Tensorion: A Tensor-Aware Generalization of the Muon Optimizer

Vladimir Bogachev [1] **Vladimir Aletov** [1 2] **Alexander Molozhavenko** [1] **Sergei Kudriashov** [1] **Maxim Rakhuba** [1]

Abstract

Common first-order optimizers, such as Adam, implicitly treat each parameter block as an unstructured vector, which disregards the multilinear weight structure present in many modern machine learning models. Recent work has shown that exploiting matrix structure can improve optimization dynamics. A notable example is Muon, which performs steepest descent under the spectral norm constraint. We take the next step and introduce Tensorion, a tensor-aware optimizer that extends Muon's constrained optimization perspective from matrices to higher-order tensors. Tensorion is built around a linear minimization oracle (LMO) over a tensor norm ball. The norm is carefully chosen to balance two objectives: tightly bounding the tensor spectral norm, while still keeping the LMO tractable. This LMO becomes computable because it reduces to operations on adaptively selected unfolding matrices. Notably, when restricted to order-2 tensors (i.e., matrices), Tensorion recovers Muon exactly. Experiments on tensor-based computer vision problems suggest that Tensorion can offer improved convergence behavior and more stable gradient updates compared with Adam-based and existing tensor-aware baselines in the evaluated settings.

1. Introduction

The field of matrix optimization is emerging as a promising paradigm for the development of next-generation, efficient training algorithms for deep neural networks. In contrast to well-established methods like Adam (Kingma and Ba, 2014), which treat weights as a flat collection of numbers (parameter vectors), matrix-based approaches preserve and leverage the natural two-dimensional structure of the weight matrices. Among these, Muon (Jordan et al., 2024) has proven to be a particularly efficient method, achieving record-breaking performance on smaller models such as NanoGPT and is recently proven scalable to larger architectures (Liu et al., 2025; Team et al., 2026).

However, many modern machine learning models are inherently tensor-valued. Flattening tensors into matrices destroys the multilinear structure and mixes independent modes, which omits information crucial for optimization. Convolutional kernels, Multi-Head Attention (MHA) (Devlin et al., 2019), Mixture-of-Experts layers (MoE) (Oldfield et al., 2024), and other architectural components allow for a natural characterization as multilinear operators, *i.e.*, higher-order tensors.

In this work, we aim to extend the success of Muon to tensor-valued parameters. This setting naturally encompasses weights in neural networks with a convolution-based architecture and also arises in tensorization, where lower-order objects (such as vectors) are reshaped into higher-dimensional tensors (Novikov et al., 2015).

To illustrate our approach and its connection to prior work, we consider the Linear Minimization Oracle (LMO) framework (Bernstein, 2025a). Given an ascent direction, for instance the gradient $\nabla\mathcal{L}(W)$ of the loss function $\mathcal{L}(\cdot)$, one selects the search direction by solving

$$\max_{X\,:\,\|X\|\leq 1} \langle \nabla\mathcal{L}\,(W), X\rangle. \tag{1}$$

The specific norm $\|\cdot\|$ chosen here essentially defines the geometry of the problem, and leads to completely different optimization trajectories (Pethick et al., 2025). If we take the Euclidean (Frobenius) norm, the solution recovers the standard gradient descent. This is unsurprising, as the Frobenius norm treats a matrix as a flat vector and ignores its inherent structure. In contrast, Muon employs the matrix spectral norm, which preserves the matrix structure and leads to an (approximate) orthogonalization of the gradient $\nabla\mathcal{L}(W)$ at each step. Other methods like signSGD (Bernstein et al., 2018) over infinite norm balls also admit this type of interpretation.

A natural step in extending this approach is to employ a norm specifically designed for tensors. However, unlike matrices, higher-order tensors exhibit fundamentally different properties, which significantly complicates such a generalization. In particular, one cannot straightforwardly use the

[1]HSE University [2]Basic Research of Artificial Intelligence Laboratory (BRAIn Lab). Correspondence to: Vladimir Bogachev <vabogachev@hse.ru>.

Preprint. June 9, 2026.

well-known tensor generalization of the spectral norm.

Our contributions are as follows:

1. We show how to implicitly define a tensor norm that ensures, on the one hand, serves as a tight upper bound to the spectral tensor norm (see Sec. 4.1), on the other hand, a tractable solution to the LMO problem (see Sec. 4.2). In a nutshell, using our norm results in adaptively choosing unfolding matrices of a tensor depending on their nuclear matrix norms. We call the resulting optimizer "Tensorion".

2. Based on theoretical results for randomly generated tensors (see Sec. 4.3) and an extensive evaluation of different unfolding strategies (see Sec. 5.1), we develop the best practical strategy for choosing the unfolding approach without computing nuclear norms at every iteration.

3. For convolutional layers, we show (see Sec. 4.5) how the proposed method relates to the LMO problem for the actual Jacobian of the convolution operation (similarly to how W is a Jacobian of $y = Wx + b$ for fully-connected layers).

4. We evaluate the proposed optimizer on multiple CV classification datasets across CNN and transformer-based architectures (see Sec. 5) and observe consistent improvements over conventional SGD and Adam-based optimizers.

2. Related work

Tensorion is related to matrix-aware and tensor-aware optimization methods. Muon introduced matrix orthogonalization into optimization (Jordan et al., 2024), with later work improving its training recipe (Liu et al., 2025), connecting it to spectral constraints (Chen et al., 2025), and extending it through manifold constraints (Bernstein, 2025b), fixed-rank fine-tuning (Bogachev et al., 2026), distributed orthonormalized updates (Ahn et al., 2025), and general norm-based LMOs (Riabinin et al., 2025). A parallel line of structured optimizers uses matrix or tensor structure for preconditioning: K-FAC uses Kronecker-factored curvature (Martens and Grosse, 2015), with extensions to convolutional layers (Grosse and Martens, 2016) and transformers (Eschenhagen et al., 2023); Shampoo maintains mode-wise tensor preconditioners (Gupta et al., 2018), with further analysis in (Morwani et al., 2024); SOAP combines Shampoo eigenspaces with Adam-like updates (Vyas et al., 2024); and (Bernstein and Newhouse, 2024) give an operator-norm view. In contrast to these preconditioning methods, Tensorion does not maintain auxiliary matrices and instead acts directly through tensor unfoldings.

Tensor methods have also been used to compress or structure neural network parameters, including tensor-structured filters (Rigamonti et al., 2013), Tensor Train layers (Novikov et al., 2015), stable tensor decompositions (Phan et al., 2020), and Tucker-style parameterizations (Peshekhonov et al., 2024). Related tensor structures have been applied to multi-head attention (Gu et al., 2025; Li et al., 2026), mixture-of-experts models (XU et al., 2026), fixed-rank tensor optimization (Mo et al., 2025). These approaches typically modify the model parameterization, whereas Tensorion preserves the original parameters and uses tensor structure only in the optimizer. Finally, unlike (Zhang et al., 2026), which stacks layerwise gradients into auxiliary tensors, Tensorion operates directly on each layer's tensor parameters.

3. Preliminaries

3.1. Norms, dual norms, and LMOs

The LMO in Eq. (1) is naturally described through dual norms. For a norm $\|\cdot\|$ on a finite-dimensional inner-product space, its dual norm is

$$\|M\|^\dagger := \max_{\|X\|\le 1} \langle M, X\rangle. \tag{2}$$

Consequently, the maximizers of the linear oracle over the unit ball are exactly the subgradients of the dual norm:

$$\begin{aligned}&\operatorname*{Argmax}_{\|X\|\le 1}\langle M, X\rangle = \partial\|M\|^\dagger = \\ &\left\{Y \;\middle|\; \|Y\| \le 1,\ \langle M, Y\rangle = \|M\|^\dagger\right\},\end{aligned} \tag{3}$$

a standard consequence of convex duality (Rockafellar, 1997). The corresponding minimization oracle is obtained by changing the sign.

For Muon, the relevant primal norm is the matrix spectral norm, $\|M\|_2 = \sigma_1(M)$, whose dual is the nuclear norm

$$\|M\|_2^\dagger = \|M\|_* = \sum_{i=1}^{r} \sigma_i(M), \tag{4}$$

where $r = \operatorname{rank}(M)$ and $\sigma_1(M) \ge \cdots \ge \sigma_r(M) > 0$ are the nonzero singular values.

Let $M = U\Sigma V^\top$ be a compact SVD. Then $UV^\top$ is a maximizer since $\langle M, UV^\top\rangle = \operatorname{Tr}(\Sigma) = \|M\|_*$. Thus, $UV^\top \in \partial\|M\|_*$, while $-UV^\top$ solves the corresponding minimization oracle. This dual-norm viewpoint is the basic mechanism behind the Muon update and will be the starting point for our tensor extension. For completeness, for a convex function $f : V \to \mathbb{R}$, the subdifferential at x is

$$\begin{aligned}&\partial f(x) := \\ &\quad \{g \in V \mid f(y) \ge f(x) + \langle g, y - x\rangle,\ \forall y \in V\}.\end{aligned} \tag{5}$$

3.2. The tensor case

Let $X, Y \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be d-th order tensors. We use the Euclidean tensor inner product

$$\langle X, Y \rangle := \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} X_{i_1,\ldots,i_d} Y_{i_1,\ldots,i_d}, \tag{6}$$

and the associated Frobenius norm $\|X\|_F := \sqrt{\langle X, X \rangle}$. The matrix spectral norm admits the variational form:

$$\|X\|_2 = \max_{\|u\|_2=\|v\|_2=1} |u^\top X v| = \max_{\|u\|_2=\|v\|_2=1} |\langle X, uv^\top \rangle|. \tag{7}$$

Then its multilinear analogue is a direct generalization of a matrix case:

$$\|X\|_\sigma := \sup_{\substack{u^{(k)} \in \mathbb{R}^{n_k},\ \|u^{(k)}\|_2=1 \\ k=1,\ldots,d}} \left| \left\langle X,\ u^{(1)} \otimes \cdots \otimes u^{(d)} \right\rangle \right|, \tag{8}$$

where $\left(u^{(1)} \otimes \cdots \otimes u^{(d)}\right)_{i_1,\ldots,i_d} = u^{(1)}_{i_1} \cdots u^{(d)}_{i_d}$. For $d = 2$, this definition reduces to the usual matrix spectral norm. For $d \geq 3$, however, the tensor spectral norm and its dual are substantially harder to compute and analyze (Friedland and Lim, 2018). This motivates the tractable relaxation developed in Section 4.

We will frequently use tensor unfoldings. For $\tau \subset \{1, \ldots, d\}$, let τ^c denote its complement and define

$$N_\tau := \prod_{k \in \tau} n_k, \qquad N_{\tau^c} := \prod_{k \notin \tau} n_k, \tag{9}$$

with the convention that an empty product equals one. The τ-unfolding of X, denoted by $X_{(\tau)}$, is the matrix obtained by grouping the modes in τ into rows and the remaining modes into columns:

$$\left(X_{(\tau)}\right)_{\mathbf{j}_\tau, \mathbf{j}_{\tau^c}} := X_{j_1,\ldots,j_d}, \qquad X_{(\tau)} \in \mathbb{R}^{N_\tau \times N_{\tau^c}}, \tag{10}$$

where $\mathbf{j}_\tau$ and $\mathbf{j}_{\tau^c}$ are the corresponding lexicographically ordered multi-indices. The standard mode-k matricization is the special case

$$X_{(k)} := X_{(\{k\})}. \tag{11}$$

4. Tensorion

This section introduces *Tensorion*, our tensor-valued extension of Muon. The central difficulty is that the natural LMO over the tensor spectral-norm ball is computationally intractable. Tensorion addresses this obstacle by replacing the tensor spectral norm with a duality-based relaxation that is both tighter than any unfolding spectral norm and computable.

4.1. Relaxed tensor spectral norm

The tensor spectral norm in Eq. (8) is the natural analogue of the matrix spectral norm and is therefore the most direct candidate for extending Muon beyond matrices. However, computing the tensor spectral norm, and hence solving the corresponding exact LMO, is NP-hard in general (Hillar and Lim, 2009). This makes the direct tensor-spectral formulation impractical for modern tensor-valued neural-network parameters.

A standard relaxation is obtained by upper-bounding the tensor spectral norm with the spectral norm of a matrix unfolding. Namely, for any $\tau \subset \{1, \ldots, d\}$, it is well known (Wang et al., 2017) that

$$\|X\|_\sigma \leq \|X_{(\tau)}\|_2. \tag{12}$$

This observation motivates unfolding-based heuristics, including the empirical strategy used in (Jordan et al., 2024) for convolutional layers. However, selecting a single unfolding hard-codes a particular tensor geometry. Different layers, and even different tensor shapes within the same architecture, may favor different groupings of modes. Thus, the choice of unfolding becomes an additional design decision rather than a property of the optimization problem itself.

A natural attempt to remove this ambiguity is to aggregate several candidate unfoldings. Given a family of mode subsets $\mathcal{T}$, for example $\mathcal{T} = \{\{1\}, \{1, 2\}\}$ or $\mathcal{T} = 2^{\{1,2,\ldots,d\}}$, one may consider

$$\min_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_2 \quad \text{or} \quad \max_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_2. \tag{13}$$

The first construction is generally not a norm: the minimum of norms need not satisfy the triangle inequality. The second construction is a norm, but it is often too conservative. In particular, it can collapse to the Frobenius norm in the presence of small or singleton modes, which are common in convolutional kernels.

To see this, consider the extreme case $X \in \mathbb{R}^{1 \times n_2 \times \cdots \times n_d}$ and suppose that $\{1\} \in \mathcal{T}$. Then

$$\max_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_2 = \|X_{(\{1\})}\|_2 = \|\operatorname{vec}(X)^\top\|_2 = \|X\|_F, \tag{14}$$

where $\operatorname{vec}(\cdot)$ denotes column-major vectorization. Hence, the relaxation becomes the Frobenius norm, which may substantially overestimate the tensor spectral norm and loses sensitivity to the multilinear structure of X. Moreover, this behavior prevents the relaxation from recovering the matrix Muon update on tensors of size $1 \times m \times n$, which should effectively behave as matrices.

Tensorion avoids this pathology by moving the relaxation to the dual side. Instead of aggregating primal unfolding spectral norms, we construct a lower bound on the dual tensor

spectral norm using unfolding nuclear norms. Specifically, we define:

$$\|X\|_{\Sigma}^{\dagger} := \max_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_{*}. \tag{15}$$

The relaxed tensor spectral norm is then defined by duality:

$$\|X\|_{\Sigma} := \left(\|X\|_{\Sigma}^{\dagger}\right)^{\dagger}. \tag{16}$$

This dual construction is the key relaxation underlying Tensorion. It preserves the desirable upper-bound property required for a spectral-norm relaxation, while avoiding the overly conservative behavior of the maximum over primal unfolding norms. Supplementary Material D provides an empirical analysis of relaxation tightness along optimization trajectories.

The following proposition formalizes the relationship between the proposed norm, the tensor spectral norm, and unfolding-based spectral relaxations. In particular, Tensorion always upper-bounds the tensor spectral norm, yet it is no looser than the best spectral norm among the selected unfoldings.

Proposition 4.1. *Let $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be any tensor. Then*

$$\|X\|_{\Sigma}^{\dagger} \leq \|X\|_{\sigma}^{\dagger}, \tag{17}$$

and consequently

$$\|X\|_{\sigma} \leq \|X\|_{\Sigma} \leq \min_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_{2}. \tag{18}$$

Proof is in Supplementary Materials, Section A, Proposition A.3.

4.2. Relaxed LMO solution

Similar in spirit to the Muon optimizer, our method solves a LMO over the $\|\cdot\|_{\Sigma}$ at each iteration. Given a tensor M (*e.g.*, the gradient or a momentum estimate at some iteration), we compute

$$X^{\text{opt}} \in \operatorname*{Arg\,max}_{X\,:\,\|X\|_{\Sigma} \leq 1} \langle M, X \rangle. \tag{19}$$

Proposition 4.2 characterizes the solution set of Eq. (19).

Proposition 4.2. *The set of solutions for LMO task in Eq. (19) is the subdifferential of the Tensorion dual norm at point M:*

$$X^{opt} \in \partial \|M\|_{\Sigma}^{\dagger} = \operatorname*{conv}_{\tau \in I} \left(\partial \|M_{(\tau)}\|_{*}\right), \tag{20}$$

where $\operatorname{conv}(\cdot)$ *denotes a convex hull and*

$$I = \left\{\tau \,\middle|\, \|M_{(\tau)}\|_{*} = \|M\|_{\Sigma}^{\dagger}\right\} \subseteq 2^{\{1,\ldots,d\}} \tag{21}$$

is a nonempty set of unfolding indices attaining the maximal nuclear norm.

Proof is in Supplementary Materials Section A, Proposition A.4. Importantly, Proposition 4.2 gives us a practical way to solve the LMO with $\|\cdot\|_{\Sigma}$ by finding an unfolding with the largest nuclear norm. In particular, for any $\tau \in I$, one admissible choice is $X^{\text{opt}} = UV^{\top}$, where U and V are the left and right singular vectors obtained from the singular value decomposition $M_{(\tau)} = U\Sigma V^{\top}$.

The best unfolding can be different for different layers and even change during the optimization process. We found that searching for optimal unfoldings during every iteration of the optimization process is excessive (see numerical experiments). Even though it is possible to do it efficiently, using, *e.g.*, Newton-Schulz iteration run in parallel, there are simpler ways that lead to similar optimization trajectory in practice.

4.3. Motivation for the choice of unfolding shape

Observe that for any unfolding $M_{(\tau)} \in \mathbb{R}^{m_\tau \times n_\tau}$

$$\|M\|_F^2 = \|M_{(\tau)}\|_F^2 = \sum_{j=1}^{r_\tau} \left(\sigma_j^{(\tau)}\right)^2, \; r_\tau = \operatorname{rank}(M_{(\tau)}), \tag{22}$$

where $\{\sigma_j^{(\tau)}\}$ denote the singular values of $M_{(\tau)}$, $m_\tau = \prod_{i \in \tau} n_i$, $n_\tau = \prod_{i \notin \tau} n_i$. Since unfolding only permutes tensor entries, the Frobenius norm of the unfolding is independent of the choice of τ. Hence, for a fixed Frobenius norm, maximizing the nuclear norm reduces to maximizing the sum of singular values subject to a fixed sum of their squares. By the Cauchy–Schwarz inequality, the maximum is attained when the norm budget is distributed uniformly across all available singular values. Consequently, the maximal nuclear norm grows with the number of potentially nonzero singular values, motivating the heuristic of choosing an unfolding with the largest possible number of nonzero singular values.

The following formal statement further supports the heuristic of choosing an unfolding that maximizes the effective number of nonzero singular values.

Proposition 4.3 (High-Probability Nuclear Norm Bound)**.** *Let $M \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a random tensor with entries $M_{i_1,\ldots,i_d} \overset{i.i.d.}{\sim} \mathcal{P}$, where $\mathcal{P}$ is a zero-mean sub-exponential distribution. Consider any unfolding $M_{(\tau)} \in \mathbb{R}^{m_\tau \times n_\tau}$ with $m_\tau \leq n_\tau$ such that $m_\tau n_\tau = \prod_{i=1}^{d} n_i$. Then, there exists a constant $C > 0$ (depending on the sub-exponential norm of $\mathcal{P}$) such that with probability at least $1 - \delta$:*

$$\left\|M_{(\tau)}\right\|_{*} \leq C \cdot m_\tau \sqrt{n_\tau} + \log(1/\delta). \tag{23}$$

Moreover, for a fixed total dimension $m_\tau n_\tau$, the value of this upper bound is maximized when the unfolding is square, i.e., $m_\tau = n_\tau = \sqrt{m_\tau n_\tau}$.

Proof is in Supplementary Materials Section B, Proposi-

Algorithm 1 Tensorion

Input: Weight tensor $W \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, step sizes $\{\eta_k\}$, momentum coefficient β.
Output: Weight tensor $W' \in \mathbb{R}^{n_1 \times \cdots \times n_d}$.
$M_0 = 0$;
$\tau := \arg\max_\tau \left(\min \left\{ \prod_{i \in \tau} n_i, \prod_{i \notin \tau} n_i \right\} \right)$;
$m_\tau, n_\tau := \prod_{i \in \tau} n_i, \prod_{i \notin \tau} n_i$;
for $k := 1, \ldots$ **do**
  Compute gradient $G_k \in \mathbb{R}^{n_1 \times \cdots \times n_d}$;
  $M_k := G_k + \beta M_{k-1}$;
  $X_k := \text{fold}\left(\text{NS}\left(\text{unfold}(M_k, \tau)\right), \tau\right)$;
  $\hat{\eta}_k := \eta_k \cdot 0.2 \sqrt{\max(m_\tau, n_\tau)}$;
  $W_{k+1} := (1 - \gamma \eta_k) W_k - \hat{\eta}_k X_k$;
end for
return W_{k+1}.

tion B.1. Therefore, unfoldings whose row and column dimensions are more balanced, in expectation admit potentially yield larger nuclear norms. Motivated by this observation, we consider the following as our primary strategy to select the unfolding index prior to the optimization loop:

$$\tau^{\text{opt}} := \arg\max_{\tau \in \mathcal{T}} \left(\min \left\{ m_\tau, n_\tau \right\} \right), \tag{24}$$

then we restrict the orthogonalization step to the τ^{opt}-unfolding.

4.4. Tensorion algorithm

The proposed theoretical framework leads to our main contribution, the *Tensorion Optimizer*. Algorithm 1 presents the version of the method used in the experimental section.

The first three lines contain the momentum initialization and the computation of the τ^{opt}-unfolding from Eq. (24), which is used to obtain a faster LMO solution. Within the main optimization loop, Tensorion follows the Muon update scheme, with the key modification appearing in line 7.

First, the function $\text{unfold}(\cdot, \tau)$ maps a d-dimensional tensor to its τ-unfolding. Then, Newton–Schulz orthogonalization, denoted $\text{NS}(\cdot)$, is applied. Finally, $\text{fold}(\cdot, \tau)$ maps the result back to the original tensor shape. Changes with respect to the Muon implementation of (Liu et al., 2025) are highlighted in blue.

4.5. Tensorion for Convolution Layers

In this subsection, let us discuss what it means to apply Tensorion to a convolutional kernel tensor. Let $X \in \mathbb{R}^{c_{\text{in}} \times n \times n}$ be the input to a convolutional layer and let $K \in \mathbb{R}^{c_{\text{out}} \times c_{\text{in}} \times h \times w}$ be its kernel. Let $Y \in \mathbb{R}^{c_{\text{out}} \times n \times n}$ denote the output tensor. The convolution operation can be viewed as a linear mapping

$$\text{vec}(Y) = T_K \, \text{vec}(X), \quad T_K \in \mathbb{R}^{c_{\text{out}} n^2 \times c_{\text{in}} n^2}, \tag{25}$$

where $\text{vec}(\cdot)$ denotes the vectorization operator and T_K is a matrix of the layer determined by the kernel K. Note that T_K is a huge matrix with additional structure, *i.e.*, it is a sparse block matrix where each block represents a two-level Toeplitz structure (Grishina et al., 2025). It is used solely for theoretical analysis purposes and is never formed explicitly. The spectral properties of this convolution operator have been shown to affect generalization and training stability (Singla and Feizi, 2019; Agarwal et al., 2019). According to (Grishina et al., 2025), the spectral norm of the linear map T_K is bounded by the spectral tensor norm as stated in Theorem 4.4.

Theorem 4.4 (Grishina et al. (2025)). *Let $T_K \in \mathbb{R}^{c_{out} n^2 \times c_{in} n^2}$ be the Jacobian matrix of a convolutional layer with stride $s = 1$ and zero padding with parameter $p \geq 0$, or circular padding. Let $K \in \mathbb{R}^{c_{out} \times c_{in} \times h \times w}$ be the convolution kernel. Then*

$$\|K\|_\sigma \leq \|T_K\|_2 \leq \sqrt{hw} \, \|K\|_\sigma. \tag{26}$$

As a corollary to this theorem, the same inequality holds for gradients:

$$\|\nabla_K \mathcal{L}(K)\|_\sigma \leq \|\nabla_{T_K} \mathcal{L}(T_K)\|_2 \leq \sqrt{hw} \|\nabla_K \mathcal{L}(K)\|_\sigma, \tag{27}$$

where $\mathcal{L}$ denotes the loss function, $\nabla_{T_K} \mathcal{L}(T_K) = T_{\nabla_K \mathcal{L}(K)} \in \mathbb{R}^{c_{\text{out}} n^2 \times c_{\text{in}} n^2}$, $\nabla_K \mathcal{L}(K) \in \mathbb{R}^{c_{\text{out}} \times c_{\text{in}} \times h \times w}$ is the loss function gradient w.r.t. T_K (with its structure constraints) and K respectively.

Applying the Muon optimizer directly to the matrix $T_{\nabla_K \mathcal{L}(K)}$ is not practical, as it breaks the sparse multilevel structure of the matrix. However, if we try to solve the LMO problem with Toeplitz-like structure constraints, the (27) becomes useful. Indeed, it allows us to relax the LMO problem to the one with $\|\nabla_K \mathcal{L}(K)\|_\sigma$ instead of $\|\nabla_{T_K} \mathcal{L}(T_K)\|_2$. Finally, using Eq. (18), we further relax the LMO and arrive to the Tensorion approach for the convolution kernel tensor. To conclude, the Tensorion approach is a relaxation of the LMO problem that is applied directly to the matrix of the actual linear convolution operator.

5. Experiments

In order to show the validity of derived heuristic, we ablate on the choice of tensor unfolding index set τ and show its effect on the optimization dynamics (Section 5.1). Afterwards we assess whether the resulting improvements translate into higher final accuracy on standard CNN (Section 5.2) and Vision Transformers (ViTs) (Section 5.3) classification benchmarks. Additional details, like hyperparameters (Secition F)

and further experiments on *e.g.*, Laplacian eigenproblem (Section C.3) or CIFAR datasets (Section C.1), are provided in the Supplementary Material.

5.1. Ablation on an Unfolding Set

We isolate the impact of the unfolding index set τ and evaluate our heuristic choice τ^{opt} from Eq. (24), referred to as the *offline* strategy. As a brute-force reference, we also compare against an *online* strategy that evaluates all admissible unfoldings at each iteration and selects the one maximizing Eq. (21). While potentially more adaptive, this online strategy is substantially more expensive in practice.

For this ablation, we evaluate all seven nontrivial unfoldings of a 4D convolutional kernel up to the complement symmetry $\tau \sim \tau^c$. Static baselines use one unfolding across all layers, while the offline strategy selects τ^{opt} per layer. To isolate unfolding selection from approximation effects, we use exact SVDs instead of NS at line 7 in Algorithm 1 and train ResNet-18 on CIFAR-10 for 25 epochs with identical settings and independently tuned constant learning rates; see Supplementary Materials, Section E.1. Figure 1 and Table 1 show that the choice of unfolding has a substantial effect on performance. The tested unfoldings separate into two accuracy regimes: the lower-performing group consists of $\tau = \{3\}$, $\tau = \{4\}$, and $\tau = \{1, 2\}$, whereas the remaining choices form a higher-performing group. The differences within each group are not statistically significant under this protocol, and therefore we do not attempt to rank individual unfoldings inside the high-accuracy group.

The online strategy, which explicitly maximizes Eq. (21) at every iteration, achieves the highest mean accuracy, as expected from its additional adaptivity. However, this comes at a substantially higher computational cost, and the observed gap to the other high-accuracy methods is not statistically resolved in this ablation. In contrast, the offline heuristic τ^{opt} is inexpensive and still falls into the high-accuracy regime. *This suggests that the heuristic succeeds at avoiding unfavorable unfoldings and recovers a competitive fixed-per-layer choice without requiring brute-force online selection.* Moreover, the elapsed time for Tensorion with offline τ-selection strategy is comparable to the Muon optimizer.

We include the fixed unfolding $\tau = \{1\}$ as the Muon comparison, since this choice recovers the original Muon update in our framework. The main conclusion of the ablation is not a fine-grained ordering among the competitive methods, but rather that Tensorion is sensitive to the unfolding choice and that the proposed heuristic provides a practical way to avoid poor unfoldings without per-iteration search. SGD-M and AdamW (Loshchilov and Hutter, 2017) are included as baselines and achieve lower test accuracy under the same constant-learning-rate protocol. To evaluate the applicability of Tensorion in a larger-scale image classification setting, we next provide a set of experiments on various computer vision problems.

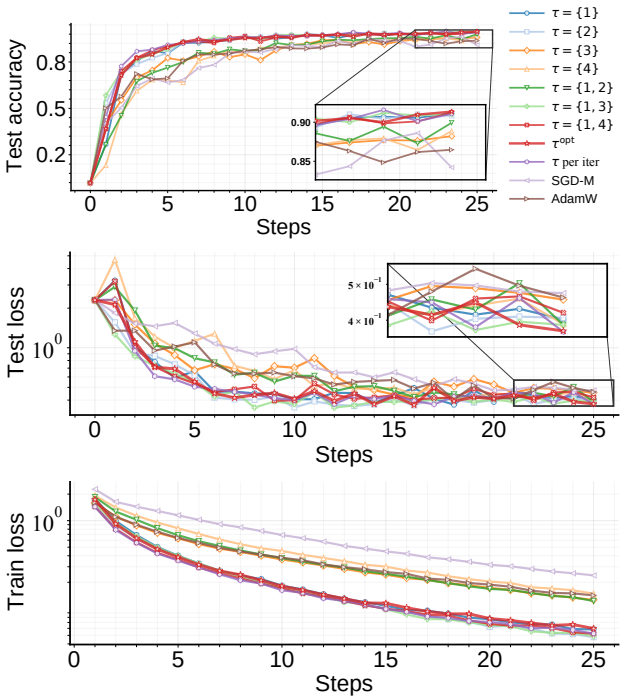


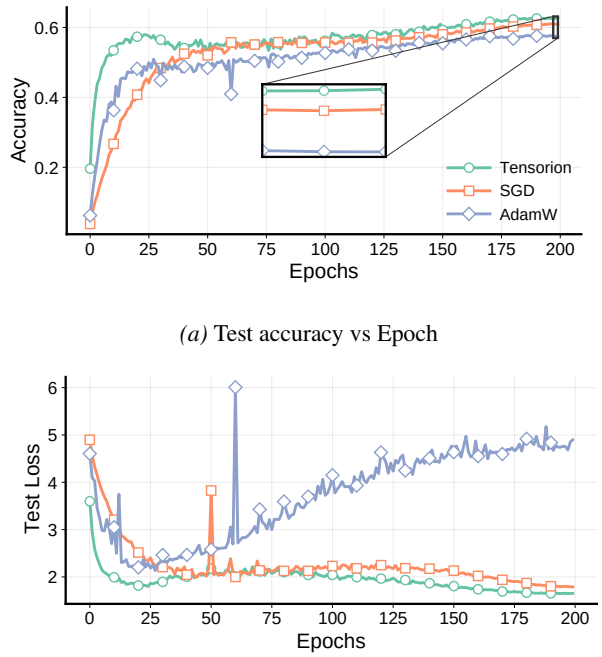
Figure 1. Ablation on the unfolding set τ for ResNet-18 on CIFAR-10 (25 epochs; constant LR, no scheduler) using exact SVD for orthogonalization. Curves labeled " $\tau = \{\cdot\}$" use a fixed unfolding index set $\tau \subset \{1, 2, 3, 4\}$."$\tau^{\text{opt}}$" selects τ once per layer via the offline heuristic in Eq. (24). "τ per iteration" performs online selection by explicitly maximizing $\|X_{(\tau)}\|_*$ over τ at each iteration as in Eq. (21). SGD-M: SGD+Momentum.

Table 1. Ablation on the unfolding set τ for ResNet-18 on CIFAR-10 after 25 epochs. Reported values are mean $\pm$ standard deviation over 10 runs; SGD-M: SGD+Momentum.

Method	**Test acc.** $\uparrow$	**Test loss** $\downarrow$
$\tau = \{1\}$	$90.67_{\pm 0.32}$	$0.424_{\pm 0.026}$
$\tau = \{2\}$	$90.62_{\pm 0.44}$	$\mathbf{0.404}_{\pm \mathbf{0.024}}$
$\tau = \{3\}$	$87.40_{\pm 1.00}$	$0.485_{\pm 0.051}$
$\tau = \{4\}$	$88.05_{\pm 0.69}$	$0.423_{\pm 0.028}$
$\tau = \{1, 2\}$	$88.83_{\pm 0.39}$	$0.430_{\pm 0.022}$
$\tau = \{1, 3\}$	$90.30_{\pm 0.51}$	$0.426_{\pm 0.031}$
$\tau = \{2, 4\}$	$90.77_{\pm 0.45}$	$0.432_{\pm 0.038}$
τ^{opt}	$90.44_{\pm 0.36}$	$0.441_{\pm 0.024}$
online τ	$\mathbf{91.10}_{\pm \mathbf{0.39}}$	$0.416_{\pm 0.026}$
SGD-M	$86.10_{\pm 1.82}$	$0.527_{\pm 0.095}$
AdamW	$86.04_{\pm 1.05}$	$0.507_{\pm 0.041}$

5.2. Tiny ImageNet with ResNet50

We train a ResNet-50 model on Tiny ImageNet (Yin et al., 2023) for 200 epochs, comparing Tensorion against SGD and AdamW baselines. Tensorion is applied only to tensor-valued parameters of order $d \geq 3$ (i.e., convolutional kernels), while all remaining parameters ($d \leq 2$) are optimized with SGD. We report mean $\pm$ std over multiple seeds.



(a) Test accuracy vs Epoch

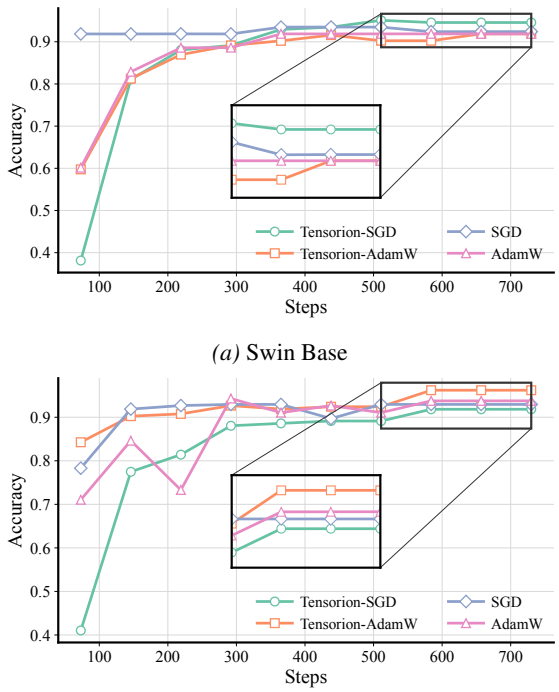
(b) Test loss vs Epoch

Figure 2. Tiny ImageNet results for ResNet-50 trained for 200 epochs. We report mean $\pm$ std over multiple seeds for test accuracy and test loss Tensorion is used only for tensor-valued blocks with order $d \geq 3$ (*e.g.*, convolutional kernels), while all remaining parameters ($d \leq 2$) are optimized with a first-order method (SGD).

Table 2. Main results. Final test loss and test accuracy. C10: CIFAR-10; C100: CIFAR-100; Tiny-IN: Tiny ImageNet; RN: ResNet; SGD-M: SGD+Momentum; T-SGD: Tensorion+SGD; T-AdamW: Tensorion+AdamW.

Dataset	Model	Opt.	Loss ↓	Acc. ↑
C10	RN-18	SGD-M	$0.25_{\pm 0.02}$	$94.1_{\pm 0.13}$
		AdamW	$0.44_{\pm 0.02}$	$93.7_{\pm 0.3}$
		T-SGD	$\mathbf{0.21_{\pm 0.01}}$	$\mathbf{95.3_{\pm 0.1}}$
		T-AdamW	$0.59_{\pm 0.08}$	$94.34_{\pm 0.3}$
C100	RN-34	SGD-M	$\mathbf{0.89_{\pm 0.13}}$	$78.2_{\pm 0.2}$
		AdamW	$1.33_{\pm 0.1}$	$74.2_{\pm 0.1}$
		T-SGD	$0.98_{\pm 0.09}$	$\mathbf{79.1_{\pm 0.1}}$
		T-AdamW	$4.89_{\pm 0.11}$	$73.8_{\pm 0.2}$
Tiny-IN	RN-50	SGD-M	$1.79_{\pm 0.06}$	$61.01_{\pm 0.02}$
		AdamW	$\mathbf{1.33_{\pm 0.05}}$	$57.4_{\pm 0.1}$
		T-SGD	$1.65_{\pm 0.06}$	$\mathbf{62.7_{\pm 0.1}}$

As shown in Figure 2, Tensorion consistently outperforms both baselines in test accuracy throughout training. The advantage is even more pronounced in terms of test loss: AdamW exhibits clear overfitting, with its test loss increasing after roughly 50 steps, whereas Tensorion achieves the lowest and monotonically decreasing loss. These findings are consistent with our ResNet-18/-34 experiments on CIFAR-10/-100 reported in Appendix C.1 and summarized alongside all main results in Table 2.



(a) Swin Base

(b) ViT Base

Figure 3. Imagenette results for Swin and ViT models trained for 5 epochs. Tensorion is used only for tensor-valued blocks with order $d \geq 3$ (*e.g.*, convolutional kernels), while all remaining parameters ($d \leq 2$) are optimized with either SGD or AdamW.

5.3. Tensorion consistently improves performance of Vision Transformers.

Although Vision Transformers mostly consist of common matrix layers, convolutional patch embedding part is crucial for optimization stability and generalization of the resulting model (Xiao et al., 2021). We aim to show that applying Tensorion to this particular module is enough to improve performance across multiple scales and architectures. Specifically, initialize ViT (Dosovitskiy et al., 2020) (Small, Base, Medium) and Swin Transformer (Liu et al., 2021) (Tiny, Small, Base) models from pretrained checkpoints in the `timm` library and finetune the models on Imagenette dataset (Howard, 2019) for 5 epochs with fixed LR scheduler and weight decay across optimizers and evaluate performance at the end of each epoch. Tensorion is only applied to the

convolutional patch embedding layer. All the others are optimized with AdamW or SGD respectively. Optimal learning rates were tuned via grid search for each setup. We report the best performance across algorithms in Fig. 3. Detailed results are reported in C.2.

6. Conclusion

We propose Tensorion, a novel optimizer that generalizes Muon to tensors via a norm that lower bounds the spectral norm of any tensor unfolding, leading to a practical relaxation of the tensor spectral-norm LMO problem. A heuristic offline method for choosing the unfolding achieves performance comparable to the online optimal choice. Experiments across multiple CV tasks and architectures show consistent improvements over conventional optimizers.

Limitations

Our study is limited primarily to computer vision models. We do not evaluate video models or architectures operating on sequential visual domains. We also do not assess whether similar behavior appears for convolutional or any other tensor-valued layers in other domains beyond vision. Finally, our experiments are limited in scale: while the observed trends are encouraging, we cannot determine from the present evidence whether the method scales as reliably as Muon or other established optimizers in substantially larger training regimes.

Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University № 139-15-2025-009. This research was supported in part through computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

References

Chirag Agarwal, Anh Nguyen, and Dan Schonfeld. Improving robustness to adversarial examples by encouraging discriminative features. In *2019 IEEE international conference on image processing (ICIP)*, pages 3801–3805. IEEE, 2019. doi: 10.1109/ICIP.2019.8803601.

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

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

Jeremy Bernstein. Modular manifolds. *Thinking Machines Lab: Connectionism*, 2025b. doi: 10.64434/tml.20250926. https://thinkingmachines.ai/blog/modular-manifolds/.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

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.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=WtbXgc9GVA.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

MMPreTrain Contributors. Openmmlab's pre-training toolbox and benchmark. https://github.com/open-mmlab/mmpretrain, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. *Advances in Neural Information Processing Systems*, 36:33624–33655, 2023.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87 (311):1255–1281, 2018. doi: 10.1090/mcom/3239.

Igor Vladimirovich Girsanov. *Lectures on mathematical theory of extremum problems*. Springer Science & Business Media, 2012.

Ekaterina Grishina, Mikhail Gorbunov, and Maxim Rakhuba. Tight and efficient upper bound on spectral norm of convolutional layers. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 19–34, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73024-5. doi: 10.1007/978-3-031-73024-5_2.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2016.

Yuxuan Gu, Wuyang Zhou, Giorgos Iacovides, and Danilo Mandic. Tensorllm: Tensorising multi-head attention for enhanced reasoning and compression in llms. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *ArXiv*, abs/0911.1393, 2009. doi: 10.1145/2512329.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://doi.org/10.48550/arXiv.1412.6980.

PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yi Li, Zhichun Guo, Miao Yin, and Bingzhe Li. Lestd: Llm compression via learning-based sparse tensor decomposition. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lek-Heng Lim and Pierre Comon. Blind multilinear identification. *IEEE Transactions on Information Theory*, 60 (2):1260–1280, 2014. doi: 10.1109/TIT.2013.2291876.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

Zhanfeng Mo, Long-Kai Huang, and Sinno Jialin Pan. Parameter and memory efficient pretraining via low-rank riemannian optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015. doi: 10.5555/2969239.2969289.

James Oldfield, Markos Georgopoulos, Grigorios G. Chrysos, Christos Tzelepis, Yannis Panagakis, Mihalis A. Nicolaou, Jiankang Deng, and Ioannis Patras. Multilinear mixture of experts: scalable expert specialization through factorization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Ivan Peshekhonov, Aleksey Arzhantsev, and Maxim Rakhuba. Training a tucker model with shared factors: a riemannian optimization approach. In *International*

Conference on Artificial Intelligence and Statistics, pages 3304–3312. PMLR, 2024.

T. Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio José Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained lmos. *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2502.07529.

Anh-Huy Phan, Konstantin Sobolev, Konstantin Sozykin, Dmitry Ermilov, Julia Gusak, Petr Tichavskỳ, Valeriy Glukhov, Ivan Oseledets, and Andrzej Cichocki. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conference on Computer Vision*, pages 522–539. Springer, 2020.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. Learning separable filters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2754–2761, 2013.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Sahil Singla and Soheil Feizi. Fantastic four: Differentiable bounds on singular values of convolution layers. *arXiv preprint arXiv:1911.10258*, 2019.

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

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.

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear algebra and its applications*, 520:44–66, 2017. doi: 10.1016/j.laa.2017.01.017.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Yuebin XU, YANHONG WANG, Xuemei Peng, Hui Zang, Chen Minghao, Pengfei Xia, and Zeyi Wen. TD-moe: Tensor decomposition for moe models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=D9cnZNZfxX.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Proceedings of the Advances*

in Neural Information Processing Systems (NeurIPS), 2023.

Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Dongyang Li, Yupeng Su, Sijia Liu, and Zheng Zhang. Teon: Tensorized orthonormalization beyond layer-wise muon for large language model pre-training, 2026. URL https://arxiv.org/abs/2601.23261.

Yunfan Zhao. Concentration inequalities for sub-weibull random tensors, 2026. URL https://arxiv.org/abs/2509.03439.

A. Primal norm

Theorem A.1 (Primal relaxed spectral norm)**.** *Let $X \in \mathbb{R}^{n_1 \times \dots \times n_d}$ then*

$$\|X\|_\Sigma = \min_{\substack{(\Xi^\tau)_{\tau\in\mathcal{T}} \\ \sum_{\tau\in\mathcal{T}} \Xi^\tau = X}} \sum_{\tau\in\mathcal{T}} \left\| \Xi^\tau_{(\tau)} \right\|_2, \tag{28}$$

where $\Xi^\tau \in \mathbb{R}^{n_1 \times \dots \times n_d}, \tau \in \mathcal{T}$ are auxiliary variables.

Proof. Since $\mathcal{T}$ is finite, the constraint

$$\|M\|^\dagger_\Sigma = \max_{\tau\in\mathcal{T}} \|M_{(\tau)}\|_* \le 1 \tag{29}$$

is equivalent to the collection of constraints

$$\|M_{(\tau)}\|_* \le 1 \qquad \forall \tau \in \mathcal{T}. \tag{30}$$

Hence, by introducing auxiliary tensors A^τ that duplicate M, we can rewrite the definition of the dual norm as

$$\|X\|_\Sigma = \max_{M,(A^\tau)_{\tau\in\mathcal{T}}} \left\{ \langle M, X\rangle : \; M = A^\tau, \; \|A^\tau_{(\tau)}\|_* \le 1 \; \forall \tau \in \mathcal{T} \right\}. \tag{31}$$

Equivalently, we consider the minimization problem

$$\min_{M,(A^\tau)_{\tau\in\mathcal{T}}} \left\{ -\langle M, X\rangle \colon M = A^\tau, \; \|A^\tau_{(\tau)}\|_* \le 1 \; \forall \tau \in \mathcal{T} \right\}. \tag{32}$$

Its optimal value is $-\|X\|_\Sigma$.

This is a finite-dimensional convex problem, and Slater's condition holds (for instance, $M = 0$ and $A^\tau = 0$ strictly satisfy all inequality constraints). Therefore, strong duality applies.

Introduce dual variables Ξ^τ for the constraints $M = A^\tau$ and nonnegative multipliers λ^τ for the constraints $\|A^\tau_{(\tau)}\|_* \le 1$. The Lagrangian is

$$\mathcal{L}(M, A, \lambda, \Xi) = -\langle M, X\rangle + \sum_{\tau\in\mathcal{T}} \Big[\lambda^\tau \big(\|A^\tau_{(\tau)}\|_* - 1 \big) + \langle \Xi^\tau, M - A^\tau \rangle \Big]. \tag{33}$$

Rearranging terms gives

$$\mathcal{L}(M, A, \lambda, \Xi) = \left\langle -X + \sum_{\tau\in\mathcal{T}} \Xi^\tau, \; M \right\rangle + \sum_{\tau\in\mathcal{T}} \Big[\lambda^\tau \|A^\tau_{(\tau)}\|_* - \langle \Xi^\tau, A^\tau \rangle - \lambda^\tau \Big]. \tag{34}$$

We now minimize over the primal variables. First, if

$$\sum_{\tau\in\mathcal{T}} \Xi^\tau \ne X, \tag{35}$$

then the term linear in M is nonzero, and minimizing over M yields $-\infty$. Thus the dual function is finite only if

$$\sum_{\tau\in\mathcal{T}} \Xi^\tau = X. \tag{36}$$

Assume from now on that this constraint holds. Then the minimization over A^τ separates across τ. For each τ, we use that unfolding preserves the Frobenius inner product:

$$\langle \Xi^\tau, A^\tau \rangle = \langle \Xi^\tau_{(\tau)}, A^\tau_{(\tau)} \rangle. \tag{37}$$

Next, recall the standard duality between the nuclear and spectral norms:

$$\langle U, V \rangle \le \|U\|_2 \, \|V\|_*. \tag{38}$$

Applying this with $U = \Xi^\tau_{(\tau)}$ and $V = A^\tau_{(\tau)}$ gives

$$\langle \Xi^\tau, A^\tau \rangle \le \|\Xi^\tau_{(\tau)}\|_2 \, \|A^\tau_{(\tau)}\|_*. \tag{39}$$

Therefore,

$$\lambda^\tau \|A^\tau_{(\tau)}\|_* - \langle \Xi^\tau, A^\tau \rangle - \lambda^\tau \ge \big(\lambda^\tau - \|\Xi^\tau_{(\tau)}\|_2\big)\|A^\tau_{(\tau)}\|_* - \lambda^\tau. \tag{40}$$

If $\lambda^\tau \ge \|\Xi^\tau_{(\tau)}\|_2$, the right-hand side is bounded below by $-\lambda^\tau$, and this bound is attained at $A^\tau = 0$. Hence the minimum over A^τ equals $-\lambda^\tau$.

If instead $\lambda^\tau < \|\Xi^\tau_{(\tau)}\|_2$, then the coefficient in front of $\|A^\tau_{(\tau)}\|_*$ is negative. Using the fact that the spectral norm is dual to the nuclear norm, one can choose a matrix B^τ with

$$\|B^\tau\|_* = 1, \qquad \langle \Xi^\tau_{(\tau)}, B^\tau \rangle = \|\Xi^\tau_{(\tau)}\|_2. \tag{41}$$

Let $\widetilde{A}^\tau$ be the tensor whose τ-matricization is B^τ. Evaluating the Lagrangian at $t\widetilde{A}^\tau$ and sending $t \to \infty$ shows that the infimum is then $-\infty$.

Thus the dual function is

$$\mathcal{G}(\lambda, \Xi) = \begin{cases} -\sum\limits_{\tau \in \mathcal{T}} \lambda^\tau, & \sum\limits_{\tau \in \mathcal{T}} \Xi^\tau = X \text{ and } \lambda^\tau \ge \|\Xi^\tau_{(\tau)}\|_2 \; \forall \tau \in \mathcal{T}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{42}$$

The dual problem is therefore

$$\max_{\lambda, \Xi} \mathcal{G}(\lambda, \Xi) = - \min_{\substack{\sum_{\tau \in \mathcal{T}} \Xi^\tau = X \\ \lambda^\tau \ge \|\Xi^\tau_{(\tau)}\|_2, \; \forall \tau \in \mathcal{T}}} \sum_{\tau \in \mathcal{T}} \lambda^\tau. \tag{43}$$

For fixed $(\Xi^\tau)_{\tau \in \mathcal{T}}$, the minimum over λ^τ is attained at

$$\lambda^\tau = \|\Xi^\tau_{(\tau)}\|_2. \tag{44}$$

Hence the dual optimal value is

$$- \min_{\substack{(\Xi^\tau)_{\tau \in \mathcal{T}} \\ \sum_{\tau \in \mathcal{T}} \Xi^\tau = X}} \sum_{\tau \in \mathcal{T}} \|\Xi^\tau_{(\tau)}\|_2. \tag{45}$$

Finally, by strong duality,

$$-\|X\|_\Sigma = - \min_{\substack{(\Xi^\tau)_{\tau \in \mathcal{T}} \\ \sum_{\tau \in \mathcal{T}} \Xi^\tau = X}} \sum_{\tau \in \mathcal{T}} \|\Xi^\tau_{(\tau)}\|_2, \tag{46}$$

and multiplying by -1 yields

$$\|X\|_\Sigma = \min_{\substack{(\Xi^\tau)_{\tau \in \mathcal{T}} \\ \sum_{\tau \in \mathcal{T}} \Xi^\tau = X}} \sum_{\tau \in \mathcal{T}} \|\Xi^\tau_{(\tau)}\|_2. \tag{47}$$

This proves the claim. □

Lemma A.2. *Let $\|\cdot\|_A, \|\cdot\|_B$ be any norms defined on the same finite-dimensional linear space V and $\|\cdot\|_A^\dagger, \|\cdot\|_B^\dagger$ are their duals. Let also*

$$\|v\|_A \le \|v\|_B, \quad v \in V, \tag{48}$$

then

$$\|u\|_A^\dagger \ge \|u\|_B^\dagger, \quad u \in V^*. \tag{49}$$

Proof. By definition of the dual norm:

$$\|u\|_A^\dagger = \max_{v \in B_A(0,1)} \langle v, u \rangle, \quad \|u\|_B^\dagger = \max_{v \in B_B(0,1)} \langle v, u \rangle, \tag{50}$$

where

$$B_A(0,1) := \{v \in V \mid \|v\|_A \leq 1\}, \quad B_B(0,1) := \{v \in V \mid \|v\|_B \leq 1\}. \tag{51}$$

From Eq. (48) it follows that $B_B(0,1) \subseteq B_A(0,1)$, therefore maximum over $B_A(0,1)$ in Eq. (50) is not smaller than maximum over $B_B(0,1)$ yielding the proposed inequality. □

Proposition A.3. *Let $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be any tensor then $\|X\|_\Sigma^\dagger \leq \|X\|_\sigma^\dagger$ and*

$$\|X\|_\sigma \leq \|X\|_\Sigma \leq \min_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_2. \tag{52}$$

Proof. We begin by comparing the dual norms. By the definition of $\|\cdot\|_\Sigma^\dagger$ (see Eq. (15)),

$$\|M\|_\Sigma^\dagger = \max_{\tau \in \mathcal{T}} \|M_{(\tau)}\|_*. \tag{53}$$

Applying Lemma A.2 to $\|X_{(\tau)}\|_2 \geq \|X\|_\sigma$ for each unfolding yields

$$\|M_{(\tau)}\|_* \leq \|M\|_* \quad \forall \tau \in \mathcal{T}. \tag{54}$$

Therefore,

$$\|M\|_\Sigma^\dagger = \max_{\tau \in \mathcal{T}} \|M_{(\tau)}\|_* \leq \|M\|_*. \tag{55}$$

Finally, the tensor nuclear norm is dual to the tensor spectral norm (Lim and Comon, 2014, Lemma 21), so $\|M\|_* = \|M\|_\sigma^\dagger$. Hence $\|M\|_\Sigma^\dagger \leq \|M\|_\sigma^\dagger$.

We now turn to the lower bound on $\|X\|_\Sigma$. Applying Lemma A.2 to

$$\|M\|_\Sigma^\dagger \leq \|M\|_\sigma^\dagger \tag{56}$$

yields

$$\|X\|_\sigma \leq \|X\|_\Sigma \tag{57}$$

It remains to establish the upper bound. By the variational representation of $\|\cdot\|_\Sigma$ (see Eq. (28)),

$$\|X\|_\Sigma = \min_{\substack{(\Xi^\tau)_{\tau \in \mathcal{T}} \\ \sum_{\tau \in \mathcal{T}} \Xi^\tau = X}} \sum_{\tau \in \mathcal{T}} \|\Xi^\tau_{(\tau)}\|_2. \tag{58}$$

Fix any $\tau' \in \mathcal{T}$, and consider the feasible decomposition

$$\Xi^\tau = \begin{cases} X, & \tau = \tau', \\ 0, & \tau \neq \tau'. \end{cases} \tag{59}$$

Substituting this choice into the objective yields

$$\|X\|_\Sigma \leq \sum_{\tau \in \mathcal{T}} \|\Xi^\tau_{(\tau)}\|_2 = \|\Xi^{\tau'}_{(\tau')}\|_2 = \|X_{(\tau')}\|_2. \tag{60}$$

Since τ' was arbitrary, we conclude that

$$\|X\|_\Sigma \leq \min_{\tau \in \mathcal{T}} \|X_{(\tau)}\|_2. \tag{61}$$

This completes the proof. □

Proposition A.4. *The set of solutions for LMO task in Eq. (19) is the subdifferential of the relaxed spectral dual norm at point M:*

$$X^{opt} \in \partial \|M\|_{\Sigma}^{\dagger} = \operatorname*{conv}_{\tau \in I} \left(\partial \|M_{(\tau)}\|_* \right), \tag{62}$$

where $\operatorname{conv}(\cdot)$ *denotes a convex hull and*

$$I = \left\{ \tau \,\middle|\, \|M_{(\tau)}\|_* = \|M\|_{\Sigma}^{\dagger} \right\} \subseteq 2^{\{1,\dots,d\}} \tag{63}$$

is a nonempty set of unfolding indices attaining the maximal nuclear norm.

Proof. First, by definition of the relaxed spectral dual norm (see Eq. (15)),

$$\|M\|_{\Sigma}^{\dagger} = \max_{\tau \in \mathcal{T}} \|M_{(\tau)}\|_*. \tag{64}$$

For every $\tau \in \mathcal{T}$, define

$$f_\tau(M) := \|M_{(\tau)}\|_*. \tag{65}$$

Since the unfolding map $M \mapsto M_{(\tau)}$ is linear and the nuclear norm is convex, each function f_τ is convex in M. Because the index set $\mathcal{T}$ is finite, we may apply the Dubovitskii-Milyutin subdifferential formula (see, *e.g.*(Girsanov, 2012)) for the maximum of finitely many convex functions:

$$\partial \left(\max_{\tau \in \mathcal{T}} f_\tau(M) \right) = \operatorname{conv} \left(\bigcup_{\tau \in I} \partial f_\tau(M) \right), \tag{66}$$

where

$$I = \left\{ \tau \in \mathcal{T} \,\middle|\, f_\tau(M) = \max_{\theta \in \mathcal{T}} f_\theta(M) \right\}. \tag{67}$$

Substituting the definition of f_τ, we obtain

$$\partial \|M\|_{\Sigma}^{\dagger} = \operatorname{conv}_{\tau \in I} \left(\partial \|M_{(\tau)}\|_* \right), \tag{68}$$

with

$$I = \left\{ \tau \in \mathcal{T} \,\middle|\, \|M_{(\tau)}\|_* = \|M\|_{\Sigma}^{\dagger} \right\}. \tag{69}$$

It remains to identify this subdifferential with the solution set of the LMO in Eq. (19). For any norm $\|\cdot\|_A$ and its dual norm $\|\cdot\|_A^{\dagger}$ (Rockafellar, 1997, Theorem 23.5)

$$\partial \|x\|_A = \operatorname{Argmax} \left\{ \langle z, x \rangle \,\middle|\, \|z\|_A^{\dagger} \le 1 \right\}, \quad x \neq 0. \tag{70}$$

Applying this general result with

$$\|\cdot\|_A = \|\cdot\|_{\Sigma}^{\dagger}, \qquad \|\cdot\|_A^{\dagger} = \|\cdot\|_{\Sigma}, \tag{71}$$

we obtain

$$\partial \|M\|_{\Sigma}^{\dagger} = \operatorname{Argmax} \left\{ \langle X, M \rangle \mid \|X\|_{\Sigma} \le 1 \right\}. \tag{72}$$

Combining this identity with the representation established above completes the proof. □

B. High-Probability Nuclear Norm Bound

Proposition B.1 (High-Probability Nuclear Norm Bound)**.** *Let $M \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a random tensor with entries $M_{i_1,\dots,i_d} \overset{i.i.d.}{\sim} \mathcal{P}$, where $\mathcal{P}$ is a zero-mean sub-exponential distribution. Consider any unfolding $M_{(\tau)} \in \mathbb{R}^{m_\tau \times n_\tau}$ with $m_\tau \le n_\tau$ such that $m_\tau n_\tau = \prod_{i=1}^{d} n_i$. Then, there exists a constant $C > 0$ (depending on the sub-exponential norm of $\mathcal{P}$) such that with probability at least $1 - \delta$:*

$$\left\| M_{(\tau)} \right\|_* \le C \cdot m_\tau \sqrt{n_\tau} + \log(1/\delta) \tag{73}$$

Moreover, for a fixed total dimension $m_\tau n_\tau$, the value of this upper bound is maximized when the unfolding is square, i.e., $m_\tau = n_\tau = \sqrt{m_\tau n_\tau}$.

Proof. Let

$$x := \mathrm{vec}(M) \in \mathbb{R}^N, \quad N = \prod_{i=1}^{d} n_i. \tag{74}$$

Since unfolding only permutes the entries of the tensor, we have

$$\|M_{(\tau)}\|_F = \|M\|_F = \|x\|_2 \tag{75}$$

for every admissible $\tau \in \mathcal{T}$.

We first recall the standard comparison between the nuclear and Frobenius norms. Let

$$A \in \mathbb{R}^{m \times n} \tag{76}$$

have singular values

$$\sigma_1 \geq \cdots \geq \sigma_r > 0, \tag{77}$$

where $r = \mathrm{rank}(A) \leq \min(m, n)$. Then

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i \tag{78}$$

and

$$\|A\|_F = \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{1/2}. \tag{79}$$

By the Cauchy–Schwarz inequality,

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i \cdot 1 \leq \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{1/2} \left(\sum_{i=1}^{r} 1^2 \right)^{1/2} = \sqrt{r}\, \|A\|_F \leq \sqrt{\min(m, n)}\, \|A\|_F. \tag{80}$$

In particular, since $m_\tau \leq n_\tau$,

$$\|M_{(\tau)}\|_* \leq \sqrt{m_\tau}\, \|M_{(\tau)}\|_F = \sqrt{m_\tau}\, \|x\|_2. \tag{81}$$

It therefore remains to control $\|x\|_2$. Since the entries of M are i.i.d., centered, unit-variance, and sub-exponential, (Zhao, 2026, Lemma 4.1) applied with $\alpha = 1$ yields the concentration bound

$$\mathbb{P}\Big(\Big| \|x\|_2 - \sqrt{N} \Big| \geq t \Big) \leq 2 \exp\big(-c \min(t^2, t) \big), \qquad t \geq 0, \tag{82}$$

for some constant $c > 0$ depending only on the sub-exponential norm of $\mathcal{P}$.

Now fix $\delta \in (0, 1)$. Choose $t_\delta > 0$ so that

$$2 \exp\big(-c \min(t_\delta^2, t_\delta) \big) \leq \delta. \tag{83}$$

Then, with probability at least $1 - \delta$,

$$\|x\|_2 \leq \sqrt{N} + t_\delta. \tag{84}$$

Since $N \geq 1$, after enlarging the constant if necessary, we may write this bound in the simpler form

$$\|x\|_2 \leq C_\delta \sqrt{N}. \tag{85}$$

Consequently, for every fixed τ we also have the slightly weaker but algebraically convenient estimate

$$\|x\|_2 \leq C_\delta \sqrt{N} + \frac{1}{\sqrt{m_\tau}} \log(1/\delta), \tag{86}$$

again with probability at least $1 - \delta$.

Combining this with the nuclear-versus-Frobenius estimate, we obtain

$$\|M_{(\tau)}\|_* \leq \sqrt{m_\tau}\, \|x\|_2 \leq \sqrt{m_\tau} \left(C_\delta \sqrt{N} + \frac{1}{\sqrt{m_\tau}} \log(1/\delta) \right). \tag{87}$$

Therefore,

$$\|M_{(\tau)}\|_* \leq C_\delta \sqrt{m_\tau N} + \log(1/\delta). \tag{88}$$

Using

$$N = m_\tau n_\tau, \tag{89}$$

we conclude that

$$\|M_{(\tau)}\|_* \leq C_\delta \, m_\tau \sqrt{n_\tau} + \log(1/\delta), \tag{90}$$

as claimed.

Finally, let us justify the statement about the "most square" unfolding. For fixed total dimension N and under the constraint $m_\tau \leq n_\tau$, we have

$$n_\tau = \frac{N}{m_\tau} \tag{91}$$

and hence

$$m_\tau \sqrt{n_\tau} = m_\tau \sqrt{\frac{N}{m_\tau}} = \sqrt{N m_\tau}. \tag{92}$$

Thus the leading term is a monotone increasing function in m_τ. Since admissible unfoldings are determined by partitions of the tensor modes, not every factorization of N need to be realizable. Therefore, among the admissible unfoldings, the largest upper bound is attained by the unfolding with the largest possible value of m_τ subject to $m_\tau \leq n_\tau$, namely by the most balanced admissible unfolding. □

C. Additional experiments

C.1. CIFAR-10 and CIFAR-100

This section presents additional training results for ResNet-18 and ResNet-34 in the CIFAR10 and CIFAR100 datasets respectively. Unlike previous experiments, here we also evaluate the Tensorion Algorithm 1 using the AdamW optimizer for the non-tensor blocks; this setting is denoted as `tensorion_adamw` in Fig. 4 and 5.

It is also important to analyze how the algorithm's behavior depends on the optimizer chosen for the non-tensor parameters of the model. Indeed, the reported experiments show that the behavior strongly depends on this choice. When SGD is used, we observe improvements in both convergence and accuracy compared to applying SGD to all model parameters. For AdamW, we observe a roughly similar picture, where both algorithms achieve approximately the same performance.

C.2. ViT & Swin models

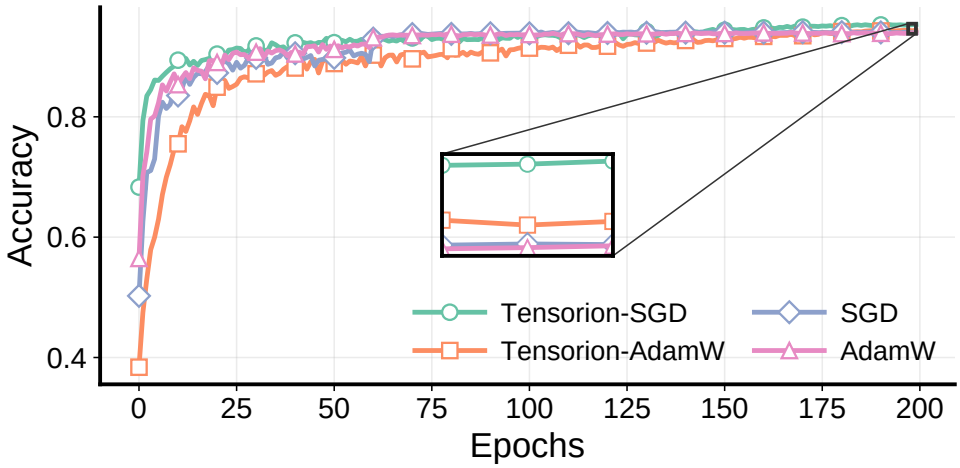
Here we present some additional details on evaluations on transformer-based models. We initialize all the models from pretrained checkpoints available in the `timm` library and finetune the models for 5 epochs on Imagenette dataset. Tensorion is only applied to convolutional layers, while we revert to either AdamW or SGD for other layers to isolate the effects. Notably, applying Tensorion to weight matrices in Transformer blocks would recover Muon updates. We fix weight decay and LR scheduler parameters for this experiment, and only sweep over optimal learning rates for the respective choice of optimizers. We report best runs for each choice of the optimizer in 6 and 7.

C.3. Beyond CNNs: Laplacian Spectrum

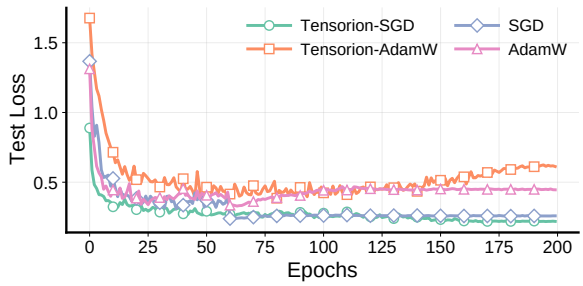
Tensor trains (TT). Tensor Train (TT) decomposition is a low-rank factorization for high-order tensors that represents a d-way tensor $x \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ as a product of *TT-cores*:

$$x(i_1, \ldots, i_d) = G^{(1)}(i_1)\, G^{(2)}(i_2) \cdots G^{(d)}(i_d), \tag{93}$$

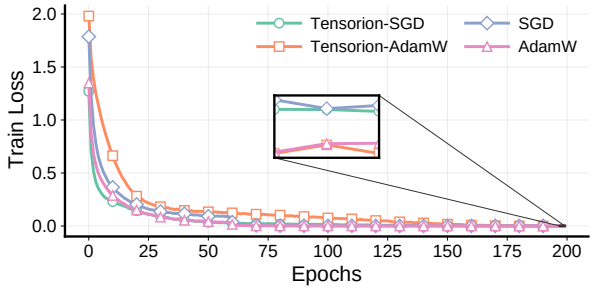
where each core $G^{(k)}(i_k) \in \mathbb{R}^{r_{k-1} \times r_k}$ is a matrix indexed by $i_k \in \{1, \ldots, n_k\}$, with boundary ranks $r_0 = r_d = 1$. The tuple $(r_1, \ldots, r_{d-1})$ are the TT-ranks; small TT-ranks yield a compressed representation that enables scalable optimization in very high dimensions.



(a) Test accuracy vs Epoch



(b) Test loss vs Epoch



(c) Train loss vs Epoch

Figure 4. CIFAR-10 results for ResNet-18 trained for 200 epochs under the standard CIFAR pipeline of (Contributors, 2023). We report mean $\pm$ std over multiple seeds for test accuracy (Figure 4a), test loss (Figure 4b), and train loss (Figure 4c). Tensorion is used only for tensor-valued blocks with order $d \geq 3$ (e.g., convolutional kernels), while all remaining parameters ($d \leq 2$) are optimized with a first-order method (SGD or AdamW).

Discrete Laplacian on $[0,1]^d$. Consider a uniform tensor-product grid on $[0,1]^d$ with n points per dimension and step size $h = \frac{1}{n+1}$ (Dirichlet boundary conditions). Let $L \in \mathbb{R}^{n\times n}$ be the standard 1D second-difference matrix

$$L = \frac{1}{h^2}\,\mathrm{tridiag}(1,-2,1), \tag{94}$$

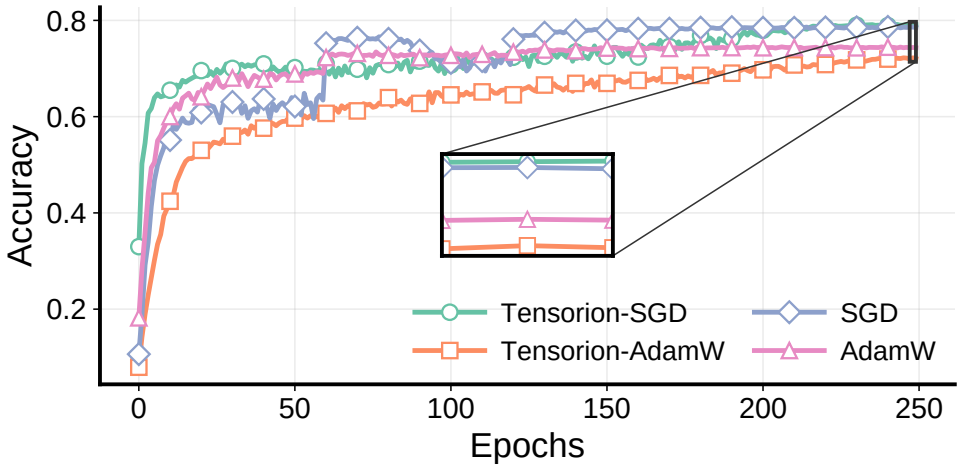
where $\mathrm{tridiag}(1,-2,1)$ denotes tridigaonal matrix.

The d-dimensional discrete Laplacian $\Delta_d \in \mathbb{R}^{n^d\times n^d}$ on the tensor-product grid is given by the Kronecker sum

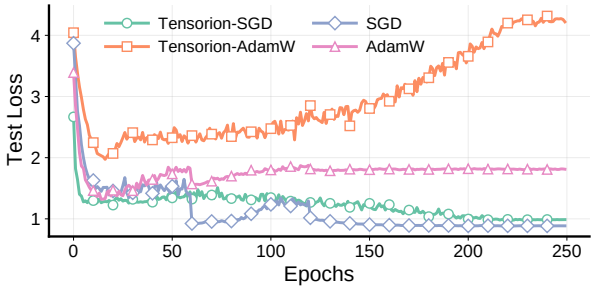
$$\Delta_d = \sum_{k=1}^{d} I^{\otimes(k-1)} \otimes L \otimes I^{\otimes(d-k)}, \tag{95}$$

where $I \in \mathbb{R}^{n\times n}$ is the identity and $\otimes$ denotes the Kronecker product. We treat Δ_d as a linear operator acting on an order-d tensor $x \in \mathbb{R}^{n\times\cdots\times n}$ via vectorization.

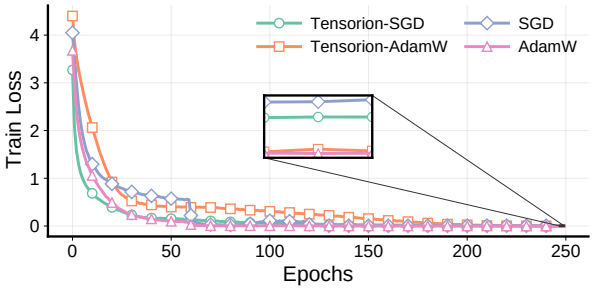
TT-operator form. Each term $I^{\otimes(k-1)} \otimes L \otimes I^{\otimes(d-k)}$ in (95) is a rank-1 Kronecker product of d factors and thus admits an exact TT-operator representation with TT-rank 1. Consequently, the full Laplacian Δ_d, being a sum of d such Kronecker terms, admits a compact TT-operator representation with small TT-ranks (bounded by a constant independent of n^d), which allows for efficient application $\Delta_d x$ and inner products $\langle x, \Delta_d x\rangle$ in TT format.



(a) Test accuracy vs Epoch



(b) Test loss vs Epoch



(c) Train loss vs Epoch

Figure 5. CIFAR-100 results for ResNet-34 trained for 250 epochs under the standard CIFAR pipeline of (Contributors, 2023). We report mean $\pm$ std over multiple seeds for test accuracy (Figure 5a), test loss (Figure 5b), and train loss (Figure 5c). Tensorion is used only for tensor-valued blocks with order $d \geq 3$ (e.g., convolutional kernels), while all remaining parameters ($d \leq 2$) are optimized with a first-order method (SGD or AdamW).

Low-rank structure of eigenvectors. Our goal is to compute the low end of the spectrum of Δ_d in a *compressed* form. We restrict the search space to eigenvectors representable with small TT-ranks, i.e., x is parameterized by TT-cores with ranks bounded by a prescribed r. This reflects the practical regime in high-dimensional PDE discretizations where one seeks approximate eigenmodes within a low-rank manifold.

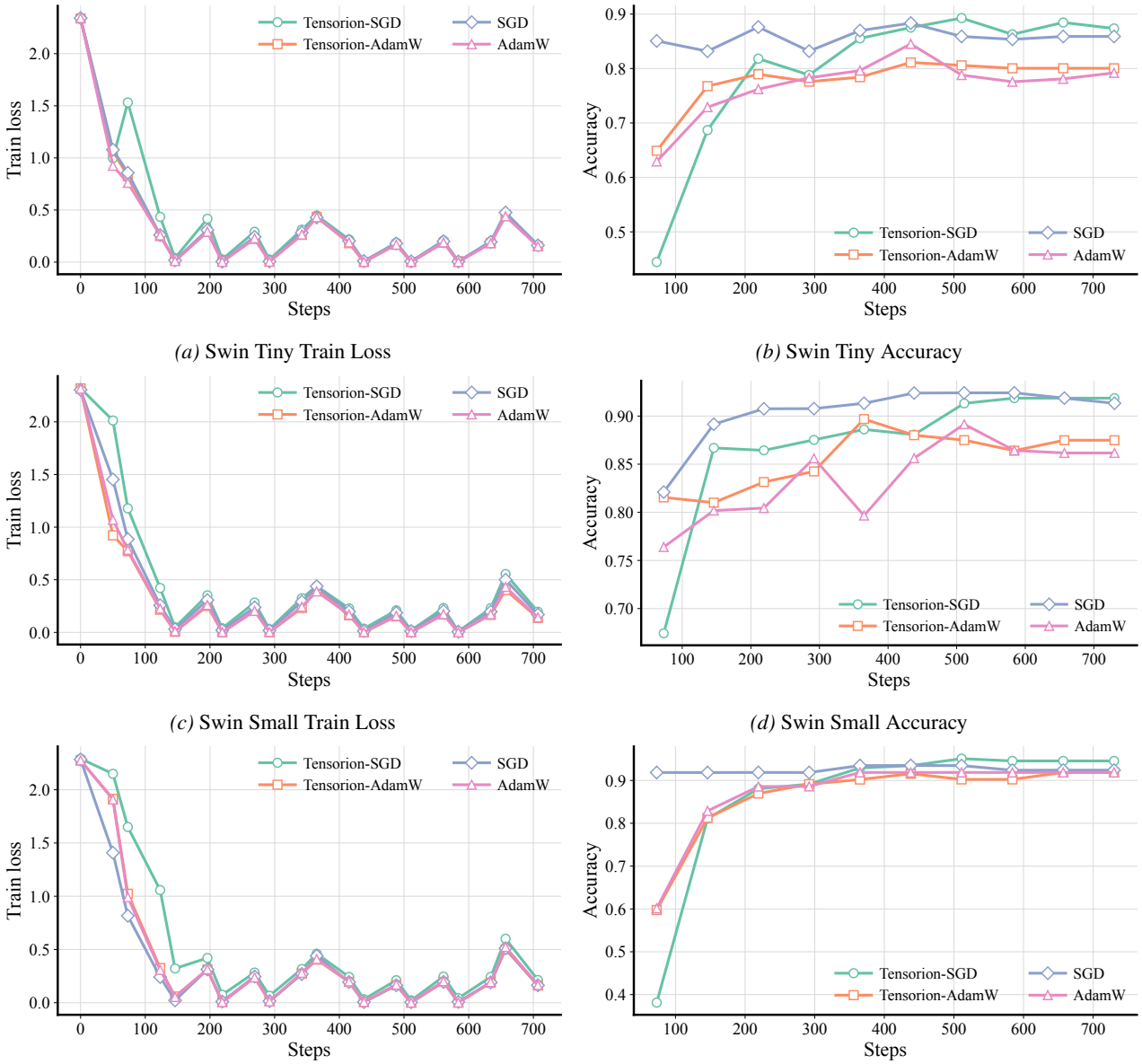
Rayleigh quotient optimization. Let $A := -\Delta_d$ be the discrete positive semidefinite operator. The smallest eigenpair of A can be obtained by minimizing the Rayleigh quotient

$$\min_{x \neq 0} \mathcal{R}(x) \;=\; \min_{x \neq 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle}, \tag{96}$$

In our experiments, we optimize (96) over the TT-parameterization of x with bounded TT-ranks, comparing Tensorion to gradient descent baselines on this non-deep, structured optimization task.

D. Relaxation Tightness

The empirical analysis of relaxation tightness along optimization trajectories of ResNet-18, CIFAR-10 is demonstrated in Fig. 9. The results show that the actual tensor spectral norm of the LMO solutions with relaxed tensor spectral norm on optimization trajectories of ResNet-18 is very close to 1.



(a) Swin Tiny Train Loss

(b) Swin Tiny Accuracy

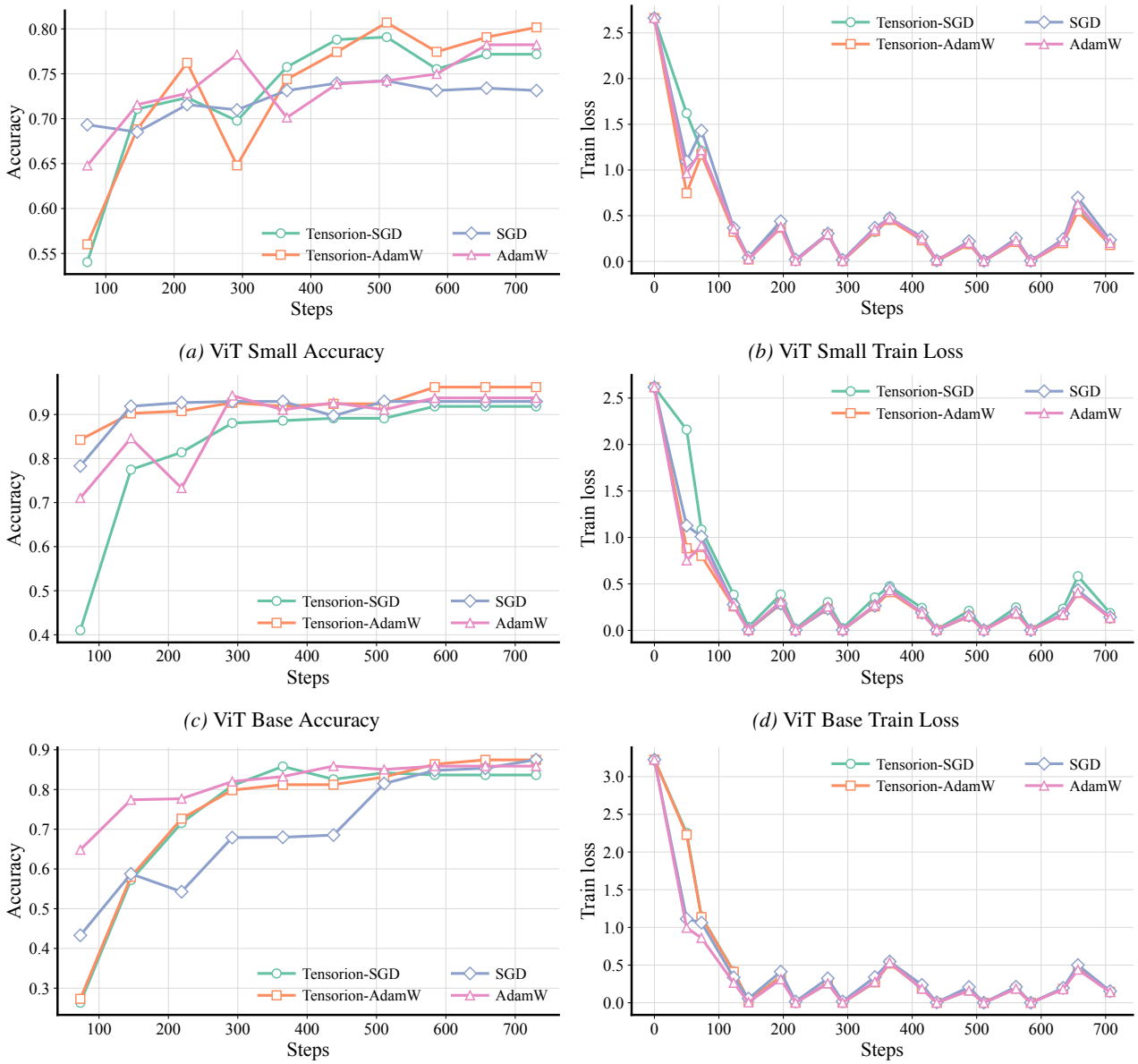
(c) Swin Small Train Loss

(d) Swin Small Accuracy

(e) Swin Base Train Loss

(f) Swin Base Accuracy

Figure 6. Test accuracy and training loss across all Swin models.



(a) ViT Small Accuracy

(b) ViT Small Train Loss

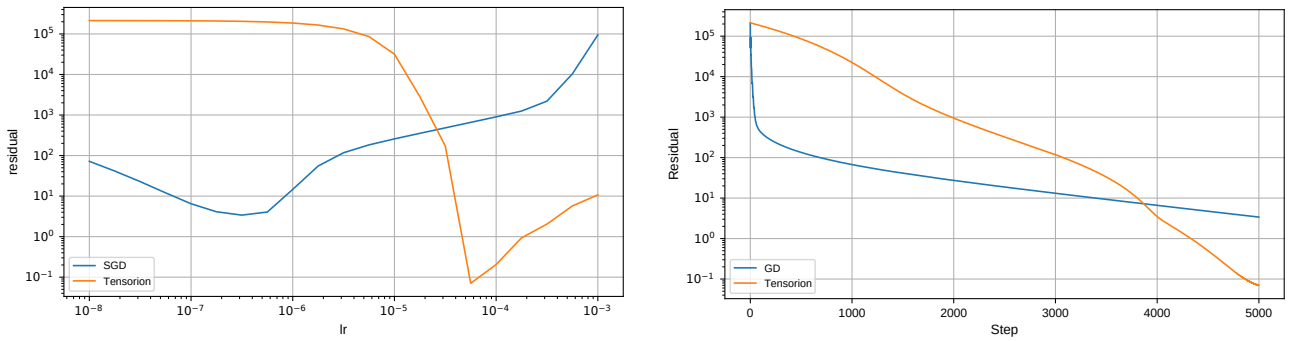
(c) ViT Base Accuracy

(d) ViT Base Train Loss

(e) ViT Medium Accuracy

(f) ViT Medium Train Loss

Figure 7. Test accuracy and training loss across ViT models.



(a) Residual vs learning rate

(b) Residual vs step, optimal learning rate

Figure 8. Minimization of the Rayleigh quotient (96) for the smallest eigenpair of $A = -\Delta_d$ on a d-dimensional tensor-product grid, with the eigenvector x parameterized in TT format with bounded TT-ranks. Left: final functional-residual as a function of the learning rate. Right: functional-residual versus optimization step at the best-tuned learning rate for each method.

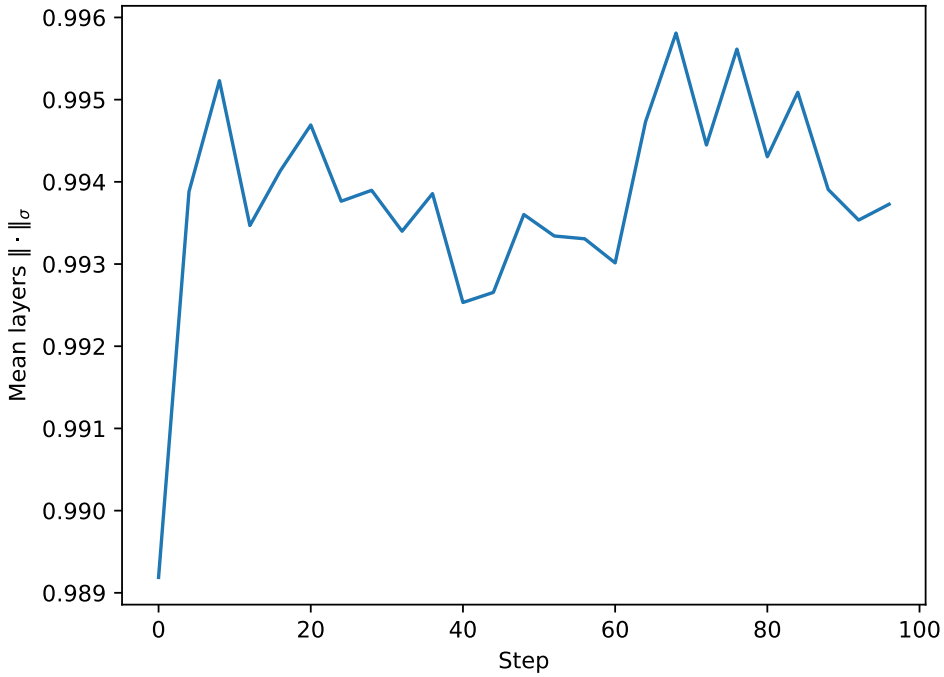


Figure 9. Actual tensor spectral norm of the LMO solutions with relaxed tensor spectral norm on optimization trajectories of ResNet-18, CIFAR-10

E. Experimental Details

E.1. Tau ablation

We enumerate all proper non-empty subsets $\tau \subsetneq \{1, 2, 3, 4\}$. Since unfolding with τ and with its complement τ^c produces matrices that are transposes of each other, and the nuclear norm is invariant to this operation, $\|X_{(\tau)}\|_* = \|X_{(\tau^c)}\|_*$, it suffices to consider τ up to the equivalence $\tau \sim \tau^c$, yielding $(2^4 - 2)/2 = 7$ distinct cases for 4-dimensional convolution kernel.

The convolutional kernel has shape $(C_{\text{out}}, C_{\text{in}}, k_h, k_w)$. In ResNet architectures, it typically holds that

$C_{\text{out}} \geq C_{\text{in}} \gg k_h, k_w$.

In case of static τ selection, the unfolding is fixed across all layers, while the offline strategy selects τ^{opt} for each layer individually. To decouple the effect of τ selection from numerical approximation artifacts, we replace the Newton–Schulz routine (line 7 in Algorithm 1) with an exact SVD throughout this ablation. AdamW was used for layers with weight tensors of order $d \leq 2$.

We train ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky et al., 2009) for a fixed budget of 25 epochs (no early stopping). Across all optimizers, we keep the batch size, weight decay, and image preprocessing identical. We do not use an LR scheduler; instead, the learning rate is tuned *independently for each method* via a grid search (see Table 4).

F. Hyperparameters

The optimal hyperparameter configurations reported throughout the experiments are summarized in Table 3 and Table 4.

Table 3. Hyperparameter configurations used across datasets and optimization setups.

Setup / Method	SGD	AdamW	Tensorion	Tensorion_AdamW
CIFAR10 Scheduler	`Multistep`	`Multistep`	`Onecycle`	`Onecycle`
CIFAR10 LR	0.075	0.0001	0.001	0.075
CIFAR10 WD	0.0005	0.0001	0.0001	0.0005
CIFAR100 Scheduler	`Multistep`	`Multistep`	`Onecycle`	`Onecycle`
CIFAR100 LR	0.05	0.00025	0.0075	0.00005
CIFAR100 WD	0.0005	0.0005	0.0005	0.0005
Tiny Scheduler	`Onecycle`	`Onecycle`	`Onecycle`	–
Tiny LR	0.01	0.0005	0.075	–
Tiny WD	0.0005	0.0001	0.0005	–
Imagenette Scheduler	`Cosine`	`Cosine`	`Cosine`	`Cosine`
Imagenette LR	0.005	0.0001	0.0005	0.0001
Imagenette WD	0.0005	0.0005	0.0005	0.0005

Table 4. Learning rates used in the ablation on the unfolding set τ for ResNet-18 on CIFAR-10 after 25 epochs. The learning rate is tuned per method via grid search.

	Tensorion (static τ)							**Offline**	**Online**	**Baselines**	
Hyperparameter	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	τ^{opt}	online τ	SGD+M.	AdamW
Learning rate	0.01	0.001	0.001	0.005	0.005	0.001	0.005	0.01	0.005	0.05	0.0005

G. Compute resources

All experiments were conducted on a shared high-performance computing infrastructure. The experiments reported in Section 5.1 required approximately 300 GPU-hours on NVIDIA V100-SXM2-32GB accelerators. All remaining experiments, including baseline training, ablation studies, and final model evaluations, consumed approximately 2,000 GPU-hours on NVIDIA A100-SXM-80GB accelerators.

In total, the computational budget for the reported results amounts to roughly 2,300 GPU-hours.

Hyperparameter search, preliminary runs, and failed attempts are included in these estimates. We confirm that the paper reports the full set of experiments conducted within this budget, and no selective reporting based on compute availability was applied.