Isotropic Curvature Model for Understanding Deep Learning Optimization: Is Gradient Orthogonalization Optimal?

Weijie Su

University of Pennsylvania

October 31, 2025

Abstract

In this paper, we introduce a model for analyzing deep learning optimization over a single iteration by leveraging the matrix structure of the weights. We derive the model by assuming isotropy of curvature, including the second-order Hessian and higher-order terms, of the loss function across all perturbation directions; hence, we call it the isotropic curvature model. This model is a convex optimization program amenable to analysis, which allows us to understand how an update on the weights in the form of a matrix relates to the change in the total loss function. As an application, we use the isotropic curvature model to analyze the recently introduced Muon optimizer and other matrix-gradient methods for training language models. First, we show that under a general growth condition on the curvature, the optimal update matrix is obtained by making the spectrum of the original gradient matrix more homogeneous—that is, making its singular values closer in ratio—which in particular improves the conditioning of the update matrix. Next, we show that the orthogonalized gradient becomes optimal for the isotropic curvature model when the curvature exhibits a phase transition in growth. Taken together, these results suggest that the gradient orthogonalization employed in Muon and other related methods is directionally correct but may not be strictly optimal. Finally, we discuss future research on how to leverage the isotropic curvature model for designing new optimization methods for training deep learning and language models.

1 Introduction

If asked before 2025 which optimization method is used for training deep learning and language models, most practitioners would have answered Adam [23]. Indeed, since its proposal in 2014, Adam has become almost synonymous with deep learning optimization. Although many potential competitors were introduced, they either failed to scale their performance to larger models, worked well only with sophisticatedly tuned hyperparameters, or the marginal improvement was not significant enough to motivate practitioners to switch from Adam [41]. The consensus in both industry and academia was that Adam's dominance was secure for years to come. For instance, the work on Adam received the Test of Time Award in early 2025 from the International Conference on Learning Representations (ICLR), one of the premier machine learning conferences.

However, Adam's dominance was challenged by a new optimizer called Muon [21] in December 2024, which was shown by its developers to outperform Adam and other existing optimizers on relatively small-scale language models. As with most optimizers when first proposed, many questioned whether the superior performance of Muon could scale to larger, commercial-grade language

models. Indeed, many, including myself, were skeptical, partly because Muon's structure is less amenable to parallelization than Adam's. However, this skepticism quickly subsided as a host of Muon variants were developed to incorporate various practical considerations [32, 3, 1, 35, 26, 18, 48, 28, 19, 20, 22, 16] and its superiority was quickly demonstrated on larger, industry-scale language models [29, 39, 42]. In particular, in February 2025, Muon was shown to require only about 52% of the FLOPs to achieve the same empirical performance as AdamW [30], a popular variant of Adam, on a 16-billion-parameter language model using 5.7 trillion tokens [29]. In July, [42] used Muon to train a frontier language model with 32 billion activated parameters and 1 trillion total parameters. Mounting evidence now suggests that Muon is becoming, and perhaps already is, the new go-to optimizer for training language models in industry, which is remarkable especially given that it all happened in less than one year.

Interestingly, the operational mechanism of Muon is both intuitive and, for many trained in mathematical optimization, surprising. The intuitive part is that Muon recognizes that weights in deep learning architectures are structured as matrices, such as in a feedforward layer or the weight matrices for query, key, and value in an attention mechanism. This is in contrast to Adam, which neglects this matrix structure by vectorizing all weights. To introduce Muon, let $f(W)$ denote the total loss function to be minimized, where W denotes parameters in the form of a matrix (we omit other parameters for simplicity). The gradient on a mini-batch of training samples, denoted G, is also a matrix of the same size. The Muon optimizer updates the weights not along the direction of the original gradient G, but rather along the direction of the orthogonalized gradient. Formally, let $G = U\Sigma V^\top$ be the (compact) singular value decomposition (SVD);[1] then

$$W \leftarrow W - \gamma UV^\top,$$

where γ is the learning rate. Equivalently, the update direction $UV^\top$, often denoted $\mathrm{msgn}(G)$, is the unitary part from the polar decomposition of the gradient G [2, 26, 13]; a variant of Muon was therefore named PolarGrad to better reflect its connection to classical numerical linear algebra [26]. While Muon is not the first optimizer to incorporate the matrix structure of gradients [8, 7, 17, 43, 44], a notable design feature of Muon is its use of polynomial methods to approximate the unitary matrix $UV^\top$ efficiently in a GPU environment.

The immediate surprise, however, is why all singular values should be made constant. A larger singular value suggests that the corresponding singular space is more "promising" for reducing the loss function. Discarding the singular value matrix Σ seems to ignore the varying levels of promise across different singular spaces. Indeed, taking a step back, it is surprising that singular values should be made closer at all, let alone perfectly uniform. A secondary but non-trivial question is why the singular spaces of the original gradient matrix should be preserved in the update. These questions are pressing, as without a theoretical foundation, the practice of deep learning optimization risks becoming an exercise in trial and error. To address these puzzles, a surge of work has emerged to understand Muon [27, 24, 40, 38, 36, 9, 49], much of which adopts a spectral norm perspective [4, 5, 14, 26, 10]. However, it is still unclear how gradient orthogonalization contributes to enhanced optimization performance, and whether it is really optimal from a theoretical viewpoint.

In this paper, we propose an analytically tractable optimization program to model Muon and, more generally, matrix-gradient methods. Our model focuses on finding an update that achieves the steepest descent in an "average" single iteration. Our rationale is that, to model the change in loss, we can assume that beyond the first-order information (the gradient), all higher-order information,

[1] In practice, the orthogonalization is applied to an exponential moving average of the gradients.

including the Hessian and other curvature information, is isotropic across all possible directions. This isotropy assumption is motivated by the immense scale of deep learning models, especially language models, where the architecture and training process do not inherently favor any particular direction. Furthermore, by averaging over a full batch with a very large number of training samples, curvature might exhibit a convergent law. We also impose certain conditions on how the curvature depends on the perturbation magnitude that are consistent with experimental observations. We call this framework the isotropic curvature model.

By analyzing the isotropic curvature model, we obtain several insights into the effective design of matrix-gradient methods. First, the isotropy of our model's design justifies the preservation of the singular spaces of the original gradient matrix in the update. More importantly, our model offers guidance on how the spectrum of the gradient matrix should be modified. By assuming a certain growth condition on the curvature, our model shows that the optimal update matrix is obtained by pushing the singular values of the original gradient matrix closer to each other while preserving their original ordering, a property we refer to as spectrum homogenization. Next, we prove that when the curvature in the isotropic curvature model transitions abruptly from small to large, optimality is attained by setting all singular values to be equal; that is, orthogonalization is optimal in this asymptotic limit.

Several implications follow from our analysis of the isotropic curvature model. Roughly speaking, the model suggests that the design rationale of Muon and many other matrix-gradient methods is conceptually sound, as they respect the matrix structure of the gradients. However, our model implies that Muon is not strictly optimal. The optimal spectrum transformation is never perfectly uniform, because in practice the curvature does not exhibit the asymptotic behavior required for the optimality of orthogonalization. Determining the optimal spectrum depends on specifying the curvature function, and a possible future direction for designing matrix-gradient optimizers is to first approximate the curvature information and then find the optimal update matrix by solving a simple but large-scale convex optimization program.

2 Isotropic Curvature Model

2.1 Derivation

Consider the objective function that we wish to minimize:

$$f(W) = \frac{1}{N}\sum_{i=1}^{N} L(W; x_i, y_i),$$

where W is a matrix parameter of size $m \times n$, L is a (cross-entropy) loss function, and (x_i, y_i) is the i-th training example and its label. This formulation includes the scenario where x_i is the context and y_i is the token for training large language models. The full objective function involves many other parameters, but here we isolate W to study how its update affects the optimization process. Let u_i be the intermediate data from x_i that serves as input to the matrix W. Recognizing that the loss function depends on x_i through the product Wu_i, we can write the loss as $L(Wu_i)$.[2]

Next, we consider how the objective f changes when W is updated to $W + \delta W$ for some small matrix perturbation δW. A good optimizer should find a δW that reduces $f(W + \delta W)$ as much

[2]Here we ignore residual connections, so the function depends on u_i only through Wu_i. We also omit the dependence on the label y_i for simplicity.

as possible. We Taylor expand the loss function L for each sample, rather than the objective f, as this allows us to leverage the matrix structure of the weights W. Assuming sufficient smoothness, we expand the loss function up to the quadratic term with a remainder:

$$\begin{aligned} L(Wu_i + \delta W u_i) =& L(Wu_i) + \langle \nabla L(Wu_i), \delta W u_i \rangle \\ &+ (\delta W u_i)^\top \left[\int_0^1 (1-t) \nabla^2 L(Wu_i + t\delta W u_i) \mathrm{d}t \right] \delta W u_i, \end{aligned}$$

where ∇L is the gradient with respect to the entire input argument of $L(\cdot)$. Note that $\int_0^1 (1-t)\nabla^2 L(x + t\delta x)\mathrm{d}t$ is a weighted average of the Hessians of L along the line segment connecting x and $x + \delta x$. Plugging this into the objective function gives

$$\begin{aligned} f(W + \delta W) =& f(W) + \frac{1}{N} \sum_{i=1}^N \langle \nabla L(Wu_i), \delta W u_i \rangle \\ &+ \frac{1}{N} \sum_{i=1}^N (\delta W u_i)^\top \left[\int_0^1 (1-t) \nabla^2 L(Wu_i + t\delta W u_i) \mathrm{d}t \right] \delta W u_i. \end{aligned}$$

The first-order term simplifies to

$$\sum_{i=1}^N \langle \nabla L(Wu_i), \delta W u_i \rangle = \sum_{i=1}^N \langle \delta W, \nabla L(Wu_i) u_i^\top \rangle = N \langle \delta W, \nabla f(W) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. Hence,

$$\begin{aligned} f(W + \delta W) =& f(W) + \mathrm{Tr}(\delta W \nabla f(W)^\top) \\ &+ \frac{1}{N} \sum_{i=1}^N (\delta W u_i)^\top \left[\int_0^1 (1-t) \nabla^2 L(Wu_i + t\delta W u_i) \mathrm{d}t \right] \delta W u_i \\ \equiv& f(W) + \mathrm{Tr}(\delta W \nabla f(W)^\top) + \frac{1}{N} \sum_{i=1}^N h(\delta W u_i, W u_i). \end{aligned}$$

We approximate $f(W + \delta W)$ by replacing the term $\frac{1}{N} \sum_{i=1}^N h(\delta W u_i, W u_i)$ with a simpler expression that allows us to analyze the optimal δW within a surrogate model. First, we assume that the singular spaces of the Hessian integral $\int_0^1 (1-t)\nabla^2 L(Wu_i + t\delta W u_i)\mathrm{d}t$, which can be viewed as random due to factors like initialization and mini-batch sampling, are independent of the vector $\delta W u_i$. This allows us to approximate

$$\begin{aligned} h(\delta W u_i, W u_i) \equiv& (\delta W u_i)^\top \left[\int_0^1 (1-t) \nabla^2 L(Wu_i + t\delta W u_i) \mathrm{d}t \right] \delta W u_i \\ \approx& \text{ average eigenvalue of } \left[\int_0^1 (1-t) \nabla^2 L(Wu_i + t\delta W u_i) \mathrm{d}t \right] \times \|\delta W u_i\|^2, \end{aligned} \tag{1}$$

where $\|\cdot\|$ throughout the paper denotes the ℓ_2 norm. It is worth noting that this approximation might not be precise for an individual i, but it becomes more faithful when averaged over all i.

To further simplify $\frac{1}{N}\sum_{i=1}^{N} h(\delta W u_i, W u_i)$, we need to make assumptions on the average spectrum of the Hessian integral. To this end, we impose the *isotropy* assumption that the spectral distribution depends on the perturbation $\delta W u_i$ only through its Euclidean norm $\|\delta W u_i\|$. This assumption is motivated by the fact that deep learning models, especially large language models, are vast systems where no direction is favored by design. Moreover, weights are often initialized from isotropic Gaussian distributions, and no specific direction is explicitly favored during training. To further simplify, we drop the dependence of $h(\delta W u_i, W u_i)$ on $W u_i$, which is a limitation of our model but may be plausible since we are conditioning on the current weights W. Taken together, under our isotropic assumption, both the average spectrum and $\|\delta W u_i\|^2$ in (1) are fully determined by $\|\delta W u_i\|$. Hence, we introduce a *curvature function* H such that $H(\|\delta W u_i\|) \approx h(\delta W u_i, W u_i)$.

This leads to the following optimization program for modeling the one-step update:

$$\min_{\delta W} f(W + \delta W) \approx f(W) + \operatorname{Tr}(\delta W \nabla f(W)^\top) + \frac{1}{N}\sum_{i=1}^{N} H(\|\delta W u_i\|).$$

However, the presence of the input data vectors u_i still complicates the analysis. To overcome this, we make another isotropy assumption by assuming all u_i are independent and identically distributed samples from a sphere centered at the origin. Again, the rationale for this isotropy is that no input directions are favored and the data are diverse enough to plausibly span all directions. This allows us to approximate the sum with an expectation:

$$\frac{1}{N}\sum_{i=1}^{N} H(\|\delta W u_i\|) \approx \mathbb{E}_{\zeta \sim \text{sphere}}\, H(\|\delta W \zeta\|).$$

Finally, for notational convenience, we let $Q = -\delta W$ denote the negative update direction (so the updated weights are $W - Q$). We use $G = \nabla f(W)$ to denote the gradient, which in practice is a stochastic gradient from a mini-batch. The update matrix Q specifies not only the direction but also incorporates the learning rate. This yields the following optimization program:

$$\min_{Q} - \operatorname{Tr}(Q G^\top) + \mathbb{E}_{\zeta \sim \text{sphere}}\, H(\|Q\zeta\|). \tag{2}$$

We call this the *isotropic curvature model*. Since the radius of the sphere can be absorbed into H, we assume for simplicity that ζ is sampled from the unit sphere in $\mathbb{R}^n$. We remark that this model is not intended to capture a specific iteration in deep learning optimization, which is subject to randomness and arbitrariness. Instead, it seeks to model, in an average or distributional sense, how a single iteration proceeds.

2.2 Assumptions on Curvature

The simplicity of the isotropic curvature model (2) is that it represents complex high-order information as a single univariate, increasing function H. To analyze this model, however, we need to make a further assumption on the curvature function H. Since H contains all high-order terms, we can assume it grows rapidly. Indeed, to ensure that (2) is well-defined, we need to assume, for example, that $H(r)/r > C$ for some sufficiently large constant C (depending on G) when r is large enough. This is almost a minimal requirement for the optimization program to have a well-defined optimal solution. Otherwise, the objective function $- \operatorname{Tr}(Q G^\top) + \mathbb{E}_{\zeta \sim \text{sphere}}\, H(\|Q\zeta\|)$ could diverge

to $-\infty$. For instance, if $G = C'I$, one could take $Q = cI$ for a large constant C' and let $c \to \infty$.[3] In practice, $H(r)$ does not grow to infinity as $r \to \infty$; for a very large r, the loss would plateau at the level of a random predictor, which is finite. To ensure the uniqueness of the solution in the generic case, we adopt another technical assumption that H is convex. This ensures that (2) is a convex program in Q, because $\|Q\zeta\|$ is a convex function of Q and H is an increasing convex function, making their composition $H(\|Q\zeta\|)$ convex.

Next, we examine the growth of H more closely. When $r > 0$ is very small, the high-order information is dominated by the second-order term (the Hessian). Hence, it is reasonable to assume $H(r) \approx cr^2$ for sufficiently small r. When r becomes larger, it is no longer reasonable to assume a constant average of all eigenvalues for the Hessian integral. To account for this, we may assume a power law $H(r) \approx c'r^{2+\alpha}$ and leave the determination of the constant α to empirical experiments.

From experiments on GPT-2 [33], as shown in Figure 1, we plot approximations of $H(r)/r^2$. When r is small, $H(r)/r^2$ is nearly constant, indicating that a quadratic approximation holds. When r exceeds approximately $10^{0.5}$, $H(r)/r^2$ grows roughly as a power law r^α, with a typical positive α around 0.2. This shows that the high-order terms grow at a super-quadratic rate in practical problems. This is not unexpected, as for larger r, third-order and other higher-order effects may kick in, causing faster growth than quadratic. In other words, this growth condition shows that, on average, the Hessian tends to have a larger spectrum when evaluated further from the current point.

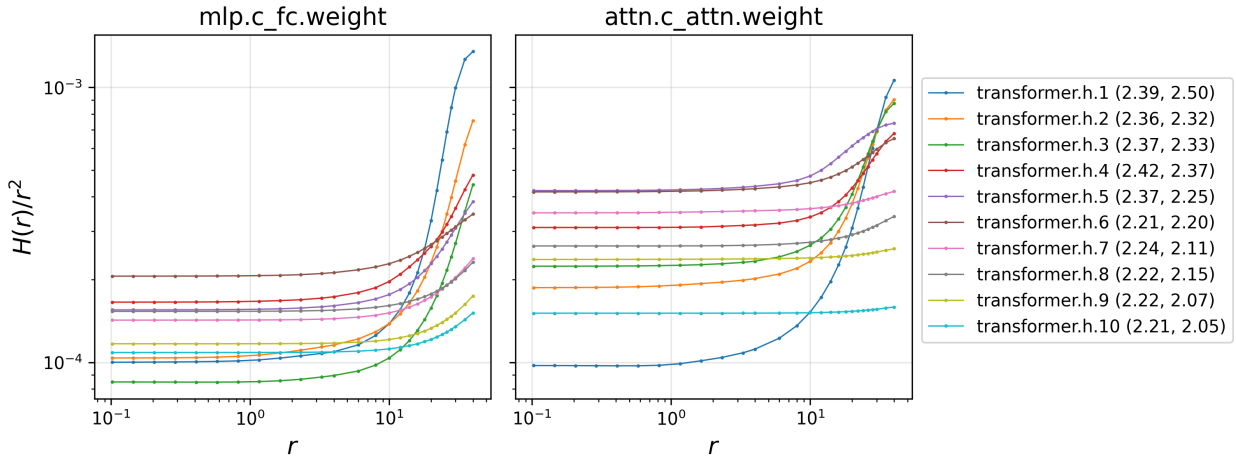

Figure 1: Numerical approximation of $H(r)/r^2$ on GPT-2 (small) [33], for fully connected layers (left) and attention layers (right). The values in parentheses in the legend indicate the estimated exponent $2 + \alpha$ of $H(r)$ starting from $r = 10^{0.5}$. For example, 2.39 and 2.50 in the first line are the exponents for the left and right panels, respectively. We sample 100 random directions δW from a standard normal distribution, sample the first 300 examples from the C4-newslike validation split [34] and truncate each example to 100 tokens. For each radius r, we construct perturbations $\delta W u_i$, where u_i is the input to the layer, and renormalize each $\delta W u_i$ to have norm r. We then compute H by subtracting $L(Wu_i) + \langle \nabla L(Wu_i), \delta W u_i \rangle$ from $L(Wu_i + \delta W u_i)$. This yields a total of $100 \times 300 \times (100 - 1) = 2{,}970{,}000$ remainder terms for approximating $H(r)/r^2$ per radius r. Note that the growth of H will eventually plateau for very large r.

[3]Similar examples can be constructed for non-square problems.

3 Analysis of the Model

We present three results in three subsections, based on progressively stronger assumptions. Proofs are provided in Section 4.

3.1 Alignment of Singular Spaces

The isotropic curvature model is rotationally invariant in the following sense: if the gradient G is replaced by $O_1 G O_2$ for two orthogonal matrices O_1 and O_2, then the objective function

$$-\operatorname{Tr}(QG^\top) + \mathbb{E}_{\zeta \sim \text{sphere}}\, H(\|Q\zeta\|)$$

remains invariant if the update matrix Q is replaced by $O_1 Q O_2$.

This observation naturally suggests an alignment of singular spaces between the gradient matrix and the optimal update from the isotropic curvature model, as stated in the following proposition. Below, we only assume that the optimization program (2) admits a finite solution, which is implied, for example, if H grows sufficiently fast. This condition is weaker than the super-quadratic growth assumption introduced in Section 3.2.

Proposition 3.1. *There exists a global solution $Q^\star$ to the optimization program of the isotropic curvature model*

$$\min_Q -\operatorname{Tr}(QG^\top) + \mathbb{E}_{\zeta \sim \text{sphere}}\, H(\|Q\zeta\|)$$

such that the singular spaces of $Q^\star$ and G are aligned.

That is, there exist orthogonal matrices U and V and diagonal matrices with nonnegative entries Σ and $\Sigma^\star$ such that $G = U\Sigma V^\top$ and $Q^\star = U\Sigma^\star V^\top$. This invariance of the row and column spaces would be lost if the update Q were treated as a vector, neglecting its matrix structure. This result is consistent with the orthogonalization in Muon, which modifies only the spectrum of the gradient matrix while leaving both its left and right singular spaces unchanged.[4]

The statement of Proposition 3.1 uses "there exists" because, in the degenerate case where G has repeated singular values, the singular vectors are defined only up to a unitary rotation within each degenerate block. In the *generic* case, every global solution $Q^\star$ necessarily has singular spaces aligned with those of G.

3.2 Spectrum Homogenization

Building on the alignment of singular spaces, we now use the isotropic curvature model to analyze the spectrum of the optimal update, which provides insight into orthogonalized gradient methods.

The discussion in Section 2.2 and the empirical observations in Figure 1 suggest that $H(r) \propto r^{2+\alpha}$ for $\alpha > 0$ over a certain range of r. Then $H(\sqrt{x}) \propto x^{1+\alpha/2}$. Note that $x^{1+\alpha/2}$ is a convex function of x. This motivates the following assumption to characterize the growth of the curvature function H.

Assumption 3.2. *The curvature function H, defined on $[0, \infty)$, is strictly increasing, and the function $x \mapsto H(\sqrt{x})$ is convex in x.*

[4]Note that this is not the case for some matrix-gradient methods such as SOAP [44].

Roughly speaking, this assumption corresponds to a super-quadratic growth of the curvature. Note that this assumption also implies that $H(r)$ is convex in r.

Any curvature function satisfying Assumption 3.2 grows sufficiently fast. This implies the assumption needed for Proposition 3.1, which allows us to simultaneously unitarily diagonalize $Q^\star$ and G: let $G = U\Sigma V^\top$ and $Q^\star = U\Sigma^\star V^\top$ be the SVDs of G and $Q^\star$, respectively. Write $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$ and $\Sigma^\star = \mathrm{diag}(\sigma_1^\star, \ldots, \sigma_{\min\{m,n\}}^\star)$. By definition, these singular values are all nonnegative.

Theorem 1. *Under Assumption 3.2, there exists a global solution $Q^\star$ to the isotropic curvature model whose spectrum preserves the ordering of the spectrum of G but is more homogeneous. That is, the following two statements hold:*

(a) For any $1 \le i, j \le \min\{m, n\}$, if $\sigma_i \ge \sigma_j$, then $\sigma_i^\star \ge \sigma_j^\star$.

(b) For all $1 \le i, j \le \min\{m, n\}$, we have[5]

$$\frac{\max\{\sigma_i^\star, \sigma_j^\star\}}{\min\{\sigma_i^\star, \sigma_j^\star\}} \le \frac{\max\{\sigma_i, \sigma_j\}}{\min\{\sigma_i, \sigma_j\}}. \tag{3}$$

Remark 3.1. In fact, Part (a) follows from the proof of Proposition 3.1 in the generic case. However, we prefer to state it along with Part (b). Conventionally, singular values in an SVD are ordered by default (e.g., in descending order). Here, we do not impose such an ordering on the SVDs of $Q^\star$ and G; otherwise, Part (a) would be trivially implied.

The second statement says that the singular values of the solution matrix $Q^\star$ are more uniform than those of the original gradient G. In this sense, the isotropic curvature model suggests that the optimal update direction should have singular values that are more homogeneous than those of the original gradient. We call this property, revealed by the theorem, *spectrum homogenization*. Spectrum homogenization is scale-invariant; thus, this theorem does not provide information about the magnitudes of the singular values or specify a learning rate, which is often determined empirically.

Theorem 1 implies that the singular values should ideally maintain the same ordering as those of the original gradient matrix. To this end, the update direction matrix should have its singular values obtained from those of the original gradient via a monotone transformation. However, this is typically not the case when implementing matrix-gradient methods with polynomial approximations, such as the Newton–Schultz iteration [21], which are generally not monotone. This suggests a potential direction for future research: developing efficient methods to solve the optimization problem in (2) while preserving the ordering of singular values.

For illustration, consider $H(r) = cr^4$ for some $c > 0$, which satisfies Assumption 3.2. The optimization program of the isotropic curvature model is equivalent to

$$\min_{\tilde\sigma_1, \ldots, \tilde\sigma_k} -\sum_{i=1}^k \sigma_i \tilde\sigma_i + \frac{c}{n(n+2)} \left[\left(\sum_{i=1}^k \tilde\sigma_i^2 \right)^2 + 2 \sum_{i=1}^k \tilde\sigma_i^4 \right],$$

where $k = \min\{m, n\}$. The optimal solution $(\sigma_1^\star, \ldots, \sigma_k^\star)$ must satisfy

$$(\sigma_j^\star)^3 + D\sigma_j^\star - \frac{n(n+2)\sigma_j}{8c} = 0 \tag{4}$$

[5]By convention, we take $\frac{0}{0} = 1$, $\frac{1}{0} = \infty$, and $\infty \le \infty$.

for all j, where $D = \frac{1}{2}\sum_{i=1}^{k}(\sigma_i^\star)^2$. It is straightforward to see that this system of cubic equations has a unique real root. Each component of this root is nonnegative, and it is positive if the corresponding $\sigma_j > 0$. It is also clear that if $\sigma_i > \sigma_j$, then we must have $\sigma_i^\star > \sigma_j^\star$. To see the second statement of Theorem 1, assume $\sigma_i > \sigma_j > 0$. From (4), it follows that

$$\frac{(\sigma_i^\star)^3 + D\sigma_i^\star}{(\sigma_j^\star)^3 + D\sigma_j^\star} = \frac{\sigma_i}{\sigma_j}.$$

Thus,

$$\frac{\sigma_i^\star}{\sigma_j^\star} = \frac{(\sigma_j^\star)^2 + D}{(\sigma_i^\star)^2 + D} \cdot \frac{\sigma_i}{\sigma_j} < \frac{\sigma_i}{\sigma_j}.$$

In contrast, if $H(r) = cr^2$, the solution $Q^\star$ to the isotropic curvature model is simply proportional to G.

3.3 Orthogonalization

Intuitively, the most extreme form of spectrum homogenization is orthogonalization, which makes all singular values equal. To study when this occurs, we next consider a scenario where the curvature function H increases very rapidly. This means its derivative $H'(r)$ increases sharply from a small value to a large value over a short range. In the limit, we consider this range to be zero, creating a jump discontinuity. We assume there is a *kink* at some radius $\tilde{r}$, such that $H'(\tilde{r}-)$ is small while $H'(\tilde{r}+)$ is large. Our next result shows that under this extreme growth condition, the optimal solution to the isotropic curvature model is orthogonalization.

To formally state our result, we consider the following assumption.

Assumption 3.3. *The curvature function H is a non-decreasing convex function such that there exists $\tilde{r} > 0$ where the left-derivative is $H'(\tilde{r}-) = A$ and the right-derivative is $H'(\tilde{r}+) = B$ for some $0 \le A < B < \infty$.*

This assumption states that the subdifferential of H at $\tilde{r}$ is the interval $[A, B]$. Throughout this subsection, we consider $m \ge n$, meaning Q has at least as many rows as columns. We assume that the gradient matrix G is of full rank; that is, $\sigma_{\min\{m,n\}} = \sigma_n > 0$.

Theorem 2. *Let H satisfy Assumption 3.3 with a sufficiently small A and a sufficiently large B. Then there exists an optimal solution $Q^\star$ to the isotropic curvature model that takes the form $Q^\star = c \cdot \mathrm{msgn}(G) \equiv cUV^\top$ for some scalar $c > 0$, where U and V are from the SVD of G.*

This theorem provides a justification for the orthogonalization in Muon and its extensions like PolarGrad [26], which can be understood as operating under an implicit assumption that the curvature function grows slowly at first and then suddenly "takes off." This theorem can be viewed as a limiting counterpart to Theorem 1, as orthogonalization is the most extreme form of homogenization.

The intuition behind this theorem can be gleaned from its proof in Section 4. The first-order optimality condition requires that G be an element of the subdifferential of the term $\mathbb{E}_{\zeta\sim\text{sphere}}\, H(\|Q\zeta\|)$. As revealed by the proof, the scalar c in Theorem 2 is $\tilde{r}$. When $r < \tilde{r}$, the derivative $H'(r)$ is small, and when $r > \tilde{r}$, it is large. The first-order condition can only be satisfied if the wide interval of the subdifferential at $\tilde{r}$ is leveraged. The stationary point must therefore occur for a $Q^\star$ such

that $\|Q^\star\zeta\| = \tilde{r}$ almost surely for ζ on the unit sphere, which implies that $Q^\star$ must have scaled orthonormal columns. This is why Theorem 2 (and Proposition 3.4 below) requires the technical condition $m \geq n$. However, as we will remark in the proof in Section 4, this condition could be dropped, albeit at the cost of more complicated mathematical statements.

The kink condition in Assumption 3.3 is also necessary for gradient orthogonalization, as shown in the next result.

Proposition 3.4. *Let H satisfy Assumption 3.2 and let G not be a (scaled) orthonormal matrix; that is, its singular values are not all equal. If the isotropic curvature model has a global solution that takes the form $Q^\star = c \cdot \operatorname{msgn}(G)$ for some scalar $c > 0$, then H must have a kink as described in Assumption 3.3.*

However, this interpretation of Theorem 2 and Proposition 3.4 requires caution. While our numerical results suggest that H grows quickly, it is unlikely to exhibit the extreme case of a derivative jumping suddenly from a small value to a large one. In light of this, the gradient orthogonalization in Muon could be interpreted as a reasonable approach from an extreme-case perspective.

In short, our analysis suggests that the optimal strategy is generally not full orthogonalization, but rather homogenization to some degree. This is consistent with numerical experiments showing that rough and exact orthogonalization yield similar performance in pretraining large language models [42]. Determining the optimal level of homogenization presents an opportunity for future research.

4 Proofs

4.1 Proofs for Sections 3.1 and 3.2

The proof of Proposition 3.1 immediately follows from von Neumann's trace inequality.

Lemma 4.1 (von Neumann's trace inequality)**.** *Let $A = U_1\Sigma_1V_1^\top$ and $B = U_2\Sigma_2V_2^\top$ be the SVDs of two matrices of the same size. Then*

$$|\operatorname{Tr}(AB^\top)| \leq \operatorname{Tr}(\Sigma_1\Sigma_2^\top).$$

Note that we assume the default SVD where the singular values in both Σ_1 and Σ_2 are in decreasing order.

Proof of Proposition 3.1. To see how Proposition 3.1 follows, note that the second term in the objective, $\mathbb{E}_{\zeta\sim\text{sphere}}\, H(\|Q\zeta\|)$, does not change if one varies the singular spaces of Q, as long as the spectrum of Q is fixed.

Let $Q^\star$ be any global solution to the isotropic curvature model and write $Q^\star = U_Q\Sigma^\star V_Q^\top$ for its SVD. Let $G = U\Sigma V^\top$ be the SVD of the gradient matrix. By the optimality of $Q^\star$, we know that $Q^\star$ maximizes $\operatorname{Tr}(QG^\top)$ while fixing its singular values, $\Sigma^\star$. By Lemma 4.1,

$$\operatorname{Tr}(Q^\star G^\top) \leq \operatorname{Tr}(\Sigma^\star\Sigma^\top),$$

and equality holds if $U_Q = U$ and $V_Q = V$. This implies that the matrix $\tilde{Q} = U\Sigma^\star V^\top$, which has the same singular values as $Q^\star$, yields an objective value less than or equal to that of $Q^\star$. Thus, $U\Sigma^\star V^\top$ must also be a global solution, and it has singular spaces aligned with those of G.

□

Remark 4.1. The equality in von Neumann's trace inequality holds if and only if the singular spaces are exactly aligned in the generic case where all singular values are distinct. In this case, the proof of Proposition 3.1 reveals that the singular values of G and $Q^\star$ have the same ordering.

Next, we prove Theorem 1.

Proof of Theorem 1. As noted in the proof of Proposition 3.1, Part (a) follows from aligning the singular spaces in the generic case. Now we prove Part (b).

For simplicity, write $q(x) = H(\sqrt{x})$, which is convex by assumption. The optimization problem becomes

$$\min_{Q} -\operatorname{Tr}(QG^\top) + \mathbb{E}_\zeta\, q(\|Q\zeta\|^2),$$

where we omit $\sim$ sphere in the subscript for brevity.

From Proposition 3.1, we can assume without loss of generality that both G and the optimal solution $Q^\star$ are diagonal. Let $Q = \operatorname{diag}(\tilde{\sigma}_1, \dots, \tilde{\sigma}_k)$, where $k = \min\{m, n\}$. Then the program becomes

$$\min_{\tilde{\sigma}_1,\dots,\tilde{\sigma}_k} -\sum_{i=1}^{k} \sigma_i \tilde{\sigma}_i + \mathbb{E}_\zeta\, q(\tilde{\sigma}_1^2\zeta_1^2 + \cdots + \tilde{\sigma}_k^2\zeta_k^2).$$

If $\sigma_i = 0$ for some i, then since q is strictly increasing, the optimal value $\sigma_i^\star$ must be 0 as well. This immediately shows that (3) holds whenever one of σ_i and σ_j is 0.

Now we consider the non-degenerate case where $\sigma_i\sigma_j \neq 0$. Let $(\sigma_1^\star, \dots, \sigma_k^\star)$ be a minimizer, which must exist since q is convex and strictly increasing. Then the first-order optimality conditions are

$$2\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_i^2 \sigma_i^\star \right] = \sigma_i$$

for each $i = 1, \dots, k$. Comparing the conditions for indices i and j:

$$2\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_i^2 \sigma_i^\star \right] = \sigma_i, \quad 2\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_j^2 \sigma_j^\star \right] = \sigma_j. \tag{5}$$

Without loss of generality, assume $\sigma_j > \sigma_i$. By Part (a), this implies $\sigma_j^\star \geq \sigma_i^\star$. However, strict inequality must hold ($\sigma_j^\star > \sigma_i^\star$), because if $\sigma_j^\star = \sigma_i^\star$, the first-order conditions (5) would imply $\sigma_j = \sigma_i$, which contradicts our assumption. To prove homogenization, we need to show $\sigma_j^\star/\sigma_i^\star \leq \sigma_j/\sigma_i$, which is equivalent to $\sigma_j^\star\sigma_i - \sigma_i^\star\sigma_j \leq 0$. We examine the sign of this difference:

$$\begin{aligned}
&\sigma_j^\star\sigma_i - \sigma_i^\star\sigma_j \\
&= \sigma_j^\star \cdot 2\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_i^2 \sigma_i^\star \right] - \sigma_i^\star \cdot 2\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_j^2 \sigma_j^\star \right] \\
&= 2\sigma_i^\star\sigma_j^\star\,\mathbb{E}_\zeta \left[\partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_i^2 - \partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right) \zeta_j^2 \right] \\
&= 2\sigma_i^\star\sigma_j^\star\,\mathbb{E}_\zeta (\zeta_i^2 - \zeta_j^2) \partial q \left(\sum_{l=1}^{k} (\sigma_l^\star)^2 \zeta_l^2 \right).
\end{aligned}$$

The proof reduces to showing that the expectation is negative. Let $S(\zeta) = \sum_{l \neq i,j} (\sigma_l^\star)^2 \zeta_l^2$. Due to symmetry between ζ_i and ζ_j,

$$\mathbb{E}_\zeta(\zeta_i^2 - \zeta_j^2)\partial q\left(S(\zeta) + (\sigma_i^\star)^2\zeta_i^2 + (\sigma_j^\star)^2\zeta_j^2\right) = \mathbb{E}_\zeta(\zeta_j^2 - \zeta_i^2)\partial q\left(S(\zeta) + (\sigma_i^\star)^2\zeta_j^2 + (\sigma_j^\star)^2\zeta_i^2\right)$$

Hence, it suffices to show that the expectation below is negative:

$$(\zeta_i^2 - \zeta_j^2)\left[\partial q\left(S(\zeta) + (\sigma_i^\star)^2\zeta_i^2 + (\sigma_j^\star)^2\zeta_j^2\right) - \partial q\left(S(\zeta) + (\sigma_i^\star)^2\zeta_j^2 + (\sigma_j^\star)^2\zeta_i^2\right)\right].$$

Let $A = S(\zeta) + (\sigma_i^\star)^2\zeta_i^2 + (\sigma_j^\star)^2\zeta_j^2$ and $B = S(\zeta) + (\sigma_i^\star)^2\zeta_j^2 + (\sigma_j^\star)^2\zeta_i^2$. Then $B - A = ((\sigma_j^\star)^2 - (\sigma_i^\star)^2)(\zeta_i^2 - \zeta_j^2)$. Since we assumed $\sigma_j^\star > \sigma_i^\star$, the sign of $B - A$ is the same as the sign of $\zeta_i^2 - \zeta_j^2$. By the convexity of q, its derivative ∂q is non-decreasing, so $\partial q(B) - \partial q(A)$ also has the same sign as $B - A$. Thus, $(\zeta_i^2 - \zeta_j^2)(\partial q(A) - \partial q(B)) \leq 0$. The inequality is strict unless $\zeta_i^2 = \zeta_j^2$ or the derivative is constant. After taking the expectation over ζ, the strict inequality holds, which completes the proof.

□

Remark 4.2. The convexity of $H(\sqrt{x})$, which ensures super-quadratic growth, can be slightly relaxed. The theorem still holds if $H(r)$ grows slowly for small r, such as $H(r) = O(r^2)$, provided $H(\sqrt{x})$ is convex for sufficiently large x.

4.2 Proofs for Section 3.3

We first state a lemma before proving Theorem 2. Recall that $\zeta = (\zeta_1, \ldots, \zeta_n)$ is uniformly sampled from the unit sphere in $\mathbb{R}^n$.

Lemma 4.2. *For any positive constants $C_1, \ldots, C_n$, there exist positive constants $a < b$ and a random variable η satisfying $a \leq \eta \leq b$ almost surely such that*

$$\mathbb{E}\,\eta\zeta_i^2 = C_i$$

for all $i = 1, \ldots, n$.

Proof of Lemma 4.2. Consider the set of all random variables η that are bounded away from 0 and ∞. Define a linear map $\mathcal{E}$ from this set to $\mathbb{R}^n$ as

$$\mathcal{E}(\eta) = (\mathbb{E}\,\eta\zeta_1^2, \ldots, \mathbb{E}\,\eta\zeta_n^2).$$

The image of this map, denoted by K, is the set of all achievable expectation vectors. Clearly, K is a convex cone in the positive orthant $\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n : x_i > 0 \text{ for all } i\}$. If K is the entire positive orthant, the proof is complete.

Now we prove by contradiction by assuming K is not the entire positive orthant. Since K is convex, the interior of its closure $\bar{K}$ is the same as the interior of K, which is a proper subset of $\mathbb{R}^n_{++}$ by assumption. Hence, the interior of $\bar{K}$ is a proper subset of $\mathbb{R}^n_{++}$. Hence, there exists a point $(C_1, \ldots, C_n) \in \mathbb{R}^n_{++}$ that is not in $\bar{K}$.

Next, we apply the separating hyperplane theorem to $\{C\}$, which is closed, convex, and compact, and $\bar{K}$, which is closed and convex. Then there must exist a vector v such that

$$v^\top C < c, \quad v^\top w \geq c$$

for all $w \in \bar{K}$ and some constant c. Since $\bar{K}$ is a cone, the hyperplane can be chosen to pass through the origin, meaning we can set $c = 0$. Thus, there exists a nonzero vector v such that

$$v^\top w \geq 0$$

for all $w \in K$, and meanwhile, $v^\top C = v_1 C_1 + \cdots + v_n C_n < 0$. Since $C_1, \ldots, C_n$ are all positive, there is at least one component of v that is negative, say, $v_{i_0} < 0$.

To finish the proof, let $w \in K$ take the form $w = \mathcal{E}(\eta) = (\mathbb{E}\eta\zeta_1^2, \ldots, \mathbb{E}\eta\zeta_n^2)$ for any η that is bounded away both from 0 and ∞. Then we get

$$v^\top w = v_1 \mathbb{E}\eta\zeta_1^2 + \cdots + v_n \mathbb{E}\eta\zeta_n^2 = \mathbb{E}\left[\eta(v_1\zeta_1^2 + \cdots + v_n\zeta_n^2)\right] \geq 0$$

for all η. However, this is clearly impossible as we can let η be very large when ζ_{i_0} is close to 1 or -1 and otherwise set η to be very small. This yields a contradiction, thereby proving that K must be the entire positive orthant.

□

Proof of Theorem 2. Since the optimization program (2) is convex, we only need to verify that $Q^\star = \tilde{r}\,\mathrm{msgn}(G)$ satisfies the first-order condition

$$G \in \mathbb{E}_\zeta\left[\partial_Q H(\|Q^\star\zeta\|)\right]. \tag{6}$$

Note that for all ζ from the unit sphere, $\|Q^\star\zeta\| = \tilde{r}$ since $m \geq n$ and the columns of $\mathrm{msgn}(G)$ are orthonormal. By the chain rule for subdifferentials,

$$\begin{aligned}
\partial_Q H(\|Q^\star\zeta\|) &= \partial H(\tilde{r})\frac{Q^\star\zeta\zeta^\top}{\|Q^\star\zeta\|} \\
&= \partial H(\tilde{r})\frac{\tilde{r}\,\mathrm{msgn}(G)\zeta\zeta^\top}{\tilde{r}} \\
&= \partial H(\tilde{r})UV^\top\zeta\zeta^\top,
\end{aligned}$$

where $\partial H(\tilde{r})$ can take any value from the subdifferential $[A, B]$.

To satisfy (6), it is sufficient to prove that there exists a random variable $\eta(\zeta)$ taking values in $[A, B]$ almost surely such that

$$\mathbb{E}_\zeta[\eta(\zeta)UV^\top\zeta\zeta^\top] = G = U\Sigma V^\top,$$

which is equivalent to

$$U\left(\mathbb{E}_\zeta[\eta(\zeta)(V^\top\zeta)(V^\top\zeta)^\top]\right)V^\top = U\Sigma V^\top.$$

Since $V^\top\zeta$ has the same distribution as ζ, the proof is complete if we can show there exists such an η satisfying

$$\mathbb{E}_\zeta[\eta(\zeta)\zeta\zeta^\top] = \Sigma.$$

The off-diagonal terms are zero by symmetry. For the diagonal terms, we need $\mathbb{E}[\eta(\zeta)\zeta_i^2] = \sigma_i$ for all i. For a sufficiently small A and large B, the convex set of achievable vectors $\{(\mathbb{E}\,\eta\zeta_1^2, \ldots, \mathbb{E}\,\eta\zeta_n^2) : A \leq \eta \leq B\}$ contains the vector $(\sigma_1, \ldots, \sigma_n)$, a consequence of Lemma 4.2. This completes the proof.

□

Remark 4.3. The practical significance lies in the finiteness of B, which roughly translates to the fact that gradient orthogonalization is optimal without requiring an infinite subdifferential. However, the proof does not specify how large A and B need to be for the kink to satisfy the first-order condition since it is not constructive. Intuitively, it depends on the dimensions and the range of the spectrum of G. It is of practical interest to obtain sharp bounds on A and B for Theorem 2.

Next, we turn to the proof of Proposition 3.4.

Proof of Proposition 3.4. Since $Q^\star = c \cdot \operatorname{msgn}(G)$ is an optimal solution to the convex program, it must satisfy the first-order condition

$$G \in \mathbb{E}_\zeta \left[\partial_Q H(\|Q^\star \zeta\|)\right]. \tag{7}$$

As in the proof of Theorem 2, we have $\|Q^\star \zeta\| = c$ for all ζ. The condition becomes

$$G \in \mathbb{E}_\zeta \left[\partial H(c) U V^\top \zeta \zeta^\top\right].$$

If H is differentiable at c, then the subdifferential $\partial H(c)$ is the singleton $\{H'(c)\}$, and we get

$$G = H'(c) U V^\top \mathbb{E}_\zeta[\zeta \zeta^\top].$$

Since $\mathbb{E}[\zeta \zeta^\top] = \frac{1}{n} I$, this simplifies to

$$U \Sigma V^\top = \frac{H'(c)}{n} U V^\top.$$

This implies $\Sigma = \frac{H'(c)}{n} I$, meaning all singular values of G must be equal. This contradicts the assumption that G is not a scaled orthogonal matrix. Therefore, H cannot be differentiable at c and must have a kink. □

Remark 4.4. Both Theorem 2 and Proposition 3.4 assume $m \geq n$. This condition ensures that $\|Q\zeta\|$ is a constant for all unit vectors ζ when the columns of $Q \in \mathbb{R}^{m \times n}$ are scaled orthonormal. The two results can be extended to the case $m < n$, but the statements might involve approximations and might not be as precise. Roughly speaking, the proof idea should continue to work by recognizing that Q preserves the norm in a subspace of dimension m. Moreover, for sufficiently large m, concentration of measure ensures that the first m components of ζ are approximately sampled from a sphere in $\mathbb{R}^m$. We leave the detailed proofs for this extension to future work.

5 Discussion

In this paper, we investigate how the gradient orthogonalization used in Muon and other related methods contributes to their superior performance in training large-scale deep learning systems such as large language models. Our approach is to propose a convex program, which we call the isotropic curvature model, to understand how optimization proceeds in a single iteration. A key ingredient in the development of this model is the Taylor expansion of the per-sample loss function instead of the total objective, thereby capturing the matrix structure of the weights. Our model shows that under a super-quadratic growth condition on the high-order terms of the loss function, the optimal update matrix shares the same singular spaces as the original gradient matrix, but its

singular values should be more homogeneous. This result immediately clarifies why treating weights as matrices is not only conceptually sound but also technically beneficial. In addition, we identify the condition under which gradient orthogonalization becomes optimal: when the high-order terms exhibit a "kink" where their growth rate sharply increases. This condition can be viewed as an extreme-case approximation, suggesting that while gradient orthogonalization is beneficial, it may not fully exploit the potential of matrix-gradient methods.

At a high level, the rationale behind the isotropic curvature model, supported by empirical evidence, points to a crucial feature of deep learning optimization. Conventionally, first- and second-order information are considered; these can in general take arbitrary values, and preconditioning methods generally require the evaluation of both.[6] However, in deep learning optimization, while first-order information (the gradient) is arbitrary and must be obtained via differentiation, it seems plausible that high-order information possesses a structure that can be leveraged without explicitly computing the Hessian. This may be due to the immense scale of systems like large language models and the isotropic properties underlying their architectures. If this hypothesis holds true, it could pave the way for a large class of *auto-preconditioned* methods. In fact, both Muon and Adam (which can be seen as a smoothed version of signed stochastic gradient descent [6]) are examples of auto-preconditioned methods, as their conditioning of first-order gradients does not explicitly leverage any second-order information. It is likely that more novel methods will emerge from exploring more fine-grained auto-preconditioners.

From a technical perspective, the isotropic curvature model is currently phenomenological, and future research is needed to refine it and unveil its underlying mechanisms. First, rigorously justifying the isotropy assumption would be of great interest, and our current assumptions may require modification to be made mathematically precise. Work characterizing the geometric properties of the optimization landscape for deep learning could provide valuable insights [31, 15, 45]. Another direction is to unveil the mechanism for the super-quadratic growth of high-order terms, which might be illuminated by connecting it to phenomena like the edge of stability [47, 12, 11] and the Hessian structure of neural networks [50, 25]. More broadly, the isotropic curvature model and its future extensions could potentially be used in conjunction with the modular interpretation of optimization [4, 5]. It would also be interesting to investigate whether the model could be extended from a single iteration to multiple iterations for convergence analysis. To better model practice, future research should also consider a noisy gradient G from a mini-batch, since it is empirically observed that the relative performance of Muon compared to Adam depends on batch size [39]. In the same spirit, incorporating momentum via exponential moving average into the model is another important direction.

From a practical standpoint, the isotropic curvature model suggests a new approach to designing optimizers for large language models. This approach would begin by approximating the curvature function H for the target neural network architecture, followed by solving the optimization program of the isotropic curvature model to find the update matrix. The first step requires care, as H may vary across different layers and over the course of training. This variability might explain why the relative empirical performance of Muon compared to Adam varies over problem instances [37, 46]. Given a curvature function H, a key problem is whether the optimal update matrix can be found efficiently using methods suitable for GPU environments, such as the Newton–Schulz iteration. An additional desired property, suggested by Theorem 1, is that the transformation of singular values

[6]Note that there are exceptions. For example, Nesterov's accelerated gradient method extracts certain second-order information from first-order gradients.

be monotone.

Acknowledgments

I am grateful to Tim Lau for introducing me to Muon in early 2025. I would like to thank Xuyang Chen, Alex Damian, Shengtao Guo, Nathan Srebro, and Zihan Zhu for helpful discussions. This research was supported in part by NSF grant DMS-2310679 and Wharton AI for Business.

References

[1] K. Ahn, B. Xu, N. Abreu, Y. Fan, G. Magakyan, P. Sharma, Z. Zhan, and J. Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

[2] N. Amsel, D. Persson, C. Musco, and R. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

[3] K. An, Y. Liu, R. Pan, S. Ma, D. Goldfarb, and T. Zhang. ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

[4] J. Bernstein and L. Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

[5] J. Bernstein and L. Newhouse. Modular duality in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[6] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[7] D. Carlson, V. Cevher, and L. Carin. Stochastic spectral descent for restricted Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[8] D. Carlson, E. Collins, Y.-P. Hsieh, L. Carin, and V. Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[9] D. Chang, Y. Liu, and G. Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

[10] L. Chen, J. Li, and Q. Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

[11] J. Cohen, A. Damian, A. Talwalkar, J. Z. Kolter, and J. D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025.

[12] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[13] M. Crawshaw, C. Modi, M. Liu, and R. M. Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

[14] C. Fan, M. Schmidt, and C. Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

[15] C. Fang, H. He, Q. Long, and W. J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

[16] K. Gruntkowska, Y. Maziane, Z. Qu, and P. Richtárik. Drop-Muon: Update less, converge faster. *arXiv preprint arXiv:2510.02239*, 2025.

[17] V. Gupta, T. Koren, and Y. Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[18] C. He, Z. Deng, and Z. Lu. Low-rank orthogonalization for large-scale matrix optimization with applications to foundation model training. *arXiv preprint arXiv:2509.11983*, 2025.

[19] C. He, S. Ren, J. Mao, and E. G. Larsson. DeMuon: A decentralized Muon for matrix optimization over graphs. *arXiv preprint arXiv:2510.01377*, 2025.

[20] F. Huang, Y. Luo, and S. Chen. LiMuon: Light and fast Muon optimizer for large models. *arXiv preprint arXiv:2509.14562*, 2025.

[21] K. Jordan, Y. Jin, V. Boza, Y. Jiacheng, F. Cecista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[22] A. Khaled, K. Ozkara, T. Yu, M. Hong, and Y. Park. MuonBP: Faster Muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

[23] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[24] D. Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

[25] F. Kunstner, R. Yadav, A. Milligan, M. Schmidt, and A. Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

[26] T. T.-K. Lau, Q. Long, and W. Su. PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

[27] J. Li and M. Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.

[28] Z. Li, L. Liu, C. Liang, W. Chen, and T. Zhao. NorMuon: Making Muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

[29] J. Liu, J. Su, X. Yao, Z. Jiang, G. Lai, Y. Du, Y. Qin, W. Xu, E. Lu, J. Yan, Y. Chen, H. Zheng, Y. Liu, S. Liu, B. Yin, W. He, H. Zhu, Y. Wang, J. Wang, M. Dong, Z. Zhang, Y. Kang, H. Zhang, X. Xu, Y. Zhang, Y. Wu, X. Zhou, and Z. Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[31] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-SGD: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015.

[32] T. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[35] A. Riabinin, E. Shulgin, K. Gruntkowska, and P. Richtárik. Gluon: Making Muon & Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025.

[36] N. Sato, H. Naganuma, and H. Iiduka. Analysis of Muon's convergence and critical batch size. *arXiv preprint arXiv:2507.01598*, 2025.

[37] A. Semenov, M. Pagliardini, and M. Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

[38] M.-E. Sfyraki and J.-K. Wang. Lions and Muons: Optimization via stochastic Frank-Wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

[39] I. Shah, A. M. Polloreno, K. Stratos, P. Monk, A. Chaluvaraju, A. Hojel, A. Ma, A. Thomas, A. Tanwer, D. J. Shah, et al. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

[40] W. Shen, R. Huang, M. Huang, C. Shen, and J. Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

[41] R.-Y. Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

[42] K. Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. Tech report of Kimi K2.

[43] M. Tuddenham, A. Prügel-Bennett, and J. Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022.

[44] N. Vyas, D. Morwani, R. Zhao, I. Shapira, D. Brandfonbrener, L. Janson, and S. Kakade. SOAP: Improving and stabilizing Shampoo using Adam. In *International Conference on Learning Representations (ICLR)*, 2025.

[45] S. Wang, F. Zhang, J. Li, C. Du, C. Du, T. Pang, Z. Yang, M. Hong, and V. Y. Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

[46] K. Wen, D. Hall, T. Ma, and P. Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025.

[47] L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

[48] M. Zhang, Y. Liu, and H. Schaeffer. Adagrad meets Muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

[49] X. Zhang and H. Gao. On provable benefits of Muon in federated learning. *arXiv preprint arXiv:2510.03866*, 2025.

[50] Y. Zhang, C. Chen, T. Ding, Z. Li, R. Sun, and Z.-Q. Luo. Why transformers need Adam: A Hessian perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.