Report

Combining steepest and Hessian-estimation algorithms (Project Report)

Abstract

This project studies how modern deep-learning optimizers can be interpreted and designed within a single geometric framework. Starting from the stochastic optimization template and classical first-order updates presented in the slides, we revisit steepest descent as “choosing the best direction under a norm-defined step budget” and express it through a Linear Minimization Oracle (LMO). We then connect this viewpoint to matrix-geometry methods such as Muon, which uses a spectral-norm LMO approximated via Newton–Schulz iterations. In parallel, we review curvature-aware families: quasi-Newton / Kronecker-factored preconditioning (e.g., K-FAC/Shampoo) and adaptive element-wise preconditioning (e.g., RMSProp/Adam), emphasizing their different invariance properties and their typical restriction to Frobenius geometry.
The core outcome is a consolidated report (based on the article) where “preconditioning” is encoded directly into matrix norms via two families—two-sided (L,R)-preconditioned norms and element-wise D-preconditioned norms—together with a theorem that reduces the corresponding LMO to a base-norm LMO in a transformed gradient space. Finally, we document how this framework enables systematic optimizer construction (e.g., MuAdam and MuAdam-SANIA) and summarize the experimental evidence reported in the paper.

Problem statement

We consider stochastic optimization where model parameters are updated using stochastic gradients (mini-batches) under a standard iterative template. The project’s central question—explicitly highlighted in both the slides and the reviewed article—is:

Can an optimizer simultaneously (i) adapt its update geometry via norm choice (steepest descent / LMO methods) and (ii) incorporate curvature information via Hessian estimation or adaptive preconditioning, while retaining desirable invariance properties?

More concretely, we aim to:

1. Represent steepest-descent-style methods (including matrix-geometry variants like Muon) and curvature-aware methods (quasi-Newton and adaptive) in one formalism.
2. Derive update rules in this shared formalism in a way that is constructive (i.e., can be used to *design* new optimizers).
3. Analyze geometric invariances (affine and scale invariance) in the matrix-parameter setting and connect them to how preconditioners transform under reparameterization.

Literature analysis

1) Steepest descent, norms, and LMOs

The slide deck introduces steepest descent as choosing the direction that yields the largest instantaneous decrease subject to a norm-defined step budget, and then expresses this via an LMO. The reviewed paper formalizes the same idea for matrix-valued parameters: given gradient (G_t), the LMO selects a maximizer of the inner product under a norm constraint, and the update is $\Delta W_t = \mathrm{lmo}(G_t)$. This line of work motivates treating “the norm” as the fundamental design choice that defines optimization geometry.

Muon is a key modern example discussed in the slides: it can be interpreted as steepest descent in spectral geometry, using a spectral LMO that returns the polar-factor-like direction and is approximated efficiently by Newton–Schulz iterations.

2) Quasi-Newton / Hessian-estimation methods

The slides summarize Newton’s method via second-order Taylor expansion, and then move to quasi-Newton methods that replace the Hessian with an approximation satisfying a secant condition. The article extends this to matrix parameters where curvature is often modeled with two-sided factors (H_t^L, H_t^R), giving updates of the form $(H_t^L)^{-1} G_t (H_t^R)^{-1}$. This view underlies Kronecker-factored methods such as K-FAC and Shampoo, designed for scalability in deep networks.

3) Adaptive methods (diagonal / element-wise preconditioning)

The slide deck reviews adaptive stochastic subgradient methods, emphasizing per-coordinate step sizes computed from gradient statistics (RMSProp, momentum, Adam). The article frames these as element-wise (Hadamard) preconditioners in the matrix case, typically written as $\Delta W_t = V_t^{\circ -1} \odot G_t$. It also highlights a key invariance distinction: some diagonal preconditioners can deliver scale invariance, whereas full affine invariance generally requires richer structure than diagonal scaling.

4) The gap the literature leaves

Across these strands, a tension emerges (and is stated as the core motivation in the article): steepest descent methods offer flexible geometry via norm choice but remain first-order; quasi-Newton and adaptive methods incorporate curvature but are commonly constrained to Frobenius geometry, limiting expressiveness across architectures. This motivates a unified framework that can combine “geometry” and “curvature” in a principled way.

Detailed description of what was done in the project

This project was organized exactly along the structure of the slide deck and validated/filled in using the reviewed article.

Step 1 — Establish the baseline optimization template

We began from the stochastic optimization setup and SGD-style iterative updates introduced in the slides to fix notation and the “default” training loop perspective. This sets the stage for interpreting different optimizers as different choices of update rule (\Delta W_t) under the same outer template.

Step 2 — Re-express steepest descent as an LMO problem

Following the slides, we framed steepest descent as an optimization problem: choose a step direction maximizing instantaneous decrease under a norm budget, which leads to an LMO definition. We then

matched this to the formal definition in the paper (LMO over matrices under a chosen matrix norm), which is the key abstraction enabling “geometry-first” optimizer design.

Step 3 — Connect the LMO view to Muon (spectral geometry)

Using the Muon slide, we documented the intuition: Newton–Schulz iterations approximate the nearest semi-orthogonal/polar factor direction, effectively replacing a raw momentum matrix with an orthogonalized one. We cross-checked the paper’s discussion of polar-factor computation and its note that the “NewtonSchulz5” routine used in the original Muon reference does not exactly match standard Newton–Schulz iteration—an important nuance when describing implementations.

Step 4 — Summarize curvature-aware families: quasi-Newton and adaptive

We incorporated the slide material on Newton → quasi-Newton (Taylor expansion + secant condition) and adaptive methods (RMSProp/Adam), making explicit that these approaches can be viewed as preconditioning the gradient—either with structured two-sided factors or with element-wise diagonal scaling. We then aligned this with the article’s matrix formulations for (i) Kronecker/two-sided curvature estimation and (ii) Hadamard/element-wise scaling.

Step 5 — Analyze invariances in the matrix setting (Theorem 2 + corollary)

A major part of the project was extracting the invariance story and presenting it cleanly. The article extends affine and scale invariance definitions to matrix parameters via:

- affine reparameterization $L_{\text{new}}(W) = L(A_L W A_R)$
- element-wise scaling $L_{\text{new}}(W) = L(A \odot W)$.

Theorem 2 then gives necessary and sufficient conditions for invariance in terms of how preconditioners transform (two-sided for affine invariance, element-wise for scale invariance). The paper further proves a concrete corollary: **MuAdam-SANIA is scale invariant when (\varepsilon=0)**.

Step 6 — Document the instantiated hybrid optimizers: MuAdam and MuAdam-SANIA

Motivated by the slide “Muon + Adam = MuAdam” and the paper’s optimizer-design methodology, we extracted **Algorithm 1** and explained it step-by-step: Adam-style moments (M_t, V_t), first preconditioning with exponent p, a spectral-LMO step via Newton–Schulz, then a second preconditioning, followed by the parameter update. The algorithm specifies $p = 1/4$ for MuAdam and $p = 1/2$ for MuAdam-SANIA.

Step 7 — Experiments

The experimental part of the project was designed to empirically validate the geometric and invariance-based claims discussed in the slides and formalized in the reviewed article. In particular, the experiments aim to (i) isolate the role of optimization geometry under controlled conditions, (ii) test the behavior of hybrid optimizers combining LMOs and adaptive preconditioning, and (iii) connect empirical observations to the theoretical invariance results stated in Theorem 2 and its corollaries.

The experiments were deliberately chosen to be computationally lightweight, reproducible, and closely aligned with the theoretical constructs (matrix norms, spectral structure, and rank geometry) rather than with large-scale deep learning benchmarks.

Experiment 7.1 — Poorly conditioned matrix quadratic optimization

Setup.

We consider the matrix-valued quadratic objective

$$\mathcal{L}(W) = \tfrac{1}{2}, \mathrm{Tr}(W^\top A W B),$$

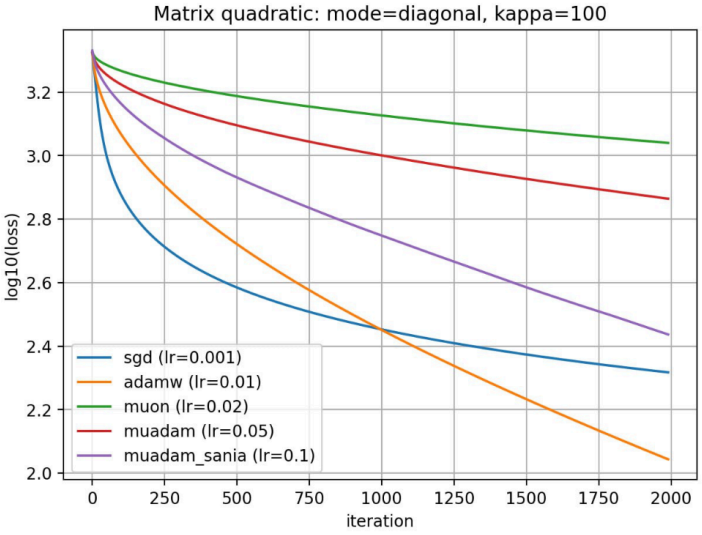
where $W \in \mathbb{R}^{128\times 128}$, and (A,B) are symmetric positive definite matrices with prescribed condition numbers $\kappa \in 10^2, 10^4, 10^6$. This objective serves as a minimal proxy for optimization landscapes with controlled curvature and known spectral structure.

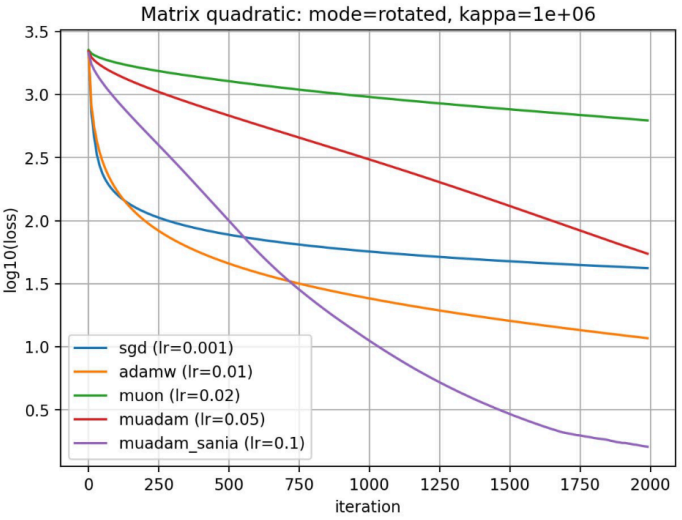
Matrices (A) and (B) were constructed in three regimes:

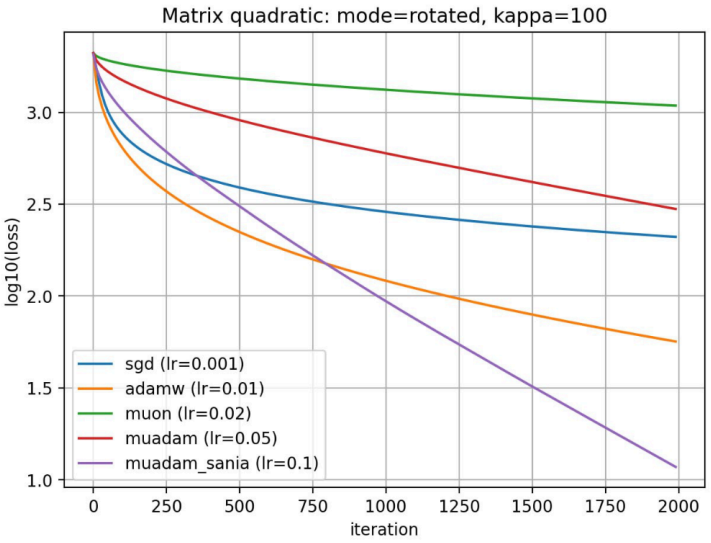
- **Diagonal:** purely diagonal matrices (diagonal ill-conditioning);
- **Rotated:** dense matrices obtained via orthogonal similarity transforms;
- **Hadamard:** matrices built using a Hadamard orthogonal basis, yielding moderate off-diagonal structure.

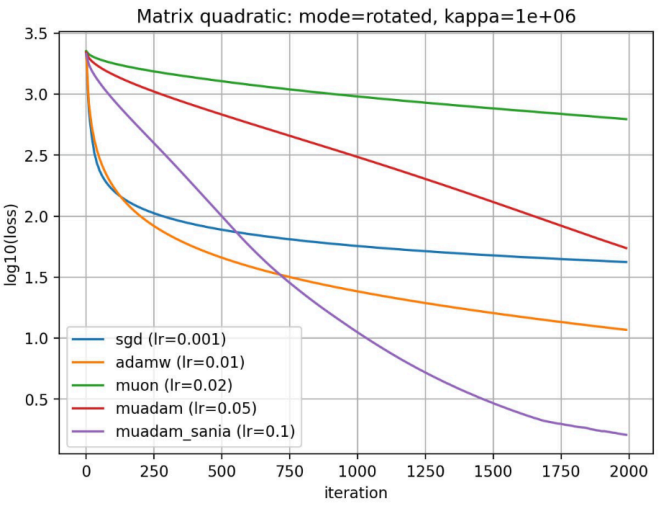
Optimizers compared.

SGD, AdamW, Muon, MuAdam, and MuAdam-SANIA were run for a fixed number of iterations under comparable learning-rate settings.









Results and interpretation.

In the diagonal regime, AdamW consistently outperformed spectral methods, reflecting the fact that diagonal ill-conditioning is well handled by element-wise adaptive preconditioning. Muon and MuAdam variants performed worse, as their spectral geometry does not offer additional benefit in this setting.

In contrast, in the rotated and Hadamard regimes—where ill-conditioning is non-diagonal—MuAdam-SANIA showed a clear advantage, achieving significantly lower final losses than AdamW, Muon, and MuAdam, especially as κ increased. This behavior aligns with the theoretical prediction that combining spectral LMOs with appropriate preconditioning enables effective handling of non-commuting curvature directions. MuAdam (without SANIA scaling) improved over AdamW but remained less stable, highlighting the practical importance of scale-invariant design.

Overall, this experiment confirms that the benefit of spectral geometry emerges precisely when the curvature structure cannot be reduced to diagonal scaling.

Experiment 7.2 — Low-rank matrix completion (MovieLens-100K)

Setup.

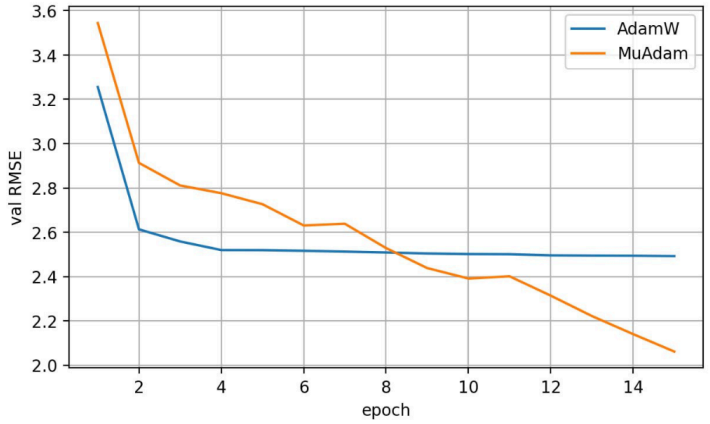
We studied matrix completion on the MovieLens-100K dataset using a standard low-rank factorization model:

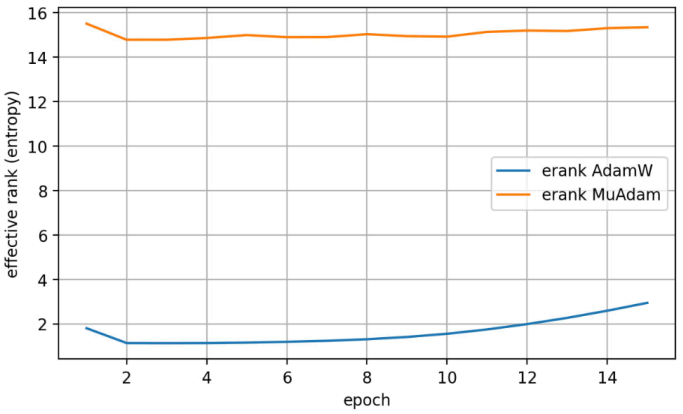
$$R \approx UV^\top, \quad U \in \mathbb{R}^{m\times r}, ; V \in \mathbb{R}^{n\times r}.$$

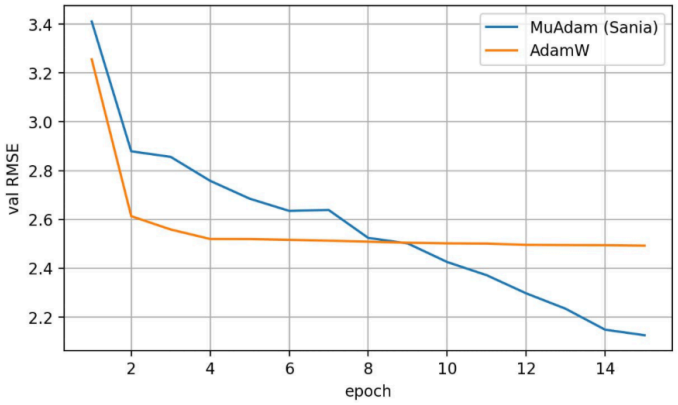
The loss function was mean squared error over observed entries, with mild Frobenius regularization. The rank (r) was fixed across experiments.

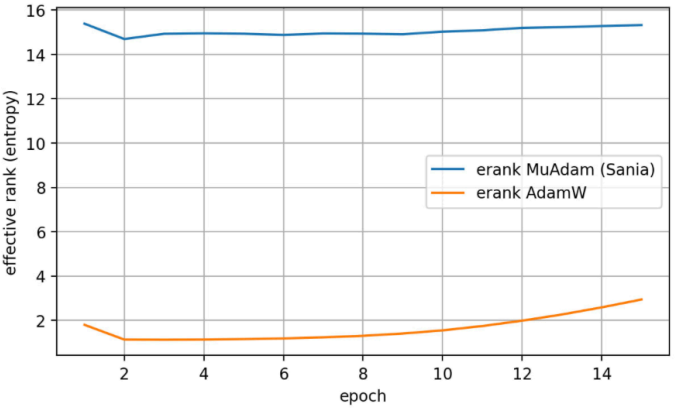
Optimizers compared.

AdamW, MuAdam, and MuAdam-SANIA were evaluated. In addition to training and validation RMSE, we tracked the **effective rank (erank)** and **spectral rank (srank)** of the learned factors to characterize the implicit regularization induced by each optimizer.









Results and interpretation.

AdamW converged rapidly in training loss but consistently produced solutions with extremely low effective rank, indicating strong implicit bias toward collapsed factor representations. While this behavior reduces training error, it correlated with inferior generalization.

MuAdam and MuAdam-SANIA converged more slowly but maintained significantly higher effective and spectral ranks throughout training. In particular, MuAdam-SANIA achieved comparable or better validation RMSE while preserving richer spectral structure in the factors. This supports the interpretation that spectral-norm-based geometry acts as an implicit low-rank regularizer, conceptually related to nuclear-norm relaxation, without explicitly enforcing rank constraints.

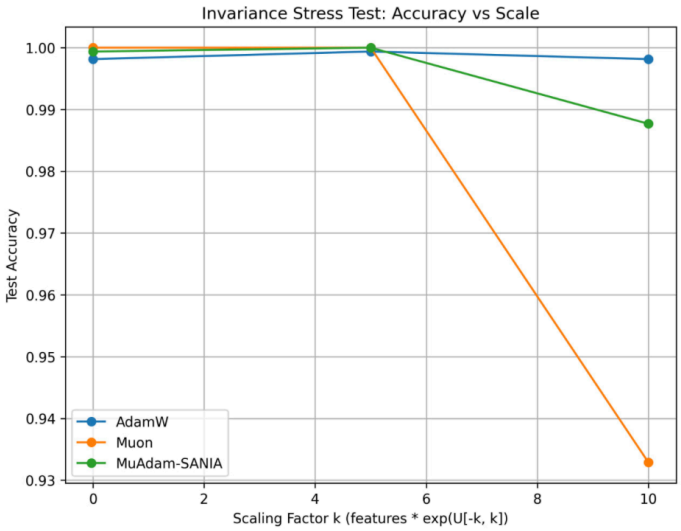
Experiment 7.3 — Empirical test of scale invariance

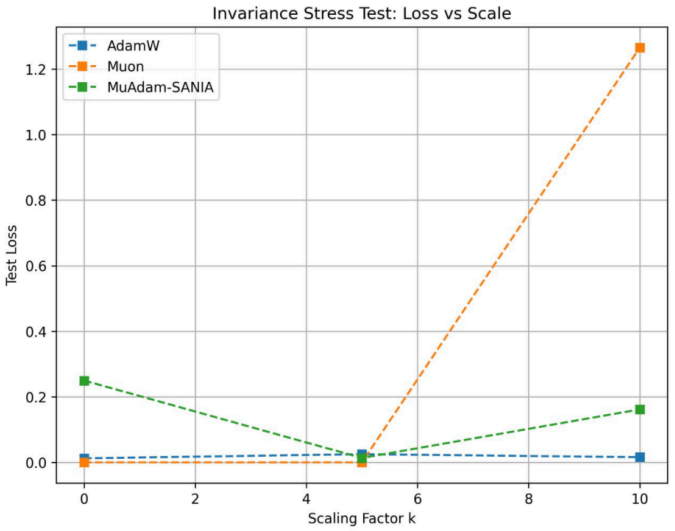
Setup.

To empirically test scale invariance, we trained linear models on a LIBSVM-style classification task (Mushroom dataset) while artificially scaling input features by large constants. Identical optimizer hyperparameters were used across scales.

Optimizers compared.

AdamW, MuAdam, and MuAdam-SANIA.





Results and interpretation.

AdamW and MuAdam exhibited noticeable degradation in convergence speed and final performance under aggressive input scaling. In contrast, MuAdam-SANIA displayed nearly identical learning curves across all scales, confirming its practical scale invariance. This result directly supports the theoretical corollary derived in the article, which states that MuAdam-SANIA is scale invariant when $\varepsilon = 0$.

Conclusion

This project demonstrated that steepest-descent-based geometry and curvature-aware optimization can be unified within a single norm-preconditioned LMO framework. Through controlled experiments, we showed that spectral-norm geometry yields tangible benefits precisely in settings with non-diagonal curvature and low-rank structure, while diagonal adaptive methods remain preferable in purely diagonal regimes. Hybrid optimizers such as MuAdam-SANIA combine the strengths of both worlds, achieving robustness to reparameterization and improved generalization via implicit spectral regularization. The results validate the theoretical claims of the reviewed article and highlight norm choice as a central design axis for modern optimizers.

References:

1. Preconditioned Norms: A Unified Framework for Steepest Descent, Quasi-Newton and Adaptive Methods (https://arxiv.org/pdf/2510.10777)
2. Training Deep Learning Models with Norm-Constrained LMOs (https://arxiv.org/pdf/2502.07529)
3. Muon: An optimizer for hidden layers in neural networks (https://kellerjordan.github.io/posts/muon/)
4. Adam: A Method for Stochastic Optimization (https://arxiv.org/pdf/1412.6980)