Two Theoretical Spectra of Modern Optimizers: Fisher-Approximated and Norm-Constrained

Yufei Gu

June 8, 2026

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

So many new optimizers are 'under proposal'.
However, where and what is the full picture?

- Extensive empirical studies are essential but not enough; Evaluating optimizer designs is increasingly burdensome for large-scale training scenarios, and conclusion consistency is hard to guarantee.
- Pure theoretical studies are also struggling to provide the correct guidance for complex network structures or the 'infinite' number of factors that influence optimization or generalization.
- We have to carefully choose the interpretations of the empirical success of popular optimizers, and investigate their connection to any universal principles.
- In this note, two key references are revisited and explored to provide a hint on the full picture of modern optimization. [1]
- Several recent optimizers (and their 'trend') will be touched toward the final, with interpretation available from both frameworks.

Key References:

- Chao Ma / MSR Cambridge / Towards Efficient Optimizer Design for LLM via Structured Fisher Approximation with a Low-Rank Extension [1]
- Thomas Pethick / EPFL / Training deep learning models with norm-constrained LMOs [2]

[1] As a brief remark, the update rule of popular optimizers is omitted in this note.

1. Fisher-Approximated Optimization

The Fisher information matrix (FIM) [3, 4] encodes how parameter perturbations affect the model's distribution, which provides:

- Curvature approximation: structured second-order approximation and avoids explicit Hessian computation.
- Geometry-aware metric: measures distance under the KL divergence, which is invariant to reparameterization.

For probabilistic models, the FIM is defined as:

$$F = \mathbb{E}_{x\sim p}\left[(\nabla_{\theta_t} \log p(x \mid \theta_t))(\nabla_{\theta_t} \log p(x \mid \theta_t))^\top\right] \tag{1}$$

Natural Gradient Descent (NGD) rescales gradients by the inverse Fisher for faster training, and as steepest descent under the KL divergence between $p(x \mid \theta_t)$ and $p(x \mid \theta^*)$:

$$\theta_{t+1} \leftarrow \theta_t - \eta F^{-1} L(\theta_t), \tag{2}$$

The **empirical Fisher** $F = \mathbb{E}[\vec{g}\vec{g}^\top]$ is widely approximated by modern optimizers as preconditioners, while **structure** is the key design choice.

- Diagonal, rotated, low-rank, blockwise, or Kronecker-factored; Better structure can capture more correlation, but also increases memory, compute, or implementation complexity.
- This spectrum is therefore a tradeoff between **fidelity to curvature** and **practical efficiency**.

$F^{-1/2}$ is also frequently preferred to F^{-1} under the gradient whitening principle [5] and acts as a softer preconditioner for noisy Fisher.

Diagonal Fisher Estimates: Cheap, Stable, and Popular

Adam and AdamW are commonly read as using a **diagonal empirical Fisher** through element-wise second-moment estimates [6, 7].

- Under the diagonal restriction $\mathcal{H} = \{\mathrm{Diag}(\mathbf{v}) : v_i > 0\}$, the optimal approximation is $\tilde{F} = \mathrm{Diag}(\mathbb{E}[\mathbf{g}^2])$.
- Adam's second-moment accumulator $v_t = \beta_2 v_{t-1} + (1-\beta_2)\mathbf{g}_t^2$ provides an exponential moving estimate of the diagonal Fisher.
- Thus Adam yields the update $\Delta\theta_t = -\eta \frac{m_t}{\sqrt{v_t}+\varepsilon} \approx -\eta \tilde{F}_t^{-1/2} m_t$.

We consider a more simple optimizer, column-normalized SGD, with a much natural Fisher interpretation with **block-diagonal** structure.

- Under the block-isotropic restriction, the Fisher matrix is approximated by assigning a single curvature scale to each column.
- Since $\alpha_j \propto \mathbb{E}[\|g_j\|_2^2]$, the column norm provides an online estimate of the Fisher trace within each column block.
- Thus column-normalized SGD yields $\Delta W_j = -\eta \frac{g_j}{\|g_j\|_2} \approx -\eta \tilde{F}_j^{-1/2} g_j$, corresponding to inverse-square-root preconditioning under a column-wise isotropic Fisher approximation.

The benefit of diagonal Fisher estimates is clear: low complexity, low overhead, and strong robustness in large-scale training.

- The limitation is equally clear: diagonal structure ignores cross-coordinate coupling inside layers or matrices.

Beyond diagonal: Kronecker product structure

Let's consider typical matrix-based optimizers: **Shampoo**.

- Under the Kronecker restriction $\mathcal{H} = \{R_n^{1/2} \otimes L_m^{1/2} : R_n, L_m \succ 0\}$, the Fisher is approximated by a Kronecker-structured metric.
- Shampoo maintains two preconditioners $R_t \approx \frac{1}{m}\mathbb{E}[G^\top G]$ and $L_t \approx \frac{1}{n}\mathbb{E}[GG^\top]$, which estimate the two Kronecker factors of the Fisher geometry [8, 9].
- Thus Shampoo yields $\Delta W_t \approx -\eta(L_t + \varepsilon I)^{-1/4} G_t (R_t + \varepsilon I)^{-1/4}$, which approximates $F^{-1/2}$ preconditioning under a Kronecker-factorized Fisher approximation.

K-FAC is arguably the canonical example of Fisher approximation.

- K-FAC maintains $A_t \approx \mathbb{E}[aa^\top], B_t \approx \mathbb{E}[\delta\delta^\top]$, where a and δ denote the input activations and output gradients, yielding $F_t \approx B_t \otimes A_t$.
- Thus K-FAC yields $\Delta W_t \approx -\eta\, B_t^{-1} G_t A_t^{-1}$, which corresponds to F^{-1} preconditioning instead.

Both Shampoo and K-FAC assume that the Fisher can be approximated by a Kronecker product $F \approx R \otimes L$, reducing the number of degrees of freedom from $O(m^2 n^2)$ to $O(m^2 + n^2)$.

- This structure preserves second-order correlations within the row and column spaces, while discarding cross-mode interactions that cannot be represented by a separable Kronecker form.
- Therefore, Kronecker-based optimizers form an intermediate class between diagonal Fisher approximations and full NGD methods.

Generalize Adam with Eigen-space Rotation

We first consider a block-diagonal matrix with a shared eigen-space U_f:

$$\mathcal{H} = \{\mathrm{Diag}_B(M_1, \ldots, M_n); M_i = U_f D_i U_f^\top\}, \tag{3}$$

where D_i is a positive eigenvalue matrix.[2]

- The general solution can be approximated by

$$U_f^* = EVD(\mathbb{E}[GG^\top]), \quad D^* = \mathrm{Diag}_M(\mathbb{E}[(U_f^{*\top} G)^{\odot 2}]), \tag{4}$$

  where EVD is the eigenvalue decomposition.
- This solution U_f^* leads to an optimizer, called **Eigen-Adam** [1], corresponding to inverse-square-root preconditioning

$$\mathrm{Mat}(\tilde{F}^{-1/2}\vec{g}) = U_f \frac{U_f^\top G}{\sqrt{\mathbb{E}[(U_f^\top G)^{\odot 2}]}}. \tag{5}$$

- The update procedures of Eigen-Adam are summarized as

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) G_t \\ Q_t &= \beta_3 Q_{t-1} + (1-\beta_3) G_t G_t^\top \\ U_{f,t} &= EVD(Q_t) \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)(U_f^\top G)^{\odot 2} \\ \nabla_t &= U_{f,t} \frac{U_{f,t}^\top m_t}{\sqrt{v_t}} \end{aligned} \tag{6}$$

- Eigen-Adam can be viewed as applying Adam's update on a space 'rotated' by eigen-matrix U_f.

The above update procedures closely relates to two related works: AdaDiag [10] and one-sided SOAP [11], which are heuristic memory-efficient variants of the full algorithms AdaDiag++ and SOAP.

[2] Adam is a special case for $U_f = \mathrm{I}$.

SOAP and AdaDiag

Let's checkout **SOAP**, which practically applies Adam-like adaptive scaling in a Shampoo-derived eigenbasis [11].

- SOAP assumes the Fisher is diagonal after a Kronecker-product change of basis, $F \approx (U_R \otimes U_L)\tilde{D}(U_R \otimes U_L)^\top$, for orthonormal matrices U_R and U_L, and positive diagonal matrix $\tilde{D}$.
- Thus SOAP performs $F^{-1/2}$ preconditioning using $\tilde{D}$, combining Kronecker eigenspaces with Adam's diagonal approximation.

AdaDiag++ is a concurrent work to SOAP [10], while disabling the EMA tracking for left and right preconditioners by storing U_R, U_L and performing two eigenvalue decompositions.

- The one-sided version AdaDiag aligned with Eigen-Adam in terms of update rule, which trades performance for memory efficiency, similar to one-sided SOAP.
- SOAP/AdaDiag++ follows a more general structural assumption to Eigen-Adam, with preconditioner minimizing an upper bound on the FIM approximation error.

Table: The underlying structure assumptions of different optimizers with practical efficiency, assuming $n \geq m$. 'Generalizes' refers to the optimizer whose structure is generalized [1].

	Adam	**Shampoo**	**Eigen-Adam/AdaDiag**	**SOAP/AdaDiag++**
Structure	$\mathrm{Diag}_{\mathbf{v}}(\boldsymbol{v})$	$\boldsymbol{R}_n^{\frac{1}{2}} \otimes \boldsymbol{L}_m^{\frac{1}{2}}$	$\mathrm{Diag}_{\mathbf{B}}(\{\boldsymbol{U}_f \boldsymbol{D}_i \boldsymbol{U}_f^T\}_i)$	$(\boldsymbol{U}_R \otimes \boldsymbol{U}_L)\,\tilde{\boldsymbol{D}}\,(\boldsymbol{U}_R \otimes \boldsymbol{U}_L)^T$
Generalizes	N/A	N/A	Adam	Eigen-Adam + Shampoo
Computation	$O(mn)$	$O(m^3 + n^3)$	$O(m^3)$	$O(m^3 + n^3)$
Memory	$3mn$	$mn + m^2 + n^2$	$3mn + 2m^2$	$3mn + 2m^2 + 2n^2$

More General Structures: Normalization and Whitening

We have previously visited column-normalization, with the operator

$$\mathrm{Norm}(G) = \frac{G}{1\sqrt{s^\top}} = GS^{-1/2},$$

where $s_i = \sum_{j=1}^{m} G_{ij}^2$ and $S = \mathrm{Diag}_v(s)$. We next consider gradient whitening, which essentially orthogonalizes G

$$\mathrm{Whitening}(G) = (GG^\top)^{-1/2}G,$$

which compute the closest orthogonal matrix under Frobenius norm.

- Under the row-covariance restriction, the Fisher matrix is approximated by $\tilde{F} \approx \mathbb{E}[GG^\top] \otimes I_n$, with the square-root natural-gradient preconditioner $\tilde{F}^{-1/2} = (\mathbb{E}[GG^\top])^{-1/2} \otimes I_n$.
- Whitening yields update corresponding to this preconditioner

$$\Delta W_t \approx -\eta\,(\mathbb{E}[GG^\top])^{-1/2} G_t \approx -\eta\,(G_t G_t^\top)^{-1/2} G_t.$$

SWAN [12] relies on two processing operators applied to raw current gradient and completely removes the internal states:

$$\mathrm{GradNorm}(G) = \frac{G - \vec{g}1_n^\top}{s1_n^\top},\ \mathrm{GradWhitening}(G) = (GG^\top)^{-1/2}G,$$

where $\vec{g} = \frac{1}{n}\sum_{i=1}^{n} G_{:,i}$ is the mean across rows, and $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G_{:,i} - \vec{g})^2}$ is the standard deviation across rows.

- SWAN derives the update from investigating the LLM dynamics.
- In practice, $(GG^\top)^{-1/2}$ is computed via Newton-Schulz iterations.

Newton-Schulz iteration and Muon

In many machine learning applications, the computation of square-root inverse of some SPD matrix is often encountered. One standard approach is to compute the EVD and take the square-root of the eigenvalue matrix. However, EVD is computationally expensive, which drives the alternative approach using Newton-Schulz iterations (NS),

$$\begin{aligned} Y_0 &= \frac{A}{\|A\|_F}, \quad Z_0 = I \\ Y_{t+1} &= \frac{1}{2} Y_t (3I - Z_t Y_t) \\ Z_{t+1} &= \frac{1}{2} (3I - Z_t Y_t) Z_t, \end{aligned} \tag{7}$$

with convergence $Y_t \to \frac{A^{1/2}}{\sqrt{\|A\|_F}}$ and $Z_t \to A^{1/2} \sqrt{\|A\|_F}$.

Muon performs whitening using NS iteration on the first-order momentum with additional memory states, wheras SWAN relies on normalization applied to the raw gradient.

- Typically, NS converges very fast with only 5 steps.
- The NS coefficients and steps have been tuned in practice for different objectives, including efficiency or accuracy [13, 14].
- Several algorithms have been introduced to replace NS and perform the msign operation, e.g., the PolarExpress [15].

Overall, structured FIM approximation served as a practical framework for optimizer design and characterizes many existing optimizers.

2. Norm-Constrained Optimization: A Geometric Reading

Instead of approximating the curvature of the landscape, choose the geometry in which the descent direction is steepest.

Conditional Gradient (CG): A series of optimization methods leverage the linear minimization oracle (lmo) to adapt to the geometry of the problem under somenorm-ball constraints.

$$x^{k+1} = (1 - \gamma_k)x^k + \gamma_k \mathrm{lmo}(\nabla f(x^k)), \tag{8}$$

with stepsizes $\gamma_k \in (0, 1)$, under the norm constraint $\|x\| \leq p$ for some $p > 0$ and some norm $\|\cdot\|$ (not necessarily the Euclidean norm).

- The linear minimization oracle (lmo) is defined as

$$\mathrm{lmo}(s) \in \arg\min_{x \in \mathcal{D}} \langle s, x \rangle, \tag{9}$$

  which is the feasible point that minimizes the linearized objective.
- Stochastic Conditional Gradient (SCG) methods

For unconstrained problems, let's consider unconstrained Stochastic Conditional Gradient (uSCG) with stepsizes $\gamma_k \in (0, 1)$ [2]

$$x^{k+1} = x^k + \gamma_k \mathrm{lmo}(d^k). \tag{10}$$

- uSCG guarantees a weight norm control of $\|x\| \leq p \sum_{k=1}^{n} \gamma_k$.
- The central object is no longer the Fisher approximation, but the **metric or norm** that defines a prescribed geometry.
- *How can we choose an appropriate norm for deep learning?*

Steepest Descent under different Norms

Table: Special instantiations of uSCG according to different choices of norm.

Method	α_k	Problem	Constraint set D	Normalization
Normalized SGD	1	Unconstrained	Euclidean $\|\cdot\|_2$-ball	$-\rho\frac{d}{\|d\|_2}$
Normalized SGDM	$[0,1]$	Unconstrained	Euclidean $\|\cdot\|_2$-ball	$-\rho\frac{d}{\|d\|_2}$
SignSGD	1	Unconstrained	Max-norm $\|\cdot\|_\infty$-ball	$-\rho\mathrm{sign}(d)$
Signum	$[0,1]$	Unconstrained	Max-norm $\|\cdot\|_\infty$-ball	$-\rho\mathrm{sign}(d)$
Muon [3]	$[0,1]$	Unconstrained	Spectral $\|\cdot\|_{S_\infty}$-ball	$-\rho UV^\top$

Steepest descent in a normed space can be expressed through a lmo

$$x^{k+1} = x^k + \frac{\gamma}{p}\|g^k\|_* \,\mathrm{lmo}(g^k). \tag{11}$$

This formulation connects uSCG to normalized updates of optimizers:

- The Euclidean $\mathrm{lmo}(g) = -\frac{g}{\|g\|_2}$ recovered normalized SGD [16, 17].
- SignSGD and Signum are typically studied under the framework of steepest descent with $\|g^k\|_1$ stepsize scaling [18].
- The LARS optimizer can be viewed as performing normalized SGD with momentum layerwise with the norm choice $\max_l \|W_l\|_F$ [19].
- The Spectral norm ball $\mathrm{lmo}(G) = -UV^\top$, where $G = U\Sigma V^\top$, recovered whitening updates and Muon [13].
- The spectral lmo used in Scion relates to the Preconditioned SGD (PSGD) methods, where the lmo is computed at each iteration [2].

Approximating the 'distance' between models through the KL divergence leads to a steepest descent method, which preconditions with the FIM, known as the **Natural Gradient**.

[3] With non-Nesterov based momentum.

Composite Norms: Multiple Geometric Biases

With the rise of Muon, many optimizers have been proposed with the simple application of composite norms as preconditioning methods.

Composite norms combine multiple normalization operators, inducing a geometry that interpolates between different steepest-descent directions. Recent optimizers can be interpreted through this lens:

- **SinkGD**: Applies Sinkhorn normalization to gradients, approximately projecting onto the set of doubly-stochastic transport maps and introducing both row- and column-wise normalization biases [20]. [4]
- **Muon+**: Combines spectral normalization with additional element-wise or dimension-wise normalization, balancing low-rank directional preference with isotropic scaling [25].
- **NorMuon:** Composes matrix whitening with norm-based scaling, yielding updates that simultaneously control singular-value structure and update magnitude [26].
- **Aurora:** Employs multiple normalization operators acting on complementary tensor dimensions, producing updates that blend spectral, channel-wise, and dimension-wise geometries [27].

However, the existing theoretical framework is struggling to characterize these new methods, and provide effective evaluation, comparison, and performance prediction in large-scale training setups.

- *I believe a valuable future direction is connect complex structural geometric assumptions to specific problem settings, empirical phenomenons, and architectures.*

[4] Many recent works shared a similar investigation to dimension-wise gradient normalization with different formulations and motivations [21–24].

Layer-dependent Norm Constraints: Principles of Architecture-Optimizer Co-design

An important perspective of future optimizer design is the Backbone-Optimizer Coupling Bias (BOCB), where the architectural dependency of optimization is often overlooked [28].

Consider a linear MLP $h_\ell(z) = W_\ell h_{\ell-1}(z) + b_\ell$, $\forall \ell \in \{1, \dots, L\}$, and layers defines as $W_\ell \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. To bound the input to any layer by the RMSNorm $\|\cdot\|_{\text{RMS}\to\text{RMS}}$, three operator norms may be considered:

- Intial layer $h_1(z)$: $\|W_1\|_{\alpha_1\to\text{RMS}} < 1$.
- Intermediary layers $h_\ell(z)$: $\|W_\ell\|_{\text{RMS}\to\text{RMS}} < 1, \forall \ell \in \{2, \dots, L-1\}$.
- Last layer $h_L(z)$: $\|W_L\|_{\text{RMS}\to\beta_L} < 1$.

Table: Example operator norms and lmos of a matrix $A \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$.

	$1 \to$ RMS (ColNorm)	RMS $\to$ RMS (SignNorm)	RMS $\to$ RMS (SpectralNorm)
Norm	$\max_j \frac{1}{\sqrt{d_{\text{out}}}} \|\text{col}_j(A)\|_2$	$\max_{i,j} \|A_{i,j}\|$	$\sqrt{d_{\text{in}}/d_{\text{out}}} \|A\|_{\mathcal{S}_\infty}$
lmo	$\text{col}_j(A) \mapsto -\sqrt{d_{\text{out}}} \frac{\text{col}_j(A)}{\|\text{col}_j(A)\|_2}$	$A \mapsto -\text{sign}(A)$	$A \mapsto -\sqrt{d_{\text{out}}/d_{\text{in}}} UV^\top$

The input norm α_1 and output norm β_L choice dependes on the application, e.g., for the language tasks:

- **Input embeddings:** For the 1-hot encoded vector embedding where $\|z\|_\infty = \|z\|_2 = \|z\|_1 = 1$, we can freely pick the operator norm (ColNorm is favored since exact computation of the lmo). [5]
- **Intermediate layers:** To prevent none of the hidden states blows up, the SpectralNorm is preferred for controlling the RMS norm.
- **LM head:** We can bound the maximal entry through ℓ_∞, which leads to a dimension scaled sign update, i.e., the SignNorm.

[5] While this theoretical implication have certain limitations, AdamW can be considered as an extended version of ColNorm from the FIM approximation perspective as we have discussed, and as suggested by empirical practises of Muon.

From Gradient Normalization to Weight-Norm Control

A stronger intervention is to constrain the weights themselves to some norm-induced hyper-surface, or

Implicit Manifold Optimization:

- **MCSD:** Manifold-Constrained Steepest Descent refers to the framework of algorithms for solving manifold optimization problems with lmo-based update directions under prescribed norms [29].

Explicit Optimizer Wrapper:

- **Hyperball:** Enforces constant weight and update norms, which constrains the optimization trajectory to lie on the surface of a hyper-sphere with radius $\mathcal{R}$ [30].

$$W_{t+1} = \mathcal{R} \cdot \mathrm{Normalize}(W_t - \eta \mathcal{R} \cdot \mathrm{Normalize}(u_t)) \tag{12}$$

- **Spectral Sphere Optimizer (SSO):** Enforces spectral constraints on both weights and updates to a spectral sphere [31], which realizes a fully μP-aligned optimization process [32].
- **Muown:** Decompose the spectral norm into a row-magnitude factor and a row-conherence factor, and explicitly learns the former under a dual norm defined in the ℓ_∞ geometry [33].

However, weight-norm control methods often trade-off peak performance for stability under optimally-tuned hyperparameters, as comparing constrained optimization with the unconstrained solutions. [6]

[6] In our benchmark experiments with a scaling focus.

Promising Future Directions?

Besides the 'never-ending' proposal of new optimizers, What kind of work is necessary and beneficial?

1. **Geometry-aware analytics.** Empirically and theoretically characterize the geometric assumptions induced by different norm constraints across applications, architectures, and training stages.
2. **Unified theory of empirical success.** Refine and consolidate the theoretical interpretation of why existing optimizers work, especially to review, compare, and guide increasingly complex optimizer designs under a common framework.
3. **Scaling-oriented analysis.** Develop analytical tools that predict large-scale training behavior beyond validation loss, since loss can be a misleading proxy for model quality, downstream capability, or post-training performance.

Toward Next-generation Optimizers? *Simple or Complex? Interpretable or Encapsulated? Overfitting or Generalizable?*

- As many promising directions co-exist, what kind of works can eventually make impact beyond optimization?

Thank you for your attention!

References I

[1] Wenbo Gong, Meyer Scetbon, Chao Ma, and Edward Meeds. Towards efficient optimizer design for llm via structured fisher approximation with a low-rank extension. *arXiv preprint arXiv:2502.07752*, 2025.

[2] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

[3] C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

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

[5] Zhirong Yang and Jorma Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 21(2-3):232–240, 2008.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[8] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[9] Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024.

[10] Son Nguyen, Bo Liu, Lizhang Chen, and Qiang Liu. Improving adaptive moment optimization via preconditioner diagonalization. *arXiv preprint arXiv:2502.07488*, 2025.

[11] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

[12] Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Sgd with normalization and whitening enables stateless llm training. *arXiv preprint arXiv:2412.13148*, 2024.

References II

[13] Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[14] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[15] Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

[16] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28, 2015.

[17] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[18] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018.

[19] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[20] Meyer Scetbon, Chao Ma, Wenbo Gong, and Edward Meeds. Gradient multi-normalization for stateless and scalable llm training. *arXiv preprint arXiv:2502.06742*, 2025.

[21] Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for llm training. *arXiv preprint arXiv:2601.23000*, 2026.

[22] Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms i: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

[23] Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

[24] Jinghui Yuan, Jiaxuan Zou, Shuo Wang, Yong Liu, and Feiping Nie. Nora: Normalized orthogonal row alignment for scalable matrix optimizer. *arXiv preprint arXiv:2605.03769*, 2026.

References III

[25] Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, and Zheng Zhang. Muon+: Towards better muon via one additional normalization step. *arXiv preprint arXiv:2602.21545*, 2026.

[26] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

[27] Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware optimizer for rectangular matrices. 2026.

[28] Siyuan Li, Juanxi Tian, Zedong Wang, Luyuan Zhang, Zicheng Liu, Weiyang Jin, Yang Liu, Baigui Sun, and Stan Z Li. Unveiling the backbone-optimizer coupling bias in visual representation learning. *arXiv preprint arXiv:2410.06373*, 2024.

[29] Kaiwei Yang and Lexiao Lai. Manifold constrained steepest descent. *arXiv preprint arXiv:2601.21487*, 2026.

[30] Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them ii: From weight decay to hyperball optimization, 11 2025.

[31] Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, et al. Controlled llm training on spectral sphere. *arXiv preprint arXiv:2601.08393*, 2026.

[32] Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

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