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos, 2025. URL https://arxiv.org/abs/2502.07529.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 3 edition, 1976. URL https://books.google.com/books?id=kwqzPAAACAAJ.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. Pmlr, 2015.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Chenxiao Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Qizheng Gu, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yang Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Haoyu Lu, Lijun Lu, Yashuo Luo, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Zeyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Lin Sui, Xinjie Sun, Flood Sung, Yunpeng Tai, Heyi Tang, Jiawen Tao, Qifeng Teng, Chaoran Tian, Chensi Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Si Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Haoning Wu, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Jing Xu, Jing Xu, Junjie Yan, Yuzi Yan, Hao Yang, Xiaofei Yang, Yi Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Siyu Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Shaojie Zheng, Longguang Zhong, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2026. URL https://arxiv.org/abs/2507.20534.

Lieven Vandenberghe. *Convex optimization*, volume 1.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International conference on machine learning*, pp. 40770–40803. PMLR, 2023.

A Proofs and further connections

A.1 RMS and Hilbert distance

Let $z \in \mathbb{R}^n$ be a preactivation vector and let $\Delta z \in \mathbb{R}^n$ be a centered preactivation perturbation, so that $\mathbf{1}^\top \Delta z = 0$. Define $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z + \Delta z)$. Then $2\,\|\Delta z\|_{\mathrm{RMS}} \le \mathcal{H}(p, q) \le \sqrt{2n}\,\|\Delta z\|_{\mathrm{RMS}}$. The softmax identity gives $\mathcal{H}(p, q) = \max_{i,j} |\Delta z_i - \Delta z_j|$. Applying the Cauchy–Schwarz inequality gives

$$|\Delta z_i - \Delta z_j| \le \sqrt{2}\sqrt{\Delta z_i^2 + \Delta z_j^2} \le \sqrt{2\sum_{k=1}^{n} \Delta z_k^2} = \sqrt{2n}\,\|\Delta z\|_{\mathrm{RMS}}\,, \qquad i \ne j.$$

Taking the largest pairwise difference proves the upper bound. For the lower bound, let $m = \min_i \Delta z_i$, $M = \max_i \Delta z_i$, and let $c = (m + M)/2$ be their midpoint. Expanding the square

and using $\sum_i \Delta z_i = 0$ gives

$$\|\Delta z\|_{\mathrm{RMS}}^2 \le \|\Delta z\|_{\mathrm{RMS}}^2 + c^2 = \frac{1}{n}\sum_i (\Delta z_i - c)^2 \le \max_i (\Delta z_i - c)^2 = \left(\frac{M-m}{2}\right)^2 = \frac{\mathcal{H}(p,q)^2}{4}.$$

Thus, Hilbert distance controls the RMS norm of a centered preactivation perturbation with a constant independent of the output width. In the reverse direction, the sharp factor grows as $\sqrt{n}$, with equality attained by $\Delta z = (a, -a, 0, \ldots, 0)$ for $a > 0$. Interestingly, this comparison resembles the standard relation between the spectral and Frobenius norms. For a matrix A of rank r, $\|A\|_{\mathrm{op}} \le \|A\|_F \le \sqrt{r}\,\|A\|_{\mathrm{op}}$. Appendix A.2 gives a centered example whose RMS tends to zero while its total variation tends to $1/2$.

A.2 Probability redistribution under softmax

The sharp total variation bound. Recall that the variation seminorm is $\gamma = \|\Delta z\|_{\mathrm{var}} = \max_\ell \Delta z_\ell - \min_\ell \Delta z_\ell$. For $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z + \Delta z)$, the Hilbert distance bound gives $\mathrm{TV}(p,q) \le \tanh(\gamma/4)$ (Cohen & Fausti, 2024, Theorem 5.1). To approach equality when $\gamma > 0$, choose i, j with $\Delta z_i - \Delta z_j = \gamma$ and set $z_i = -\gamma/2$, $z_j = 0$. The difference between these preactivation coordinates therefore changes from $-\gamma/2$ to $\gamma/2$. Let all other preactivations tend to $-\infty$, concentrating both distributions on i, j. The probability gained by coordinate i then gives

$$\mathrm{TV}(p,q) \longrightarrow \frac{e^{\gamma/2}}{1+e^{\gamma/2}} - \frac{1}{1+e^{\gamma/2}} = \tanh\left(\frac{\gamma}{4}\right).$$

This approaches the upper bound through strictly positive softmax distributions, proving the claimed supremum. For $\gamma = 0$, the update is a common shift and both sides are zero. The following example shows that centering does not prevent RMS from vanishing while probability redistribution remains substantial.

Vanishing RMS with nonvanishing probability redistribution. Take $n \ge 3$, initial preactivation vector $z = 0$, and $a_n = \log(n-1)$. The initial probabilities and centered update are

$$p = \frac{\mathbf{1}}{n}, \qquad \Delta z = a_n \left(1 - \frac{1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n}\right)^\top, \qquad \mathbf{1}^\top \Delta z = 0.$$

By softmax's common-shift invariance, $q = \mathrm{softmax}(\Delta z)$ satisfies

$$q_1 = \frac{e^{a_n}}{e^{a_n} + n - 1} = \frac{1}{2}, \qquad q_i = \frac{1}{e^{a_n} + n - 1} = \frac{1}{2(n-1)} \quad (i > 1).$$

Yet, as $n \to \infty$,

$$\|\Delta z\|_{\mathrm{RMS}} = \frac{\sqrt{n-1}}{n}\log(n-1) \longrightarrow 0, \qquad \mathrm{TV}(p,q) = \frac{1}{2} - \frac{1}{n} \longrightarrow \frac{1}{2}.$$

Even after centering, vanishing RMS does not ensure vanishing probability redistribution uniformly across widths. Meanwhile, $\mathcal{H}(p,q) = \log(n-1)$ captures the growing relative change.

A.3 The exact dual of the diameter constraint

The diameter constraint compares pairs of update rows, and its dual will use one vector for each of these pairs. Let $G \in \mathbb{R}^{V \times d}$ have rows $g_i^\top$, with $V \ge 2$ and $\sum_i g_i = 0$. For $\rho_{\mathrm{head}} > 0$, we return to

$$\min_{\Delta W} \langle G, \Delta W \rangle_F \qquad \text{subject to} \qquad \|\Delta w_i - \Delta w_j\|_2 \le \rho_{\mathrm{head}} \quad \text{for all } i < j.$$

We derive an exact dual formulation of this minimum.

Pairwise flow variables. For each pair $i < j$, introduce an independent vector $\varphi_{ij} \in \mathbb{R}^d$, and write Φ for their collection. Define $\varphi_{ji} = -\varphi_{ij}$ and $\varphi_{ii} = 0$ to simplify the indexing. The dual requires these vectors to reconstruct the gradient rows,

$$\sum_{j=1}^{V} \varphi_{ij} = g_i, \qquad i = 1, \ldots, V. \tag{15}$$

For a pair $i < j$, the vector φ_{ij} contributes to row i and its negative contributes to row j. These contributions cancel when summed over all rows,

$$\sum_i \sum_j \varphi_{ij} = \sum_{i<j} (\varphi_{ij} + \varphi_{ji}) = 0.$$

The zero-sum condition on G is therefore necessary. It also guarantees that the balance equations are feasible. For example, take $\varphi_{iV} = g_i$ for $i < V$ and set all other independent vectors to zero. The remaining equation holds because $g_V = -\sum_{i<V} g_i$.

The exact dual identity. Among all representations satisfying Equation 15, we minimize the sum of the vector norms. The resulting identity is

$$\min_{D(\Delta W) \le \rho_{\text{head}}} \langle G, \Delta W \rangle_F = -\rho_{\text{head}} \min_{\substack{\Phi \\ \sum_j \varphi_{ij} = g_i \ \forall i}} \sum_{i<j} \|\varphi_{ij}\|_2 . \tag{16}$$

Both minima are attained. The minimum cost on the right, multiplied by ρ_{head}, equals the greatest first-order decrease $\Delta_\star$.

Proof. We first check that the original problem has a finite, attained minimum despite its freedom under common translations. Subtracting Δw_V from every update row preserves the diameter and changes the objective by $-\langle \sum_i g_i, \Delta w_V \rangle = 0$. For this existence argument, we may therefore choose a representative with $\Delta w_V = 0$. Every other row then has norm at most ρ_{head}. The feasible representatives form a nonempty compact set, on which the continuous objective attains its minimum.

To derive the dual, introduce a variable ξ_{ij} for each row difference,

$$\xi_{ij} = \Delta w_i - \Delta w_j, \qquad \|\xi_{ij}\|_2 \le \rho_{\text{head}}, \qquad i < j.$$

Attach the vector multiplier φ_{ij} to each equality and retain the ball constraints on ξ. The Lagrangian becomes

$$\begin{aligned} \Lambda(\Delta W, \xi, \Phi) &= \langle G, \Delta W \rangle_F + \sum_{i<j} \langle \varphi_{ij}, \xi_{ij} - \Delta w_i + \Delta w_j \rangle \\ &= \sum_i \left\langle g_i - \sum_j \varphi_{ij}, \Delta w_i \right\rangle + \sum_{i<j} \langle \varphi_{ij}, \xi_{ij} \rangle . \end{aligned}$$

For fixed Φ, we minimize Λ over all ΔW and over ξ within the balls. The difference equalities are now represented by the Lagrangian terms, so they are not imposed during this inner minimization. Any nonzero coefficient of an unrestricted row Δw_i makes the infimum $-\infty$. A finite bound therefore requires exactly the balance equations in Equation 15. Under these equations, the terms involving ΔW vanish and the remaining infimum is

$$\inf_{\substack{\xi \\ \|\xi_{ij}\|_2 \le \rho_{\text{head}} \ \forall i<j}} \sum_{i<j} \langle \varphi_{ij}, \xi_{ij} \rangle = -\rho_{\text{head}} \sum_{i<j} \|\varphi_{ij}\|_2 .$$

For $\varphi_{ij} \neq 0$, this value is attained at $\xi_{ij} = -\rho_{\text{head}} \varphi_{ij} / \|\varphi_{ij}\|_2$. When $\varphi_{ij} = 0$, we may choose $\xi_{ij} = 0$. The dual maximizes this lower bound over Φ. Because $-\rho_{\text{head}} < 0$, this amounts to minimizing the sum of norms.

The reformulated problem is convex, and $\Delta W = 0$, $\xi = 0$ satisfies every equality and lies strictly inside every ball. Slater's condition gives strong duality and attainment of the dual optimum (Vandenberghe, Section 5.2.3). This establishes Equation 16. □

The connection to vector transport. The pair vectors can now be read as vector flows, with Equation 15 specifying the net flow at each token. They carry gradient vectors, whose coordinates may have either sign. The cost sums their Euclidean norms and counts each pair once. Assign distance one to each pair of distinct tokens. The diameter constraint then says that $i \mapsto -\Delta w_i / \rho_{\text{head}}$ is 1-Lipschitz, meaning that the distance between its values never exceeds the distance between the tokens. Equation 16 is the finite form of vector transport duality (Ciosmak, 2021, Theorem 2).

The identity characterizes the optimal value without requiring a unique minimizing update. Its transport formulation contains $\binom{V}{2}$ independent vectors in $\mathbb{R}^d$. This quadratic growth in the number of variables motivates the common-ball approximation in the main text.

A.4 THE ROWNORM APPROXIMATION

The main text obtains RowNorm by requiring all update rows to lie in a common ball and then moving that ball to the origin. We first justify why this change of variable preserves the minimum. We then bound the suboptimality introduced by replacing the diameter constraint with the ball constraint.

A.4.1 MOVING THE BALL TO THE ORIGIN

Fix a gradient $G \in \mathbb{R}^{V\times d}$ with rows $g_i^\top$ satisfying $\sum_i g_i = 0$, and let $\rho_{\text{head}} > 0$. Recall that $\mathcal{B}(\zeta, \rho_{\text{head}}/2)$ denotes the closed Euclidean ball with center ζ and radius $\rho_{\text{head}}/2$. In Equation 7, we choose both the update ΔW and the center ζ,

$$\min_{\Delta W,\zeta} \langle G, \Delta W\rangle_F \qquad \text{subject to} \qquad \Delta w_i \in \mathcal{B}(\zeta, \rho_{\text{head}}/2) \quad \forall i.$$

Any center that encloses all the update rows is allowed. In particular, the center is not required to equal the mean row, and the mean row remains unrestricted.

Since the gradient rows sum to zero, translating every update row by the same vector leaves the objective unchanged. Using this freedom, we invoke a change of variable

$$\widehat{\Delta W} = \Delta W - \mathbf{1}\zeta^\top.$$

For any feasible pair $(\Delta W, \zeta)$, the translated rows satisfy

$$\left\|\widehat{\Delta w_i}\right\|_2 = \|\Delta w_i - \zeta\|_2 \le \frac{\rho_{\text{head}}}{2} \qquad \forall i.$$

Thus, every row now lies in a ball centered at the origin. The objective is preserved because

$$\langle G, \Delta W\rangle_F = \left\langle G, \widehat{\Delta W}\right\rangle_F + \underbrace{\left\langle \sum_i g_i, \zeta\right\rangle}_{=0} = \left\langle G, \widehat{\Delta W}\right\rangle_F.$$

After this change of variable, neither the row constraints nor the objective depends on ζ. Equation 7 therefore has the same minimum as

$$\min_{\widehat{\Delta W}} \left\langle G, \widehat{\Delta W}\right\rangle_F \qquad \text{subject to} \qquad \widehat{\Delta w_i} \in \mathcal{B}(0, \rho_{\text{head}}/2) \quad \forall i,$$

which is Equation 8. To complete the equivalence, we also need to return from this problem to the original one. Given any feasible $\widehat{\Delta W}$, choose any $\zeta \in \mathbb{R}^d$ and set

$$\Delta W = \widehat{\Delta W} + \mathbf{1}\zeta^\top.$$

The resulting rows lie in $\mathcal{B}(\zeta, \rho_{\text{head}}/2)$, and the same objective identity shows that their value is unchanged. Choosing $\zeta = 0$ simply gives $\Delta W = \widehat{\Delta W}$. The two problems therefore attain exactly the same objective values.

The feasible set for the problem centered at the origin is a finite product of closed, bounded balls. It is nonempty and compact, so the continuous linear objective attains its minimum. The reverse construction then provides a minimizing pair for the original problem as well. Consequently, every minimizer of the problem centered at the origin gives a minimizer of the common-ball problem, and every minimizing pair of the common ball problem translates back to one.

This change of variable centers the enclosing ball. The rows inside it can still have a nonzero mean. The RowNorm update in Equation 10 is therefore already one minimizing choice for the common ball problem, with $\zeta = 0$. Subtracting its mean row is optional and preserves the objective by the same translation argument.

Moving the ball to the origin introduces no additional suboptimality. The gap considered next comes entirely from requiring the update rows to fit inside a ball of radius $\rho_{\text{head}}/2$ instead of imposing only the original diameter constraint.

A.4.2 THE SUBOPTIMALITY GAP

Moving the ball to the origin preserves the minimum. We now bound the first-order decrease lost by replacing the diameter constraint with the common-ball constraint. Assume $G \neq 0$, $\mathbf{1}^\top G = 0$, and $\rho_{\text{head}} > 0$. Let $\Delta_\star$ and Δ_{ball} denote the greatest first-order decreases under these two constraints, respectively. Since RowNorm attains the common-ball optimum,

$$\Delta_{\text{ball}} = \frac{\rho_{\text{head}}}{2} \sum_i \|g_i\|_2 > 0.$$

Bounding the gap. Jung's theorem bounds the radius needed to enclose a set using its diameter and dimension. Here we can work in the space spanned by the gradient rows, $\mathcal{S} = \text{span}\{g_1, \ldots, g_V\}$. By definition of matrix rank, the dimension of this row space is $k = \text{rank}(G)$.

To see why this space suffices, consider the part of an update row perpendicular to $\mathcal{S}$. It has zero inner product with every gradient row and therefore contributes nothing to the objective. Removing these parts from all rows is an orthogonal projection, which cannot increase pairwise distances. Every feasible update can thus be replaced by one whose rows lie in $\mathcal{S}$, with the same objective value and no larger diameter.

Jung's theorem (Jung, 1901) guarantees that any such rows with diameter at most ρ_{head} lie in a ball of radius at most

$$r_k = \rho_{\text{head}} \sqrt{\frac{k}{2(k+1)}}.$$

Let ζ be the center of such a ball, so that $\|\Delta w_i - \zeta\|_2 \leq r_k$ for every row. We only need to know that this center exists. As in the preceding subsection, $\sum_i g_i = 0$ allows us to subtract ζ without changing the objective. Cauchy–Schwarz then gives

$$-\langle G, \Delta W \rangle_F = -\sum_i \langle g_i, \Delta w_i - \zeta \rangle \leq r_k \sum_i \|g_i\|_2 .$$

This bounds $\Delta_\star$ because every feasible update can be brought into $\mathcal{S}$ without changing its objective value. Every common-ball update also satisfies the diameter constraint, so

$$0 < \Delta_{\text{ball}} \leq \Delta_\star \leq \sqrt{\frac{2k}{k+1}}\, \Delta_{\text{ball}}.$$

Rearranging proves the relative suboptimality bound in Equation 9.

Sharpness. A regular simplex gives a gradient for which equality holds. Choose $g_1, \ldots, g_{k+1}$ as the unit vertices of a regular k-dimensional simplex centered at the origin in $\mathbb{R}^d$. For a vocabulary with $V > k+1$, set $g_i = 0$ for every $i > k+1$. The simplex rows satisfy

$$\sum_{i=1}^{k+1} g_i = 0, \qquad \|g_i\|_2 = 1, \qquad \langle g_i, g_j \rangle = -\frac{1}{k} \quad (i \neq j).$$

The full gradient therefore satisfies $\mathbf{1}^\top G = 0$ and $\text{rank}(G) = k$. Consider the update $\Delta W = -r_k G$. For every $i > k+1$, this sets $\Delta w_i = -r_k g_i = 0$. Distinct simplex update rows have distance

$$\|\Delta w_i - \Delta w_j\|_2 = r_k \sqrt{\frac{2(k+1)}{k}} = \rho_{\text{head}}.$$

Each simplex update row is also at distance $r_k \leq \rho_{\text{head}}$ from every zero update row. The entire update is therefore feasible. It achieves a decrease of $r_k(k+1)$, matching the upper bound on $\Delta_\star$. Since $\Delta_{\text{ball}} = \rho_{\text{head}}(k+1)/2$, the relative gap is

$$\frac{\Delta_\star - \Delta_{\text{ball}}}{\Delta_\star} = 1 - \frac{\rho_{\text{head}}}{2r_k} = 1 - \sqrt{\frac{k+1}{2k}}.$$

This proves sharpness for every permissible rank $1 \leq k \leq \min(d, V-1)$. RowNorm therefore attains the optimal first-order decrease when $k = 1$ and retains at least $1/\sqrt{2}$ of it for every rank. The next subsection describes the practical update with momentum and numerical stabilization.

A.5 Practical RowNorm

In training, we apply row normalization to a first-moment estimate of the gradient. Let $P = I_V - \mathbf{1}\mathbf{1}^\top/V$ denote the projection that subtracts the mean row of a matrix. Starting from $B_{-1} = 0$, for $t = 0, 1, \ldots$ let

$$B_t = \beta B_{t-1} + (1-\beta)G_t, \qquad \widehat{B}_t = \frac{B_t}{1-\beta^{t+1}}, \qquad \widetilde{B}_t = P\widehat{B}_t,$$

where $0 \le \beta < 1$. For numerical stability, define $\mathsf{RN}_{\delta_{\mathrm{num}}}$ by dividing row $\widetilde{b}_{t,i}^\top$ by $\left\|\widetilde{b}_{t,i}\right\|_2 + \delta_{\mathrm{num}}$, with $\delta_{\mathrm{num}} > 0$. With learning rate $\eta_t \ge 0$, the update is

$$\Delta W_t = -\eta_t P\,\mathsf{RN}_{\delta_{\mathrm{num}}}(\widetilde{B}_t), \qquad W_{t+1} = W_t + \Delta W_t. \tag{17}$$

Under our assumptions, the momentum estimate is already centered in exact arithmetic, so the first projection only removes any accumulated numerical drift. Row normalization can introduce a nonzero mean, which the final projection removes. Every normalized row has norm at most one, and centering preserves pairwise differences, so $D(\Delta W_t) \le 2\eta_t$. Consequently, $0 \le \eta_t \le \varepsilon/(2R)$ ensures the required Hilbert distance bound for every $\|h\|_2 \le R$. The computation takes $O(Vd)$ time, uses one first-moment buffer, and implements each projection by subtracting the mean row. Momentum and numerical stabilization preserve the diameter bound on the applied update. The forward and backward estimates therefore remain applicable under their stated assumptions. The first-order suboptimality guarantee does not automatically extend to these practical modifications.

B Forward and backward perturbation bounds

A spectral constraint, as used by Muon, controls perturbations in both the forward and backward computations of a linear layer. For a fixed input h and incoming gradient r,

$$\|\Delta W h\|_2 \le \|\Delta W\|_{\mathrm{op}} \|h\|_2, \qquad \left\|\Delta W^\top r\right\|_2 \le \|\Delta W\|_{\mathrm{op}} \|r\|_2.$$

We establish analogous bounds for our softmax modules below, comparing each module before and after an update at the same inputs. Write $\tau = \tanh(\varepsilon/4)$. The inequality relating Hilbert distance and total variation gives $\mathrm{TV}(p,q) \le \tau$ whenever $\mathcal{H}(p,q) \le \varepsilon$ (Cohen & Fausti, 2024, Theorem 5.1).

B.1 The LM head

The diameter constraint controls the head's probability output and also bounds the change in the gradient passed to its hidden state. Let $p = p_W(h)$ and $q = p_{W+\Delta W}(h)$ at a fixed input $\|h\|_2 \le R$. The forward bound is

$$\mathcal{H}(p,q) \le R D(\Delta W) \le \varepsilon.$$

For a target token c, denote the gradients passed to h by λ_h and λ_h^+. Cross-entropy gives

$$\lambda_h = W^\top(p - e_c), \qquad \lambda_h^+ = (W + \Delta W)^\top(q - e_c),$$

where e_c is the target's one-hot vector. Writing $w_j^\top$ and $\Delta w_j^\top$ for the rows of W and ΔW, subtraction gives

$$\lambda_h^+ - \lambda_h = \sum_j q_j(\Delta w_j - \Delta w_c) + \sum_j (q_j - p_j) w_j.$$

The first sum averages differences between update rows, so its norm is at most $D(\Delta W)$. The second sum accounts for the redistributed probability. Because p and q each have unit mass, the positive and negative coefficients $q_j - p_j$ have the same total mass, namely $\mathrm{TV}(p,q)$. After normalization, they form two weighted averages of the head rows. Their distance cannot exceed $D(W)$, giving

$$\left\|\lambda_h^+ - \lambda_h\right\|_2 \le D(\Delta W) + D(W)\mathrm{TV}(p,q) \le \frac{\varepsilon}{R} + \tau D(W).$$

The current head appears through $D(W)$ because it transmits the changed probability gradient.

B.2 ATTENTION

For attention, we again seek bounds on the change in the output distributions and the resulting change in the gradients passed to the hidden states. The backward calculation now includes both the query and key uses of each hidden state. We keep the inputs $\|h_\ell\|_2 \le R$ and visible sets $\mathcal{J}_i$ fixed. Let $M = W_Q W_K^\top$, and let ΔM denote its change. Equation 12 gives

$$\max_i \mathcal{H}(p_i, p_i^+) \le \frac{2R^2}{\sqrt{d_h}} \|\Delta M\|_{\mathrm{op}} \le \varepsilon.$$

The total variation bound recalled above then gives $\mathrm{TV}(p_i, p_i^+) \le \tau$, where $\tau = \tanh(\varepsilon/4)$.

To connect this probability bound to backpropagation, let r_i denote the incoming gradient on p_i. We hold these incoming gradients fixed when comparing the module before and after the QK update. Differentiating softmax gives the gradient with respect to the scores,

$$a_{ij} = p_{ij}(r_{ij} - \mu_i), \qquad \mu_i = \sum_{j \in \mathcal{J}_i} p_{ij} r_{ij}.$$

Each score $s_{ij} = h_i^\top M h_j / \sqrt{d_h}$ sends this gradient to both its query input and its key input. Denote the corresponding contributions by λ_i^{Q} and $\lambda_{ij}^{\mathrm{K}}$:

$$\lambda_i^{\mathrm{Q}} = \frac{M \sum_{j \in \mathcal{J}_i} a_{ij} h_j}{\sqrt{d_h}}, \qquad \lambda_{ij}^{\mathrm{K}} = \frac{a_{ij} M^\top h_i}{\sqrt{d_h}}.$$

The total gradient passed to h_ℓ adds its query contribution and its key contributions from all queries that can see it,

$$\lambda_\ell = \lambda_\ell^{\mathrm{Q}} + \sum_{i : \ell \in \mathcal{J}_i} \lambda_{i\ell}^{\mathrm{K}}.$$

There are two sources of change in these expressions: the QK product changes, and so does the score gradient. Writing $\Delta a_i = a_i^+ - a_i$ and $\Delta\lambda = \lambda^+ - \lambda$, we obtain

$$\Delta\lambda_i^{\mathrm{Q}} = \frac{\Delta M \sum_{j \in \mathcal{J}_i} a_{ij}^+ h_j + M \sum_{j \in \mathcal{J}_i} \Delta a_{ij} h_j}{\sqrt{d_h}},$$
$$\Delta\lambda_{ij}^{\mathrm{K}} = \frac{a_{ij}^+ \Delta M^\top h_i + \Delta a_{ij} M^\top h_i}{\sqrt{d_h}}.$$

Since every hidden state has norm at most R, the remaining quantities to bound are $\|a_i^+\|_1$ and $\|\Delta a_i\|_1$. The first accompanies the product change, and the second accompanies the current product.

We establish these estimates for one row, temporarily omitting the index i. Write $a_j(p) = p_j(r_j - \mu)$ with $\mu = \sum_j p_j r_j$. Recall that $\mathrm{osc}(r) = \max_j r_j - \min_j r_j$. For the size of $a(p)$, Cauchy–Schwarz reduces the calculation to the variance of the entries of r under p. With $m = (\max_j r_j + \min_j r_j)/2$ denoting the midpoint of their range,

$$\|a(p)\|_1^2 \le \sum_j p_j (r_j - \mu)^2 = \sum_j p_j (r_j - m)^2 - (\mu - m)^2 \le \frac{\mathrm{osc}(r)^2}{4}.$$

For the change in a, we use the probability redistribution measured by total variation. Any change from p to q can be decomposed into transfers from coordinates that lose probability to coordinates that gain it. The total mass transferred in this way is $\mathrm{TV}(p, q)$. It therefore suffices to bound the effect of one transfer and add these bounds.

Suppose an amount α moves from coordinate k to coordinate j, giving $q = p + \alpha(e_j - e_k)$. Here e_j, e_k are coordinate vectors and $0 \le \alpha \le p_k$. Subtracting the score gradients at fixed r gives

$$a(q) - a(p) = \alpha\big[(r_j - \mu)e_j - (r_k - \mu)e_k - (r_j - r_k)q\big].$$

Because q has unit mass,

$$\|a(q) - a(p)\|_1 \le \alpha\big(|r_j - \mu| + |r_k - \mu| + |r_j - r_k|\big) \le 2\alpha\ \mathrm{osc}(r).$$

The last step uses the fact that the three absolute differences sum to twice the range of r_j, r_k, μ. All three values lie within the range of r, since μ is a weighted average of its entries. Adding the bounds over the transfers proves $\|a(q) - a(p)\|_1 \le 2\,\mathrm{osc}(r)\mathrm{TV}(p,q)$.

Returning to the attention rows, the two estimates become

$$\left\|a_i^+\right\|_1 \le \frac{\mathrm{osc}(r_i)}{2}, \qquad \|\Delta a_i\|_1 \le 2\,\mathrm{osc}(r_i)\tau.$$

Apply these to the query and key expressions above, using $\|h_\ell\|_2 \le R$ and the triangle inequality:

$$\begin{aligned}\sum_\ell \|\Delta\lambda_\ell\|_2 &\le \frac{2R}{\sqrt{d_h}} \sum_i \left(\|\Delta M\|_{\mathrm{op}} \left\|a_i^+\right\|_1 + \|M\|_{\mathrm{op}} \|\Delta a_i\|_1\right) \\ &\le \frac{R}{\sqrt{d_h}} \left(\|\Delta M\|_{\mathrm{op}} + 4\,\|M\|_{\mathrm{op}}\,\tau\right) \sum_i \mathrm{osc}(r_i).\end{aligned}$$

The factor of two in the first line accounts for the query and key contributions. Both are included when a position attends to itself. For fixed M and incoming gradients, the bound tends to zero as ε tends to zero under the stated spectral constraint. For finite weight updates, ΔM includes the quadratic term $\Delta W_Q \Delta W_K^\top$. The comparison concerns the QK probability module and does not include changes in the incoming gradients or the value pathway.

C Reproducing Figure 1

Figure 1 holds the initial distribution fixed and varies only the centered update. We construct the initial logits by centering a quadratic profile over $i = 1, \dots, 32$:

$$a_i = -\left(\frac{i - 16.5}{4}\right)^2, \qquad z_i = a_i - \frac{1}{32}\sum_{j=1}^{32} a_j, \qquad p = \mathrm{softmax}(z).$$

The three updates below either shift a central block together or perturb individual preactivation coordinates within it. Here, $\mathbf{1}_{\{\cdot\}}$ is the indicator function, equal to one when its condition holds and zero otherwise:

$$\begin{aligned}\Delta z_i^{\text{blue}} &= 21_{\{13 \le i \le 16\}} - 21_{\{17 \le i \le 20\}}, \\ \Delta z_i^{\text{purple}} &= 21_{\{9 \le i \le 24\}} - 1, \\ \Delta z_i^{\text{red}} &= 21_{\{i=16\}} - 21_{\{i=19\}}.\end{aligned}$$

Each update has zero mean and produces $q = \mathrm{softmax}(z + \Delta z)$. Using

$$\mathrm{RMS}(\Delta z) = \sqrt{\frac{1}{32}\sum_{i=1}^{32} (\Delta z_i)^2}, \qquad \mathcal{H}(p,q) = \max_i \Delta z_i - \min_i \Delta z_i,$$

we obtain the following values, together with the total variation $\mathrm{TV}(p,q) = \frac{1}{2}\sum_i |p_i - q_i|$:

	Blue	Purple	Red
$\mathrm{RMS}(\Delta z)$	1	1	1/2
$\mathcal{H}(p,q)$	4	2	4
$\mathrm{TV}(p,q)$	≈ 0.514	≈ 0.004	≈ 0.430

Purple adds a common increment to coordinates 9–24, which initially carry more than 99.5% of the probability mass. The pairwise odds among coordinates $9, \dots, 24$ remain unchanged, and only a small amount of probability mass moves from the tails into this block. Blue and red instead alter these ratios, producing much larger redistribution under the same RMS bound.