Cut Less, Fold More: Model Compression through the Lens of Projection Geometry

Olga Saukh[1,2] Dong Wang[1] Haris Šikić[1] Yun Cheng[3] Lothar Thiele[4]

[1]Graz University of Technology [2]Complexity Science Hub [3]Swiss Data Science Center [4]ETH Zurich

Motivation

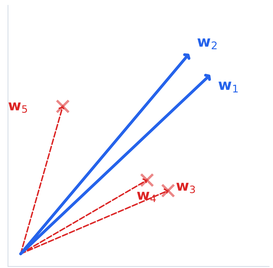
- Neural network compression is critical for deployment at scale.
- **Structured pruning** (the dominant approach) removes neurons/channels entirely — this is an **axis-aligned projection** that throws away directional information.
- **Model folding** (recently introduced) clusters similar weights and replaces each cluster with its mean — this is a **cluster-structured projection** that preserves directional information.
- **Key question**: Can we formally compare these two compression strategies through the lens of projection geometry?

Key Contributions

1. **Unified projection framework:** Pruning = axis-aligned projection; Folding = cluster-structured projection. Both are orthogonal projections onto fundamentally different subspaces.
2. **Provable dominance:** Folding achieves smaller reconstruction error (Theorems 2.1 & 2.2) with negligible rank slack (<1%).
3. **Large-scale validation:** >1,000 checkpoints across CNNs, ViTs, and LLMs on CIFAR-10, ImageNet-1K, and C4.
4. **Ablation insights:** FOLD excels at moderate–high compression and flatter landscapes; gap narrows at extremes.

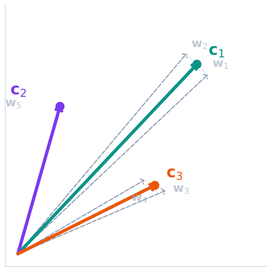
Structured Pruning VS Model Folding

Structured Pruning
Axis-aligned projection, rank k=2



Keep top-k neurons by $\|\mathbf{w}_i\|$, set remaining $m-k$ to $\mathbf{0}$.
$P_{\text{prune}} = \text{diag}(\mathbf{s})$, $s_i \in \{0, 1\}$

Model Folding
Cluster-structured projection, rank k+1=3



k-means clusters all m neurons into k+1 groups, each $\mathbf{w}_i \mapsto$ cluster mean $\mathbf{c}_{a(i)}$.
$P_{\text{fold}} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$

$\|\mathbf{W} - \hat{\mathbf{W}}\|_F = 1.262 \gg \|\mathbf{W} - \tilde{\mathbf{W}}\|_F = 0.127$ (90% less error with only +1 rank)

Kept (pruning) Pruned → **0** Original (faded) Centroids (folding)

Theory and Key Results

Theorem 2.1
For any pruning of rank k_p (at least one vector pruned), there exists a folding of rank $k_f = k_p + 1$ such that:

$$\|\mathbf{W} - \mathbf{W}_p\|_F^2 \geq \|\mathbf{W} - \mathbf{W}'_f\|_F^2.$$

Theorem 2.2
Optimal k-means folding with k_f clusters always satisfies:

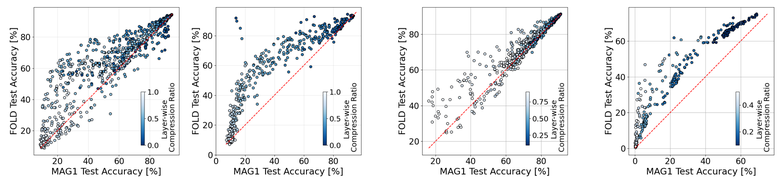
$$\|\mathbf{W} - \mathbf{W}_p\|_F^2 \geq \|\mathbf{W} - \mathbf{W}_f^*\|_F^2.$$

for any pruning of rank $k_p = k_f - 1$.

Key implication:

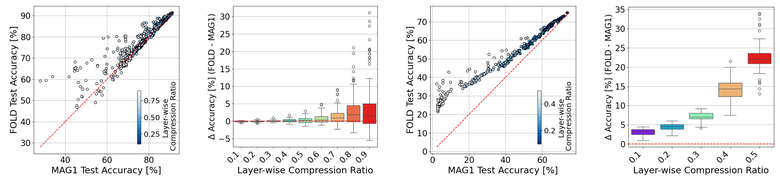
- The rank difference $k_f = k_p + 1$ is **negligible** in practice (e.g., 0.78% for ResNet-18 at 50% compression, 0.26% for ViT-B/32)
- Via Lipschitz continuity of the loss: smaller Frobenius error → tighter bound on loss perturbation
- Ordering: error($\mathbf{W}_p$) ≥ error($\mathbf{W}'_f$) ≥ error($\mathbf{W}_f^*$) — pruning always worst, optimal folding always best

Main Empirical Results — Folding vs Pruning



ResNet18 PreActResNet ViT CLIP

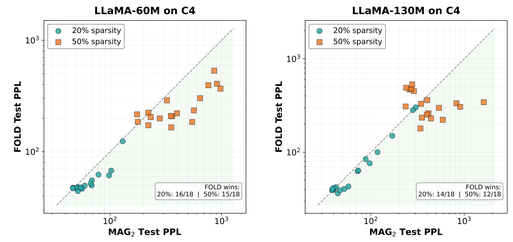
After LayerNorm-only fine-tuning:



ViT-B/32 CLIP ViT-B/32

MAG_1 vs FOLD after LayerNorm-only fine-tuning. Bar plots show accuracy gain $\Delta = \text{Acc}(\text{FOLD}) - \text{Acc}(\text{MAG}_1)$, which is positive — FOLD retains advantage even under LayerNorm adaptation.

LLM Results



18 LLaMA models (60M and 130M) trained on C4 with diverse learning rates, warmup, and weight decay. FFN blocks compressed at 20% and 50% sparsity.

Key findings:

- FOLD wins in **16/18** (60M) and **14/18** (130M) at 20% sparsity
- Advantage grows with sparsity and learning rate
- At low LR with long warmup, MAG_2 can match FOLD

Ablation Studies — What Training Conditions Favor Folding?

>1'000 checkpoints spanning ResNet18, PreActResNet18, ViT-B/32, and CLIP ViT-B/32

Factor	Effect on FOLD advantage
Learning rate	FOLD leads at moderate-low LR; gap shrinks/reverses at very high LR
SAM training	SAM lifts both methods; FOLD benefits more (especially with Adam)
Data augmentation	On ResNets: narrows gap. On ViTs: **increases** FOLD advantage
Optimizer	Adam: larger variance, more pronounced FOLD advantage. SGD: tighter, smaller gap
Weight decay	Does not diminish FOLD advantage

Intuitive explanation (callout): Conditions that produce **flatter, more structured** weight landscapes (moderate LR, SAM) → **amplify** FOLD's advantage, because when weights are well-aligned, clustering reduces projection error more than coordinate removal.

Code & Reproducibility

- Core Framework & Vision Models:
  `https://github.com/osaukh/folding_as_projection`
- LLM Compression (LLaMA):
  `https://github.com/nanguoyu/simple_model_folding_public`

References

Olga Saukh, Dong Wang, Haris Šikić, Yun Cheng, and Lothar Thiele. Cut less, fold more: Model compression through the lens of projection geometry. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=JV9CEtKLQF`.

Dong Wang, Haris Šikić, Lothar Thiele, and Olga Saukh. Forget the data and fine-tuning! just fold the network to compress. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=W2Wkp9MQsF`.