A Missing proof and results

Lemma A.1 and proof

Lemma A.1. *For any tensor $\mathcal{V}$ that has the same shape as $\mathcal{G}_m$,*

$$\langle \Phi(\ldots, \mathcal{G}_{m-1}, \mathcal{V}, \mathcal{G}_{m+1}, \ldots), \nabla_{\mathcal{T}} f \rangle_F = \langle \mathcal{V}, \nabla_{\mathcal{G}_m} f \rangle_F ,$$

where $\mathcal{T} = \Phi(\mathcal{G}_1, \ldots, \mathcal{G}_m, \ldots, \mathcal{G}_K)$.

Proof. Let $F(\mathcal{G}_1, \ldots, \mathcal{G}_K) := f(\Phi(\mathcal{G}_1, \ldots, \mathcal{G}_m, \ldots, \mathcal{G}_K))$, and fix all G_j except G_m. The directional derivative of F at $\mathcal{G}_m$ in the direction of $\mathcal{V}$ is given by:

$$\begin{aligned} D_{\mathcal{G}_m} F(\mathcal{G}_m; \mathcal{V}) &= \frac{d}{d\epsilon} F(\ldots, \mathcal{G}_m + \epsilon \mathcal{V}, \ldots)\Big|_{\epsilon=0} \\ &= \left\langle \nabla_{\mathcal{T}} f, \frac{d}{d\epsilon} \Phi(\ldots, \mathcal{G}_m + \epsilon \mathcal{V}, \ldots)\Big|_{\epsilon=0} \right\rangle_F \\ &= \langle \nabla_{\mathcal{T}} f, \Phi(\ldots, \mathcal{V}, \ldots) \rangle_F . \end{aligned}$$

By the definition of gradient and chain rule, we have:

$$D_{\mathcal{G}_m} F(\mathcal{G}_m; \mathcal{V}) = \langle \nabla_{\mathcal{G}_m} f, \mathcal{V} \rangle_F .$$

Then, we arrive at the desired result:

$$\langle \Phi(\ldots, \mathcal{V}, \ldots), \nabla_{\mathcal{T}} f \rangle_F = \langle \nabla_{\mathcal{G}_m} f, \mathcal{V} \rangle_F . \tag{1}$$

This completes the proof. □

Proof of Theorem 1

Proof. For any core $\mathcal{G}_k^{(t)}, \forall k \in [K]$, the dynamics of its squared Frobenius norm is given by:

$$\frac{d}{dt} \|\mathcal{G}_k^{(t)}\|_F^2 = 2 \left\langle \mathcal{G}_k^{(t)}, \frac{d}{dt} \mathcal{G}_k^{(t)} \right\rangle_F = -2 \left\langle \mathcal{G}_k^{(t)}, \nabla_{\mathcal{G}_k^{(t)}} f \right\rangle_F .$$

By the lemma (A.1), we have:

$$\begin{aligned} \left\langle \mathcal{G}_k^{(t)}, \nabla_{\mathcal{G}_k^{(t)}} f \right\rangle_F &= \left\langle \Phi(\ldots, \mathcal{G}_{k-1}^{(t)}, \mathcal{G}_k^{(t)}, \mathcal{G}_{k+1}^{(t)}, \ldots), \nabla_{\mathcal{T}} f \right\rangle_F \\ &= \langle \mathcal{T}, \nabla_{\mathcal{T}} f \rangle_F . \end{aligned}$$

Therefore, all cores have the same dynamics as follows:

$$\frac{d}{dt} \|\mathcal{G}_1^{(t)}\|_F^2 = \cdots = \frac{d}{dt} \|\mathcal{G}_K^{(t)}\|_F^2 = -2 \langle \mathcal{T}, \nabla_{\mathcal{T}} f \rangle_F .$$

This completes the proof. □

Proof of Theorem 2

Proof. We have denoted gradients at the perturbed point and unperturbed point as $\tilde{g}_k^{(t)}$ and $g_k^{(t)}$ respectively in (4). Denote $\mathcal{T}^{(t)} = \Phi(\mathcal{G}_1^{(t)}, \ldots, \mathcal{G}_K^{(t)})$, $\tilde{\mathcal{T}}^{(t)} = \Phi(\tilde{\mathcal{G}}_1^{(t)}, \ldots, \tilde{\mathcal{G}}_K^{(t)})$, and $\tilde{f} = f(\tilde{\mathcal{T}}^{(t)})$. Note that perturbed point $\tilde{\mathcal{G}}_k^{(t)}$ satisfies Lemma A.1. By letting $\mathcal{V} = \tilde{\mathcal{G}}_k^{(t)}$, we have:

$$\left\langle \tilde{\mathcal{T}}^{(t)}, \nabla_{\mathcal{T}} \tilde{f} \right\rangle_F = \left\langle \tilde{\mathcal{G}}_k^{(t)}, \tilde{g}_k^{(t)} \right\rangle_F .$$

Next, we expand the dynamics of $\forall i, j \in [K]$ with $i \neq j$ as follows:

$$\begin{aligned} &\frac{d}{dt} \left(\|\mathcal{G}_i^{(t)}\|_F^2 - \|\mathcal{G}_j^{(t)}\|_F^2 \right) \\ &= -2 \left(\left\langle \mathcal{G}_i^{(t)}, \tilde{g}_i^{(t)} \right\rangle_F - \left\langle \mathcal{G}_j^{(t)}, \tilde{g}_j^{(t)} \right\rangle_F \right) \\ &= -2 \Bigg(\Bigg[\underbrace{\left\langle \tilde{\mathcal{G}}_i^{(t)}, \tilde{g}_i^{(t)} \right\rangle_F}_{= \left\langle \tilde{\mathcal{T}}^{(t)}, \nabla_{\mathcal{T}} \tilde{f} \right\rangle_F} - \rho u^{(t)} \left\langle g_i^{(t)}, \tilde{g}_i^{(t)} \right\rangle_F \Bigg] \\ &\quad - \Bigg[\underbrace{\left\langle \tilde{\mathcal{G}}_j^{(t)}, \tilde{g}_j^{(t)} \right\rangle_F}_{= \left\langle \tilde{\mathcal{T}}^{(t)}, \nabla_{\mathcal{T}} \tilde{f} \right\rangle_F} - \rho u^{(t)} \left\langle g_j^{(t)}, \tilde{g}_j^{(t)} \right\rangle_F \Bigg] \Bigg) \\ &= 2\rho u^{(t)} \left(\left\langle g_i^{(t)}, \tilde{g}_i^{(t)} \right\rangle_F - \left\langle g_j^{(t)}, \tilde{g}_j^{(t)} \right\rangle_F \right) \\ &= 2\rho u^{(t)} \left(\|g_i^{(t)}\|_F^2 - \|g_j^{(t)}\|_F^2 \right) \\ &\quad + \underbrace{2\rho u^{(t)} \left(\left\langle g_i^{(t)}, \tilde{g}_i^{(t)} - g_i^{(t)} \right\rangle_F - \left\langle g_j^{(t)}, \tilde{g}_j^{(t)} - g_j^{(t)} \right\rangle_F \right)}_{=: R_{ij}^{(t)}} . \end{aligned}$$

To bound the term $R_{ij}^{(t)}$, we can bound the term for $g_i^{(t)}$ and $g_j^{(t)}$ separately. First, we have:

$$\begin{aligned} 2\rho u^{(t)} \left\langle g_i^{(t)}, \tilde{g}_i^{(t)} - g_i^{(t)} \right\rangle_F &\leq 2\rho \underbrace{u^{(t)} \|g_i^{(t)}\|_F}_{\leq 1} \|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F \\ &\leq 2\rho \|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F . \end{aligned}$$

That is, we need to bound the term $\|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F$. With assumption (1), we have:

$$\|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F \leq L \|\tilde{\mathcal{T}}^{(t)} - \mathcal{T}^{(t)}\|_F .$$

Define $\Delta_k = \tilde{\mathcal{G}}_k^{(t)} - \mathcal{G}_k^{(t)}$. We have that $\|\Delta_k\|_F = \rho u^{(t)} \|g_k^{(t)}\|_F = O(\rho)$. We can expand

$$\begin{aligned} \tilde{\mathcal{T}}^{(t)} - \mathcal{T}^{(t)} = &\underbrace{\sum_{k=1}^{K} \Phi(\mathcal{G}_1^{(t)}, \ldots, \Delta_k, \ldots, \mathcal{G}_K^{(t)})}_{=: A^{(1)}} \\ &+ \underbrace{\sum_{1 \leq i < j \leq K} \Phi(\ldots, \Delta_i, \ldots, \Delta_j, \ldots)}_{=: A^{(2)}} \\ &+ A^{(>2)}, \end{aligned}$$

where $A^{(>2)}$ collects all terms involving at least three perturbations. The norm of $A^{(1)}$ can be bounded as follows:

$$\begin{aligned} \|A^{(1)}\|_F &\leq \sum_{k=1}^{K} \|\Phi(\mathcal{G}_1^{(t)}, \ldots, \Delta_k, \ldots, \mathcal{G}_K^{(t)})\|_F \\ &\leq \sum_{k=1}^{K} \left(\prod_{j \neq k} \|\mathcal{G}_j\|_F \right) \cdot \|\Delta_k\|_F \\ &= O(\rho). \end{aligned}$$

For the term $A^{(2)}$, we have:

$$\begin{aligned}\|A^{(2)}\|_F &\le \sum_{1\le i<j\le K} \|\Phi(\dots,\Delta_i,\dots,\Delta_j,\dots)\|_F \\ &\le \sum_{1\le i<j\le K} \left(\prod_{k\neq i,j} \|\mathcal{G}_k\|_F\right)\cdot \|\Delta_i\|_F \cdot \|\Delta_j\|_F \\ &= O(\rho^2).\end{aligned}$$

Similarly, higher-order terms are all $O(\rho^k)$ with $k \ge 2$, and we can bound:

$$\begin{aligned}\|\tilde{\mathcal{T}}^{(t)} - \mathcal{T}^{(t)}\|_F &\le \|A^{(1)}\|_F + \|A^{(2)}\|_F + \|A^{(>2)}\|_F \\ &\le O(\rho) + O(\rho^2) + O(\rho^3) + \cdots \\ &= O(\rho).\end{aligned}$$

Thus, we have:

$$\|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F \le L\|\tilde{\mathcal{T}}^{(t)} - \mathcal{T}^{(t)}\|_F = O(\rho L).$$

Combining the bounds, we have:

$$\begin{aligned}R_{ij}^{(t)} &\le 2\rho\|\tilde{g}_i^{(t)} - g_i^{(t)}\|_F + 2\rho\|\tilde{g}_j^{(t)} - g_j^{(t)}\|_F \\ &= O(\rho^2 L).\end{aligned}$$

The proof is complete. □

Proof of Corollary 1

Proof. By Theorem 1, we have that $\forall k, j \in [K]$,

$$\frac{d}{dt}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \|\mathcal{G}_j^{(t)}\|_F^2\right) = 0.$$

Therefore, the Norm Deviation Q is constant:

$$\frac{dQ}{dt} = \frac{d}{dt}\frac{1}{2K}\sum_{i,j=1}^{K}\left(\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2\right)^2 = 0.$$

This completes the proof. □

Lemma A.2. *For the cores $\{\mathcal{G}_k\}_{k=1}^K$, with infinitesimal step-size $\eta \to 0$ and update step* (4)*,*

$$\begin{aligned}&\frac{d}{dt}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \frac{1}{K}\sum_{i=1}^{K}\|\mathcal{G}_i^{(t)}\|_F^2\right) \\ &= 2\rho u^{(t)}\left(\|g_k^{(t)}\|_F^2 - \frac{1}{K}\sum_{i=1}^{K}\|g_i^{(t)}\|_F^2\right) + O(\rho^2 L).\end{aligned}$$

Proof. By Theorem 2, we have:

$$\begin{aligned}&\frac{d}{dt}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \|\mathcal{G}_j^{(t)}\|_F^2\right) \\ &= 2\rho u^{(t)}\left(\|g_k^{(t)}\|_F^2 - \|g_j^{(t)}\|_F^2\right) + O(\rho^2 L).\end{aligned}$$

Summing over j and dividing by K gives:

$$\begin{aligned}&\frac{d}{dt}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \frac{1}{K}\sum_{i=1}^{K}\|\mathcal{G}_i^{(t)}\|_F^2\right) \\ &= 2\rho u^{(t)}\left(\|g_k^{(t)}\|_F^2 - \frac{1}{K}\sum_{i=1}^{K}\|g_i^{(t)}\|_F^2\right) + O(\rho^2 L).\end{aligned}$$

This completes the proof. □

Proof of Theorem 3

Proof. For simplicity, let $\bar{s} := \frac{1}{K}\sum_{i=1}^{K}\|\mathcal{G}_k^{(t)}\|_F^2$ and $\bar{\gamma} := \frac{1}{K}\sum_{i=1}^{K}\|g_k^{(t)}\|_F^2$. The dynamics of Q can be expressed as:

$$\begin{aligned}\frac{dQ}{dt} &= \frac{d}{dt}\sum_{k=1}^{K}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \bar{s}\right)^2 \\ &= \sum_{k=1}^{K} 2\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \bar{s}\right)\frac{d}{dt}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \bar{s}\right) \\ &= 4\rho u^{(t)}\sum_{k=1}^{K}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \bar{s}\right)\left(\|g_k^{(t)}\|_F^2 - \bar{\gamma}\right) + O(\rho^2 L) \\ &= 4\rho u^{(t)} K \cdot \mathrm{Cov}\left(\|\mathcal{G}_k^{(t)}\|_F^2, \|g_k^{(t)}\|_F^2\right) + O(\rho^2 L),\end{aligned}$$

where the third equality follows from Lemma A.2. □

Proposition A.1 and proof

Proposition A.1 (Local Pairwise Norm Shrinkage in General Scale-invariant Models)**.** *We suppose that $\|g_i\|_F^2 > 0$, $\|g_j\|_F^2 > 0$, and small enough $\rho > 0$. Under SAM gradient flow, the pairwise norm gap $\mathcal{B}_{ij} := \left|\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2\right|$ satisfies:*

$$\frac{d}{dt}\mathcal{B}_{ij} < 0 \quad \textit{whenever } \mathcal{B}_{ij} > \bar{\mathcal{B}}_{ij}^{(\rho)} := \left(\bar{\alpha}^2 + \delta\right) - \frac{1}{\bar{\alpha}^2 + \delta},$$

where

$$\bar{\alpha}^2 := \sqrt{\frac{C_i}{C_j}},\ \delta = \mathcal{O}(\rho),$$

for some constant C_i, C_j.

Proof. Without loss of generality, suppose that $\|\mathcal{G}_i\|_F^2 > \|\mathcal{G}_j\|_F^2$ and $\mathcal{B}_{ij} = \|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2$. By Theorem 2, the evolution of the pairwise norm difference is given by:

$$\begin{aligned}&\frac{d}{dt}\left(\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2\right) \\ &\quad = 2\rho u^{(t)}\left(\|g_i\|_F^2 - \|g_j\|_F^2\right) + O(\rho^2 L).\end{aligned}$$

Under the re-parameterization $\mathcal{G}_i = \alpha\bar{\mathcal{G}}_i, \mathcal{G}_j = \bar{\mathcal{G}}_j/\alpha$ with $\|\bar{\mathcal{G}}_i\|_F = \|\bar{\mathcal{G}}_j\|_F = G$, we have:

$$\|\mathcal{G}_i\|_F^2 = \alpha^2 G^2, \quad \|\mathcal{G}_j\|_F^2 = \frac{G^2}{\alpha^2} \Rightarrow \mathcal{B}_{ij} = (\alpha^2 - \frac{1}{\alpha^2})\cdot G^2.$$

Since $\Phi(\cdot)$ is multilinear and the loss f is fixed under this re-parameterization, the gradients scale accordingly:

$$\|g_i\|_F^2 = \frac{C_i}{\alpha^2}, \quad \|g_j\|_F^2 = \alpha^2 C_j,$$

for constants $C_i = \|\nabla_{\bar{\mathcal{G}}_i} f\|_F^2$ and $C_j = \|\nabla_{\bar{\mathcal{G}}_j} f\|_F^2$.

We now show that the leading-order term in the SAM flow dominates the higher-order correction. Define

$$\epsilon := \frac{C_i}{\alpha^2} - \alpha^2 C_j.$$

Then:

$$\frac{d\mathcal{B}_{ij}}{dt} = 2\rho u^{(t)}\epsilon + \mathcal{O}(\rho^2 L).$$

If $\epsilon < -c\rho L$ for some constant $c > 0$, the first-order term dominates and we obtain:

$$\frac{d\mathcal{B}_{ij}}{dt} < 0.$$

This occurs whenever $\alpha^2 > \bar{\alpha}^2 := \sqrt{C_i/C_j} + \delta$ for some small margin $\delta = \mathcal{O}(\rho)$. In that case,

$$\mathcal{B}_{ij} > (\bar{\alpha}^2 - \frac{1}{\bar{\alpha}^2}) \cdot G^2 \Rightarrow \frac{d\mathcal{B}_{ij}}{dt} < 0.$$

which completes the proof. □

Results for Norm Deviations of multi-layer models

We extend the results in the main part of the paper following (Li, Zhang, and He 2024) and let $l \in [D]$ be the layer index. Denote the multilinear reconstruction function of l-th layer be Φ_l and the corresponding tensor core set be $\{\mathcal{G}_{1,l}^{(t)}, \dots, \mathcal{G}_{K_l,l}^{(t)}\}$. Then the update of SAM for layer l can be written as:

$$\begin{aligned} g_{k,l}^{(t)} &= \nabla_{\mathcal{G}_{k,l}} f(\{\Phi_l(\mathcal{G}_{1,l}^{(t)}, \dots, \mathcal{G}_{K_l,l}^{(t)})\}_l), \\ \tilde{\mathcal{G}}_{k,l}^{(t)} &= \mathcal{G}_{k,l}^{(t)} + \rho u_D^{(t)} g_{k,l}^{(t)}, \\ \tilde{g}_{k,l}^{(t)} &= \nabla_{\mathcal{G}_{k,l}} f(\{\Phi_l(\tilde{\mathcal{G}}_{1,l}^{(t)}, \dots, \tilde{\mathcal{G}}_{K_l,l}^{(t)})\}_l), \\ \mathcal{G}_{k,l}^{(t+1)} &= \mathcal{G}_{k,l}^{(t)} - \eta \tilde{g}_{k,l}^{(t)}, \end{aligned} \tag{2}$$

where $u_D^{(t)} = (\sum_{l=1}^{D} \sum_{k=1}^{K_l} \|\}_{k,l}^{(t)} \|_F^2)^{-1/2}$ is the normalization factor. We use the following assumption:

Assumption A.1 (Layer-wise Smoothness)**.** *There exists $\hat{L} > 0$ such that for any real tensors $\mathcal{X}_l, \mathcal{Y}_l$ having the same shape as $\Phi_l(\mathcal{G}_{1,l}^{(t)}, \dots, \mathcal{G}_{K_l,l}^{(t)})$ for all l, it holds that*

$$\|\nabla_l f(\mathcal{X}_l) - \nabla_l f(\mathcal{Y}_l)\|_F \le \hat{L} \|\mathcal{X}_l - \mathcal{Y}_l\|_F,$$

where $\nabla_l f$ is the gradient on $\mathcal{X}_l$.

Theorem 2 can be extended to

Theorem A.1 (Layer-wise Pairwise Norm Dynamics under SAM)**.** *Applying the update steps* (2) *with infinitesimal stepsize $\eta \to 0$, the gradient flow of SAM satisfies that $\forall i, j \in [K]$ with $i \neq j$:*

$$\frac{d}{dt}\left(\|\mathcal{G}_i^{(t)}\|_F^2 - \|\mathcal{G}_j^{(t)}\|_F^2\right) = 2\rho u_D^{(t)} \left(\|g_i^{(t)}\|_F^2 - \|g_j^{(t)}\|_F^2\right) + O(\rho^2 \hat{L}).$$

Proof. Using similar arguments in the proof of Theorem 2, $\forall i, j \in [K_l], l \in [D]$ with $i \neq j$ we have

$$\begin{aligned} &\frac{d}{dt}\left(\|\mathcal{G}_{i,l}^{(t)}\|_F^2 - \|\mathcal{G}_{j,l}^{(t)}\|_F^2\right) \\ &= 2\rho u_D^{(t)} \left(\|g_{i,l}^{(t)}\|_F^2 - \|g_{j,l}^{(t)}\|_F^2\right) \\ &\quad + \underbrace{2\rho u_D^{(t)} \left(\left\langle g_{i,l}^{(t)}, \tilde{g}_{i,l}^{(t)} - g_{i,l}^{(t)} \right\rangle_F - \left\langle g_{j,l}^{(t)}, \tilde{g}_{j,l}^{(t)} - g_{j,l}^{(t)} \right\rangle_F\right)}_{=:R_{ij,l}^{(t)}}. \end{aligned}$$

To bound the term $R_{ij,l}^{(t)}$, we bound the following term:

$$\begin{aligned} 2\rho u_D^{(t)} \left\langle g_{i,l}^{(t)}, \tilde{g}_{i,l}^{(t)} - g_{i,l}^{(t)} \right\rangle_F &\le 2\rho \|\tilde{g}_{i,l}^{(t)} - g_{i,l}^{(t)}\|_F \\ &\le 2\rho \hat{L} \|\tilde{\mathcal{T}}_l^{(t)} - \mathcal{T}_l^{(t)}\|, \end{aligned}$$

where the second $\le$ uses Assumption A.1 and we denote $\mathcal{T}_l^{(t)} = \Phi(\mathcal{G}_{1,l}^{(t)}, \dots, \mathcal{G}_{K_l,l}^{(t)})$, $\tilde{\mathcal{T}}_l^{(t)} = \Phi(\tilde{\mathcal{G}}_{1,l}^{(t)}, \dots, \tilde{\mathcal{G}}_{K_l,l}^{(t)})$. Define $\Delta_{k,l} = \tilde{\mathcal{G}}_{k,l}^{(t)} - \mathcal{G}_{k,l}^{(t)}$. We have that $\|\Delta_{k,l}\|_F = \rho u_D^{(t)} \|g_{k,l}^{(t)}\|_F = O(\rho)$. We can expand

$$\begin{aligned} \tilde{\mathcal{T}}_l^{(t)} - \mathcal{T}_l^{(t)} = &\underbrace{\sum_{k=1}^{K_l} \Phi(\mathcal{G}_{1,l}^{(t)}, \dots, \Delta_{k,l}, \dots, \mathcal{G}_{K_l,l}^{(t)})}_{=:A_l^{(1)}} \\ &+ \underbrace{\sum_{1 \le i < j \le K_l} \Phi(\dots, \Delta_{i,l}, \dots, \Delta_{j,l}, \dots)}_{=:A_l^{(2)}} \\ &+ A_l^{(>2)}, \end{aligned}$$

where $A_l^{(>2)}$ collects all terms involving at least three perturbations. The norm of $A_l^{(1)}$ can be bounded as follows:

$$\begin{aligned} \|A_l^{(1)}\|_F &\le \sum_{k=1}^{K_l} \|\Phi(\mathcal{G}_{1,l}^{(t)}, \dots, \Delta_{k,l}, \dots, \mathcal{G}_{K_l,l}^{(t)})\|_F \\ &\le \sum_{k=1}^{K_l} \left(\prod_{j \neq k} \|\mathcal{G}_{j,l}\|_F\right) \cdot \|\Delta_{k,l}\|_F \\ &= O(\rho). \end{aligned}$$

For the term $A_l^{(2)}$, we have:

$$\begin{aligned} \|A_l^{(2)}\|_F &\le \sum_{1 \le i < j \le K_l} \|\Phi(\dots, \Delta_{i,l}, \dots, \Delta_{j,l}, \dots)\|_F \\ &\le \sum_{1 \le i < j \le K_l} \left(\prod_{k \neq i,j} \|\mathcal{G}_{k,l}\|_F\right) \cdot \|\Delta_{i,l}\|_F \cdot \|\Delta_{j,l}\|_F \\ &= O(\rho^2). \end{aligned}$$

Similarly, higher-order terms are all $O(\rho^k)$ with $k \ge 2$, and we can bound:

$$\begin{aligned} \|\tilde{\mathcal{T}}_l^{(t)} - \mathcal{T}_l^{(t)}\|_F &\le \|A_l^{(1)}\|_F + \|A_l^{(2)}\|_F + \|A_l^{(>2)}\|_F \\ &\le O(\rho) + O(\rho^2) + O(\rho^3) + \cdots \\ &= O(\rho). \end{aligned}$$

Thus, we have:

$$\|\tilde{g}_{i,l}^{(t)} - g_{i,l}^{(t)}\|_F = O(\rho \hat{L}).$$

Combining the bounds, we have:

$$\begin{aligned} R_{ij,l}^{(t)} &\le 2\rho \|\tilde{g}_{i,l}^{(t)} - g_{i,l}^{(t)}\|_F + 2\rho \|\tilde{g}_{j,l}^{(t)} - g_{j,l}^{(t)}\|_F \\ &= O(\rho^2 \hat{L}). \end{aligned}$$

The proof is complete. □

Then we can extend to a multi-layer version of Theorem 3. We define layer-wise Norm Deviation $Q_l := \sum_{k=1}^{K_l} \left(\|\mathcal{G}_{k,l}\|_F^2 - \frac{1}{K_l} \sum_{i=1}^{K_l} \|\mathcal{G}_{i,l}\|_F^2 \right)^2$.

Theorem A.2 (Norm Deviation Dynamics under SAM)**.** *Applying the update steps* (2) *with infinitesimal stepsize* $\eta \to 0$*, the gradient flow of SAM satisfies:*

$$\frac{dQ_l}{dt} = 4\rho u_D^{(t)} K_l \cdot \mathrm{Cov}\left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2, \|g_{k,l}^{(t)}\|_F^2 \right) + O(\rho^2 \hat{L}).$$

Proof. For simplicity, let $\bar{s} := \frac{1}{K} \sum_{i=1}^{K_l} \|\mathcal{G}_{k,l}^{(t)}\|_F^2$ and $\bar{\gamma} := \frac{1}{K_l} \sum_{i=1}^{K_l} \|g_{k,l}^{(t)}\|_F^2$. The dynamics of Q can be expressed as:

$$\begin{aligned}
\frac{dQ_l}{dt} &= \frac{d}{dt} \sum_{k=1}^{K_l} \left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2 - \bar{s} \right)^2 \\
&= \sum_{k=1}^{K_l} 2 \left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2 - \bar{s} \right) \frac{d}{dt} \left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2 - \bar{s} \right) \\
&= 4\rho u_D^{(t)} \sum_{k=1}^{K_l} \left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2 - \bar{s} \right) \left(\|g_{k,l}^{(t)}\|_F^2 - \bar{\gamma} \right) + O(\rho^2 \hat{L}) \\
&= 4\rho u_D^{(t)} K_l \cdot \mathrm{Cov}\left(\|\mathcal{G}_{k,l}^{(t)}\|_F^2, \|g_{k,l}^{(t)}\|_F^2 \right) + O(\rho^2 \hat{L}),
\end{aligned}$$

where the third equality holds after summing the results of Theorem A.1. □

B Experiment details

All experiments are implemented with `torch` (Paszke et al. 2019) library in python. The tensor completion experiments are done using CPU AMD EPYC 7413 (24C/48T, 2.65GHz, 128M cache). The experiments on tensorized ResNet-20/-32, FLoRA fine-tuning, and some of LoRETTA fine-tuning tasks are performed on either single NVIDIA RTX A5000 with 24 gigabytes (GB) of GPU memory (VRAM) or single NVIDIA RTX A6000 with 48 GB VRAM. ImageNet experiments are run on 7 NVIDIA RTX A5000's using multiple GPU recipes. Full fine-tuning experiments of LoRETTA are done on a single NVIDIA H100 Tensor Core GPU with 80GB VRAM when the VRAM requirements exceed 48 GB.

Tucker Decomposition for Tensor Completion

Details for the real-world data COVID dataset can be found in the directory https://tensorly.org/stable/modules/generated/tensorly.datasets.load_covid19_serology.html.

The data set is formatted as a three-mode tensor of samples, antigens, and receptors with shape $(438, 6, 11)$. We fix $\rho = 0.01$ for SAM and $\alpha = 0.001$ for DAS.

Metric. We use R^2 metrics.

$$R^2 = 1 - \frac{SSE}{SST}$$

where:

- SSE is the Sum of Squared Errors (also called the residual sum of squares). It represents the unexplained variance. The formula is: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Here, y_i is the actual observed value and $\hat{y}_i$ is the value predicted by the model.

Method	HOOI	ADAM
R^2 score	$0.9268_{\pm 0.0063}$	$0.9482_{\pm 0.0012}$
Method	SAM	DAS
R^2 score	$0.9485_{\pm 0.0013}$	$0.9484_{\pm 0.0017}$

Table 1: Results of Tucker on COVID dataset.

- SST is the Total Sum of Squares. It represents the total variance in the data. The formula is: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ Here, $\bar{y}$ is the mean of the observed data.

Optimizers. We use a base optimizer, ADAM. The hyperparameter for ADAM is the default setting of torch, *i.e.*, `lr=0.001, betas=(0.9, 0.999), weight_decay=0.01`. Each Tucker decomposition is optimized for 50000 iterations.

Experiment results with standard deviation. Due to space limitations, we did not show the standard deviation. Here we show the results with standard deviation in Table 1.

Training Tensorized Neural Networks from Scratch

A cosine annealing scheduler is applied with an initial learning rate of 0.1 and batch size 128. We use a weight decay of 0.0005 and a momentum of 0.9. Hyperparameter optimization is done using trials trained with 80 epochs. We have used CP and Tensor-Ring (TR) in the tensorized ResNet-20. Convolution layer weights are inherently 4-order tensors, and we use 4-order CP decompositions to parameterize the layer. For Tensor-Ring, we follow existing works (Wang et al. 2018; Li et al. 2022; Cao et al. 2024) and use the tensorization as in Table 2, to reshape the 4-order convolution layer weights to 5-order or 7-order tensors and use TR cores to parameterize layers. The uncompressed baseline for ResNet-32 on CIFAR-10 is 91.65.

As discussed in the main paper, for both SAM and DAS, the corresponding hyperparameters ρ and α are tuned independently over the shared search space $\{0.001, 0.005, 0.01, 0.05, 0.1\}$ using a validation split from the training set. We use a fixed validation split with 5000 samples from the train set of CIFAR-10. This results in the train/valid set having a size of 45000/5000. After the best hyperparameters (with the largest validation accuracy) are found, models are trained from scratch using the best choices using the whole 50000 training samples. Hyperparameter optimizations are implemented using Optuna (Akiba et al. 2019).

Tensorized Neural Networks under Label Noise

We use a ResNet-32 with each convolution layer tensorized and parameterized with Tensor-Ring using the tensorization shape in Table 2. We use a uniform rank of 15. This results in a TR-ResNet-32 with 0.20M parameters, compared to the original model with 0.46M parameters. We use $\rho = 0.05$ for SAM and $\alpha = 0.1$ for DAS. We use a cosine annealing decreasing scheduler with an initial stepsize of 0.1. Momentum

Layer	Tensorization	ResNet-32	-20
CONV	$(3, K^2, 4, 2, 2)$		
CONV	$(4, 2, 2, K^2, 4, 2, 2)$	$\times 10$	$\times 5$
CONV	$(4, 2, 2, K^2, 4, 4, 2)$		
CONV	$(4, 4, 2, K^2, 4, 4, 2)$	$\times 9$	$\times 4$
CONV	$(4, 4, 2, K^2, 4, 4, 4)$		
CONV	$(4, 4, 4, K^2, 4, 4, 4)$	$\times 9$	$\times 4$
FC	Not compressed		

Table 2: The tensorization and details of ResNet model in TR formats. The numbers of blocks in ResNet-32 and ResNet-20 are shown. K is the kernel size and $K = 3$ in ResNets. We keep the final output layer uncompressed.

Layer	**TT-ranks**
`layer1.0.conv1`	`[1, 64, 64, 1]`
`layer1.0.conv2`	`[1, 64, 64, 1]`
`layer1.1.conv1`	`[1, 64, 64, 1]`
`layer1.1.conv2`	`[1, 64, 64, 1]`
`layer2.0.conv1`	`[1, 120, 60, 1]`
`layer2.0.conv2`	`[1, 100, 100, 1]`
`layer2.1.conv1`	`[1, 100, 100, 1]`
`layer2.1.conv2`	`[1, 100, 100, 1]`
`layer3.0.conv1`	`[1, 200, 150, 1]`
`layer3.0.conv2`	`[1, 135, 135, 1]`
`layer3.1.conv1`	`[1, 135, 135, 1]`
`layer3.1.conv2`	`[1, 135, 135, 1]`
`layer4.0.conv1`	`[1, 320, 200, 1]`
`layer4.0.conv2`	`[1, 170, 170, 1]`
`layer4.1.conv1`	`[1, 170, 170, 1]`
`layer4.1.conv2`	`[1, 170, 170, 1]`

Table 3: TT-ranks for TT-ResNet-18.

is set as 0.9, models are trained for 200 epochs, and weight decay is 0.0001.

Finetuning Pre-trained Models after Compression. The pre-trained ResNet-18 is available in torchvision https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html. The TT decomposition ranks suggested in (Yin et al. 2021) is available in https://openaccess.thecvf.com/content/CVPR2021/supplemental/Yin_Towards_Efficient_Tensor_CVPR_2021_supplemental.pdf. Here we show the detail of ranks in Table 3. Specifically, the original weight tensor of size $O \times I \times K \times K$ is reshaped and transposed to size $O \times K^2 \times I$, and then parameterized using TT decompositions. We use $\rho = 0.05$ for SAM and $\alpha = 0.01$ for DAS. We use a batch size of 2048, a peak stepsize of 0.008 with a one-cycle scheduler (Smith and Topin 2019), a weight decay of 0.00001, and a momentum of 0.9.

Finetuning with FLoRA

FLoRA can be used for the adaptation of high-order tensors, such as convolution layer weights. Since our experiments require only tuning linear layers, we introduce only the 2-order case. For a pre-trained linear layer with shape $O \times I$, FLoRA (Si et al. 2025) models the update $\Delta W = AGB^\intercal$ with $A \in \mathbb{R}^{O \times r}$, $G \in \mathbb{R}^{r \times r}$, and $B \in \mathbb{R}^{I \times r}$, resulting in a three-core tensorized model. FLoRA requires $r \times r$ extra parameters compared to LoRA with the sa.me rank, which is significantly small with small r. We experimented on RoBERTa-large (Liu et al. 2019). In the LoRA and FLoRA experiments, the target modules of fine-tuning are query linear and value linear layers.

Hyperparameter	Value
Number of iterations	1000
Batch size	16
Stepsize	0.0001, 0.0003, 0.0005
LoRA/FLoRA rank	8
LoRA/FLoRA α	16
α for DAS	0.1, 0.5, 0.7
ρ for SAM	0.001, 0.01, 0.1

Table 4: Hyperparameters for fine-tuning RoBERTa-large.

Our codebase is heavily built on (Malladi et al. 2023). For the detailed scripts for building the datasets, please refer to https://github.com/princeton-nlp/MeZO/tree/main/medium_models. The hyperparameters for our experiments are summarized in Table 4. We report the test accuracy on the model trained on the best hyperparameter with the best validation performance for each few-shot datasets. Full fine-tuning baseline uses a batch size of 16, and a stepsize from $\{0.00001, 0.00003, 0.00005\}$.

Finetuning with LoRETTA

Specifically, LoRETTA first reshapes the weight matrix with shape $O \times I$ to $O_1 \times \cdots \times O_n \times I_1 \times \cdots \times I_m$, where $O = O_1 \times \cdots \times O_n$ and $I = I_1 \times \cdots \times I_m$, and then parameterizes the high-order tensor with TT cores. The hidden size of OPT-6.7B is 4096, and we use a TT decomposition with shape $[4, 4, 16, 16, 16, 16, 4, 4]$ with rank 16 to parameterize the update for the 4096×4096 query and value layer matrices. Our code base is built upon the repository of LoRETTA https://github.com/yifanycc/loretta.

We fine-tune OPT-6.7B (Zhang et al. 2022), an autoregressive language model with 6.7B parameters on the SuperGLUE tasks (CB, BoolQ, WSC, COPA, ReCoRD) (Wang et al. 2019) and generation tasks including SQuAD (Rajpurkar et al. 2016) and DROP (Dua et al. 2019). Note that BoolQ and COPA use the accuracy metric, and the others use the F1 metric. We run all experiments for 3 epochs. Hyperparameters are summarized in Table 5. Note that we use a batch size of 8 for COPA and WSC, 2 for BoolQ, and 1 for the other datasets. Except for the stepsize given in the Table 5, we use ADAM as base optimizers with default settings, *i.e.*, `betas=(0.9, 0.999), weight_decay=0.01`.

Hyperparameter	Value
Stepsize	0.0001, 0.0003, 0.0005
LoRA/LoRETTA rank	16
LoRA/FLoRA α	16
α for DAS	0.5, 1, 2
ρ for SAM	0.05, 0.1, 0.2

Table 5: Hyperparameters for fine-tuning OPT-6.7B.

C Limitations and Future Work

A limitation of our theoretical analysis is the simplification of the optimization dynamics. The provided theorems, particularly Corollary 1, are based on a standard gradient flow assumption for SGD, which predicts that the Norm Deviation is conserved over time

However, our experiments—specifically those involving training from scratch and fine-tuning compressed models—utilize SGD with momentum for improved convergence. The introduction of a momentum term means the optimizer update is influenced by an accumulated velocity of past gradients, not just the gradient at the current step. Consequently, the Norm Deviation Q is not strictly preserved in our empirical settings, creating a small gap between the theoretical model and practical implementation.

A rigorous analysis of the Norm Deviation dynamics under momentum would require modeling the system with a second-order ordinary differential equation, often referred to as a "heavy-ball" analysis. Such an investigation exceeds the scope of this paper but represents a crucial direction for future work. This would provide a more complete understanding of how implicit regularization manifests in the presence of more complex, practical optimizers.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Cao, T.; Sun, L.; Nguyen, C. H.; and Mamitsuka, H. 2024. Learning low-rank tensor cores with probabilistic L0-regularized rank selection for model compression. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 3780–3788.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. 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)*, 2368–2378.

Li, B.; Zhang, L.; and He, N. 2024. Implicit regularization of sharpness-aware minimization for scale-invariant problems. *Advances in Neural Information Processing Systems*, 37: 44444–44478.

Li, N.; Pan, Y.; Chen, Y.; Ding, Z.; Zhao, D.; and Xu, Z. 2022. Heuristic rank selection with progressively searching tensor ring network. *Complex & Intelligent Systems*, 8(2): 771–785.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Malladi, S.; Gao, T.; Nichani, E.; Damian, A.; Lee, J. D.; Chen, D.; and Arora, S. 2023. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36: 53038–53075.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Si, C.; Wang, X.; Yang, X.; Xu, Z.; Li, Q.; Dai, J.; Qiao, Y.; Yang, X.; and Shen, W. 2025. Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Wang, W.; Sun, Y.; Eriksson, B.; Wang, W.; and Aggarwal, V. 2018. Wide compression: Tensor ring nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9329–9338.

Yin, M.; Sui, Y.; Liao, S.; and Yuan, B. 2021. Towards efficient tensor decomposition-based dnn model compression with optimization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10674–10683.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.