STRUCTURED SAMPLE-WISE GRADIENTS IN NEURAL NETWORK OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent (SGD) aggregates gradients from individual training examples, yet the organization of these sample-wise optimization directions remains poorly understood. We analyze per-example gradients using pairwise cosine similarity and clustering throughout training. We find that gradient clusters are internally coherent, with highly aligned within-cluster gradients, while gradients from different clusters exhibit much lower similarity. This structure is reproducible across independently trained models and is stronger in deeper layers. Gradient clusters are strongly aligned with class labels early in training, but this correspondence weakens over time. Besides, lower within-cluster similarity is associated with better generalization performance, and pretrained models begin downstream training with lower within-cluster similarity than randomly initialized models. Our results reveal a structured sample-wise organization underlying SGD updates and connect its evolution to class structure, pretraining, and generalization.

1 INTRODUCTION

Stochastic gradient descent (SGD) updates a model by aggregating gradients induced by individual training examples. Existing analyses have largely characterized SGD through the lens of optimization trajectories or aggregate measures of gradient stochasticity, such as gradient noise (Simsekli et al., 2020; Wu et al., 2020; Smith et al., 2020). However, this aggregate perspective leaves the relationships among the individual examples contributing to each update less understood (Figure 1A). The geometry of sample-wise gradients determines how individual optimization signals combine: aligned gradients reinforce one another, whereas diverse or conflicting directions can partially cancel (Yin et al., 2018). Recent work has begun to examine per-example gradient information beyond the mini-batch average, including its statistical properties and role in optimization (Roulet & Agarwala, 2025). Yet, the structured organization of sample-wise optimization directions, and how this organization changes throughout training, remains insufficiently understood. Characterizing this structure may provide a finer-grained perspective on how individual examples collectively shape SGD updates and how their relationships evolve during learning.

We begin by asking whether sample-wise gradients exhibit a meaningful organization. We characterize this organization through the directional geometry of per-example gradients: for each example, we compute its gradient and measure pairwise cosine similarities across examples, layers, and training stages. The resulting similarity matrices reveal a clear block structure, which is further captured by gradient-based clustering. Examples within the same gradient cluster exhibit substantially higher directional alignment, often reaching cosine similarities above 0.7, whereas examples across clusters are much less aligned and can even exhibit negative cosine similarity. Thus, sample-wise gradients are neither uniformly aligned nor unstructured. Instead, they exhibit a clustered geometry in which distinct groups of examples induce coherent but different optimization signals that are subsequently aggregated by SGD.

A natural question is whether this organization is specific to a particular random initialization or optimization trajectory. We find that it is highly reproducible across independently trained models: gradient-based cluster assignments remain similar across runs with different random initializations. This reproducibility suggests that the observed organization reflects systematic relationships among training examples rather than being solely determined by a particular initialization or stochastic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

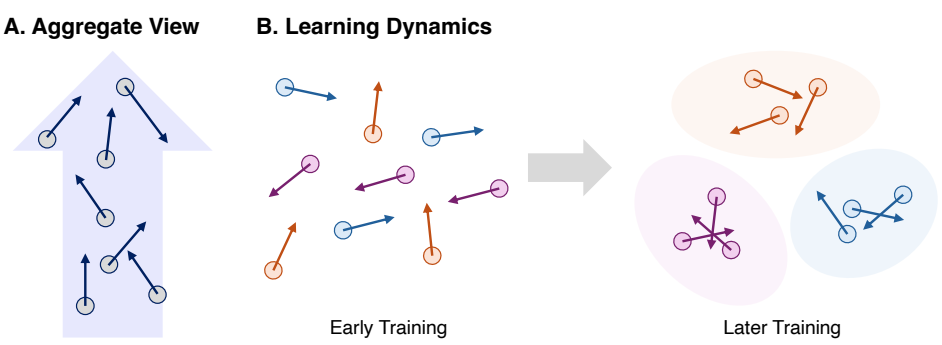

Figure 1: **Sample-wise gradients exhibit a structured and evolving geometry.** **(A)** Standard SGD is typically viewed through the aggregate gradient, which combines optimization signals from individual training examples and obscures their relationships. **(B)** Viewing gradients at the sample level reveals structured directional relationships that evolve during training: gradients are initially organized into coherent groups and progressively differentiate as optimization proceeds.

training trajectory. The result therefore indicates that sample-wise gradient geometry captures consistent properties of how different examples shape parameter updates.

We next ask what determines these relationships and how the gradient organization evolves during learning. Early in training, gradient clusters strongly align with class labels, particularly in deeper layers, indicating that class identity provides a coarse semantic organization of sample-wise optimization directions. However, this class-level correspondence becomes progressively weaker as training proceeds. The alignment between gradient clusters and class labels gradually decreases, accompanied by a reduction in directional similarity among examples within the same gradient cluster (Figure 1B). Learning therefore does not simply preserve or strengthen the initial class-level organization. Instead, the gradient clusters progressively differentiate, with examples within a cluster becoming less directionally homogeneous over training. This evolution is more pronounced in shallow layers, whereas deeper layers retain more stable gradient organization over time.

Finally, we ask whether this sample-wise differentiation is associated with the quality of the learned solution. Across training configurations, lower within-cluster gradient similarity is consistently associated with higher final test accuracy. Moreover, pretrained models begin downstream fine-tuning with lower within-cluster gradient similarity than randomly initialized models, indicating that differentiated sample-wise gradient geometry can already be present before task-specific fine-tuning. Together, these findings show that stronger generalization is associated with a more differentiated sample-wise gradient geometry rather than increasingly uniform optimization directions within a cluster.

Our contributions are as follows:

- We develop a layer-wise, time-resolved framework for characterizing the geometry of sample-wise gradients through pairwise directional similarity and gradient-based clustering.
- We show that sample-wise gradients form coherent and reproducible groups that initially reflect class structure but become increasingly differentiated throughout training.
- We find that stronger generalization is associated with lower within-cluster gradient similarity, and that pretrained models exhibit this differentiated gradient geometry at the onset of downstream fine-tuning.

2 Related Work

Gradient alignment and interference. Gradient alignment has been widely studied as a measure of compatibility between optimization objectives. In multi-task learning, prior work uses gradient alignment or conflict to characterize and mitigate interference between task updates (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021). Gradient diversity has also been linked to the convergence and generalization behavior of minibatch optimization (Yin et al., 2018), while gradient confusion

connects sample-gradient dispersion to optimization efficiency (Sankararaman et al., 2020). These studies establish the importance of directional relationships among gradients, but primarily focus on aggregate alignment, conflict, or diversity. We instead study the organization of individual examples in gradient space and how this organization evolves during training.

Per-example gradient analysis. Recent work has increasingly examined gradients at the level of individual training examples, including their role in optimizer analysis and modification (Roulet & Agarwala, 2025). Per-example gradients have also been used to quantify sample influence and value, such as through influence functions, TracIn, and gradient-based data valuation (Koh & Liang, 2017; Pruthi et al., 2020; Evans et al., 2024). These approaches primarily use gradients to characterize individual sample importance or to improve optimization. In contrast, we study the relational structure among examples, rather than assigning an importance score to individual gradients.

Gradient geometry and sample relationships. Gradient similarity has been used to characterize relationships between training examples, including their relevance to target data and their semantic structure (Evans et al., 2024). However, existing approaches typically treat gradient similarity as a tool for attribution, valuation, or sample comparison, rather than studying the organization of the gradient population itself. Our work explicitly characterizes this organization through gradient-based clustering, examines its reproducibility across independent training runs, and tracks its evolution across layers and training stages. We further relate these gradient structures to class identity and final generalization.

3 METHODOLOGY

3.1 SAMPLE-WISE GRADIENT GEOMETRY

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and a neural network f_θ, let $\theta^{(\ell)}$ denote the parameters of layer or block ℓ. We define the sample-wise gradient at layer ℓ as

$$g_i^{(\ell)}(\theta) = \nabla_{\theta^{(\ell)}} \ell(f_\theta(x_i), y_i). \tag{1}$$

To focus on directional relationships, we normalize each gradient as

$$\tilde{g}_i^{(\ell)}(\theta) = \frac{g_i^{(\ell)}(\theta)}{\|g_i^{(\ell)}(\theta)\|_2 + \epsilon}. \tag{2}$$

Unless otherwise stated, we analyze the final feature block before the classification head. Layer-wise analyses apply the same procedure independently to each block.

3.2 GRADIENT SIMILARITY AND CLUSTERING

For two examples i and j, we measure their directional similarity using cosine similarity,

$$s_{ij}^{(\ell)}(\theta) = \left\langle \tilde{g}_i^{(\ell)}(\theta), \tilde{g}_j^{(\ell)}(\theta) \right\rangle. \tag{3}$$

For a gradient cluster C_k, we define its within-cluster similarity as

$$S_{\text{within}}^{(\ell)}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{P}_k|} \sum_{(i,j)\in\mathcal{P}_k} s_{ij}^{(\ell)}(\theta), \tag{4}$$

where $\mathcal{P}_k = \{(i,j) : i < j,\ i,j \in C_k\}$. Lower values indicate greater differentiation among examples within the same gradient group. We analogously compute between-cluster similarity using pairs from different clusters.

To characterize the global organization of sample-wise gradients, we first apply PCA to the normalized gradients and retain the leading 150 principal components. This fixed dimensionality is used throughout the main-text analyses to improve computational efficiency and ensure comparability across layers, checkpoints, and random seeds. We then apply K-means clustering in the resulting PCA space, producing a partition $C^{(\ell)}(\theta) = \{C_1, \ldots, C_K\}$.

Unless otherwise stated, we set the number of clusters to $K = 10$ throughout the main-text analyses and keep it fixed across layers, checkpoints, and random seeds. PCA and clustering are performed independently at each checkpoint. To assess the robustness of our results, Appendix B and C reports additional analyses varying both the PCA dimensionality and the number of clusters. Specifically, in addition to the 150 principal components used in the main text, we consider 50 and 500 principal components, and we evaluate alternative cluster numbers of $K = 15$ and $K = 20$.

Adjusted Rand Index. We use the adjusted Rand index (ARI) to quantify agreement between two partitions of the same examples. ARI equals 1 for identical partitions and is approximately 0 under random agreement. We use ARI to measure cross-run agreement between independently trained models, temporal changes in cluster organization across training stages, and alignment between gradient-based clusters and class labels.

4 Experimental Results

We organize our experiments around a sequence of questions. We first establish whether per-example gradients form coherent groups and whether this organization is reproducible across independent training runs. We then examine how the organization evolves during training and to what extent it is explained by class labels. Finally, we study whether within-cluster gradient geometry is associated with generalization and whether it is systematically affected by pretrained initialization. Unless otherwise stated, gradients are clustered independently for each layer, checkpoint, and run using the procedure described in Section 3.

4.1 Sample-wise Gradients Form Reproducible Groups

We first ask whether per-example gradients exhibit a structured organization rather than being uniformly aligned or randomly distributed. Immediately after initialization, we compute the gradient induced by each training example and cluster the resulting gradient representations into 10 groups. As shown in Figure 2A, gradients within the same cluster exhibit strong directional alignment, with mean cosine similarities of approximately 0.8, whereas similarities between different clusters are substantially lower, ranging from near zero to approximately -0.4. Thus, the sample-wise gradient population exhibits a clear clustered geometry: examples within the same group induce similar optimization directions, while different groups induce substantially distinct directions.

We next ask whether these gradient groups are specific to a particular random initialization. We independently initialize two models, compute per-example gradients for the same training examples, and cluster the gradients from each model separately. Because the cluster indices are arbitrary across the two runs, we match clusters between the two partitions by maximizing their sample overlap using the Hungarian algorithm. Figure 2B shows the resulting overlap matrix. After matching, the diagonal entries dominate, indicating that the two independently initialized models consistently group many of the same training examples together.

We quantify this agreement using the ARI. As shown in Figure 2C, the clustering agreement is consistently high across layers. This result indicates that the reproducibility of sample-wise gradient organization is not uniform throughout the network: deeper representations give rise to more consistent gradient groupings across independent initializations.

Finally, Figure 2D visualizes the gradient geometry underlying these clusters. We project the final-layer per-example gradients into two dimensions using PCA and color each example according to its cluster assignment. The resulting visualization shows distinct regions corresponding to different gradient groups, providing a direct geometric view of the organization quantified in Figure 2A.

Together, these results show that sample-wise gradients form coherent and reproducible groups. The consistency of the grouping across independent random initializations suggests that the observed organization is not solely determined by a particular initialization, but reflects systematic relationships among training examples in gradient space.

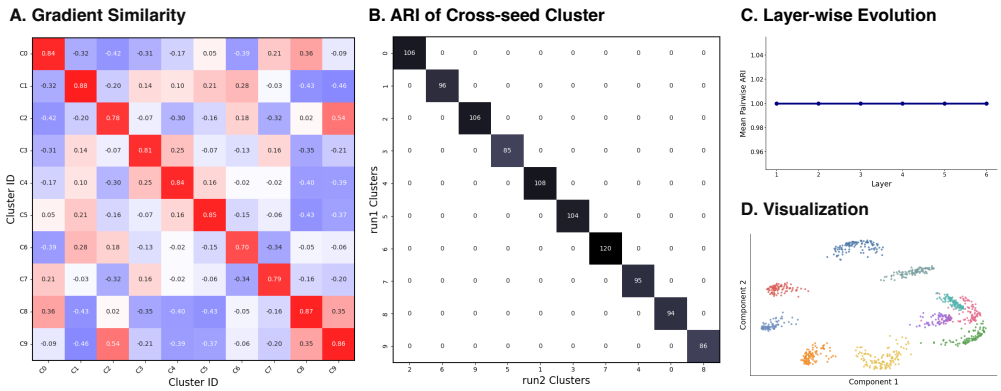

Figure 2: **Sample-wise gradient groups are coherent and reproducible across random initializations. (A)** Mean pairwise cosine similarity of per-example gradients within and between ten gradient clusters immediately after initialization. Gradients within the same cluster exhibit substantially higher directional similarity than gradients from different clusters. **(B)** Overlap matrix between gradient clusters obtained from two independently initialized models. Clusters are matched using the Hungarian algorithm to maximize sample overlap, resulting in the strongest correspondences along the diagonal. **(C)** ARI between the two clusterings across network layers, showing increasing cross-run agreement in deeper layers. **(D)** Two-dimensional PCA visualization of final-layer per-example gradients, with colors indicating cluster assignments. Distinct regions in gradient space correspond to the identified gradient groups.

4.2 Gradient Organization Evolves During Training

Having established that sample-wise gradient clusters are reproducible at initialization, we next examine how their organization changes during training. For each layer and epoch, we cluster the per-example gradients and quantify the similarity between clusterings using the ARI.

Figure 3A shows pairwise ARI values between gradient clusterings obtained at different epochs within the same training run. The clusterings from the early epochs are generally highly similar to one another. As training proceeds, however, the clustering structure gradually diverges from its early-training organization. In particular, the clusterings at later epochs show lower ARI values when compared with those from the early stages of training, while the agreement between late-training epochs also becomes lower. This indicates that training progressively reorganizes the groups of samples that induce similar gradient patterns, rather than maintaining a fixed partition throughout optimization.

The magnitude of this reorganization depends on network depth. The decline in ARI is more pronounced in shallow layers, whose gradient clusterings change substantially over the course of training. Gradient clusterings in deeper layers also become less similar as training progresses, but their organization remains considerably more stable: even at the end of training, the ARI between different epochs remains above approximately 0.5. Thus, deeper layers preserve a larger component of their initial sample-wise gradient organization.

We further evaluated whether this stability is reproducible across independently initialized models. Figure 3B presents pairwise ARI values between gradient clusterings obtained from two training runs with different random initializations, evaluated across training epochs. The cross-run agreement follows a pattern similar to the within-run comparison. Clusterings become progressively less similar as training proceeds, with the strongest reduction in shallow layers. In contrast, the deeper layers maintain substantially higher agreement across the two runs, despite starting from different initial conditions. These results suggest that training reorganizes sample-wise gradient structure in a depth-dependent manner: shallow-layer organization is more strongly shaped by the specific training trajectory, whereas deeper-layer organization remains comparatively stable and reproducible.

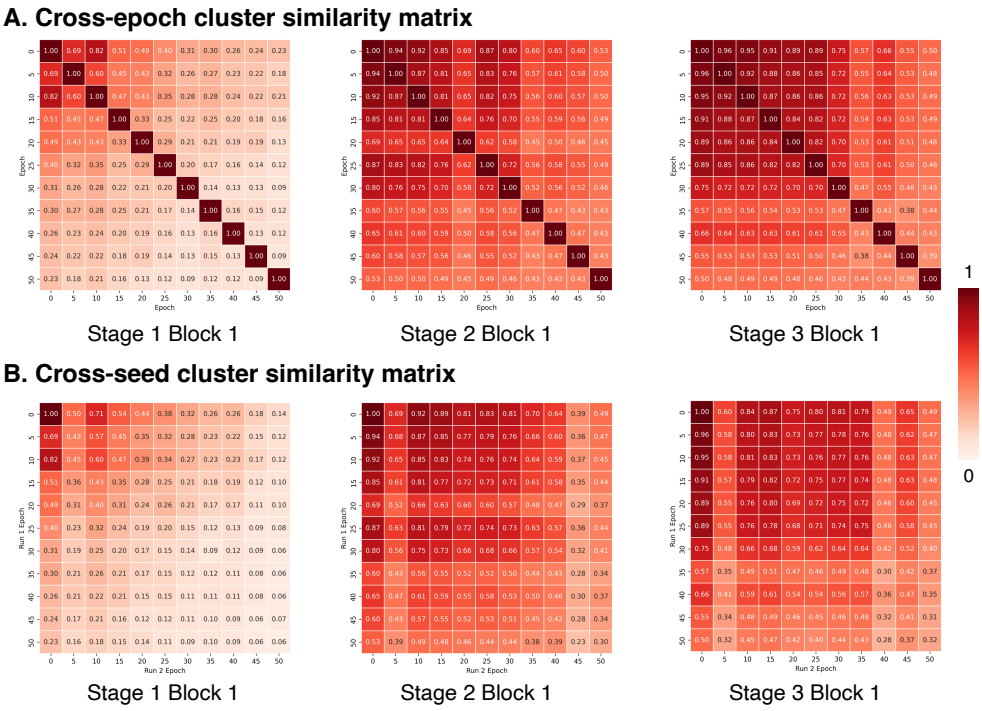

Figure 3: **Gradient organization progressively reorganizes during training, with greater stability in deeper layers. (A)** Pairwise ARI between gradient clusterings at different epochs within a training run, showing increasing temporal divergence during training. **(B)** Pairwise ARI between clusterings from two independently initialized runs across different epochs, showing reduced cross-run agreement as training progresses.

4.3 Class Labels Explain the Initial Gradient Organization

Having established that gradient-based groups are reproducible across independent initializations, we next ask whether they correspond to class structure. At each checkpoint, we compare the partition induced by sample-wise gradients with the partition defined by ground-truth labels using the ARI.

Figure 4A shows the sample overlap between the gradient-based and label-based partitions after matching clusters with the Hungarian algorithm. The strong diagonal correspondence at early training indicates that gradient clusters substantially overlap with class-defined groups. Figure 4B quantifies this correspondence across layers and training stages. Label alignment increases with network depth and is strongest in deeper layers. Across training, however, the ARI generally decreases, with a more pronounced reduction in shallow layers. Deeper layers also show a decline, but retain relatively high alignment, with late-training ARI remaining above approximately 0.5.

To examine whether the observed correspondence depends on the original assignment of labels, we repeat the analysis after randomly shuffling the labels across samples. As shown in Figure 4C, the alignment between gradient-based clusters and the shuffled-label partition remains substantial, particularly in deeper layers. Unlike the alignment with the original labels, this correspondence does not show a pronounced decrease over training. One possible explanation is that training with randomly assigned labels is more difficult, leading to different optimization dynamics and preserving the organization of gradient clusters for longer. Nevertheless, the persistence of strong alignment with the shuffled labels indicates that the organization of sample-wise gradients is not determined solely by the semantic class partition.

Taken together, these results show that gradient-based groups are strongly structured and become increasingly aligned with deeper network representations. Their alignment with the original class labels is already evident early in training and remains relatively strong in deeper layers. However,

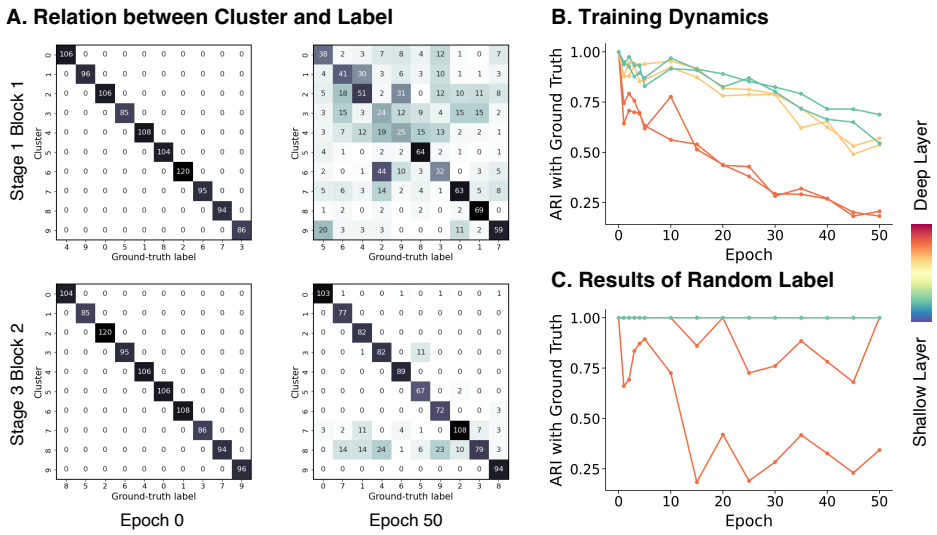

Figure 4: **Gradient organization is initially aligned with class structure but progressively departs from it during training.** **(A)** Sample overlap between gradient-based clusters and class-defined groups after Hungarian matching. **(B)** ARI between gradient-based and label-based partitions across layers and training epochs, showing stronger alignment in deeper layers and a gradual decrease during training. **(C)** Corresponding analysis using randomly shuffled class labels, providing a control for the observed label alignment.

comparable alignment can also be observed after labels are randomly shuffled, especially in deeper layers, and does not necessarily decay over training. This suggests that sample-wise gradient organization is determined to a substantial extent by the output targets and the resulting optimization dynamics, rather than reflecting only the semantic class structure present in the original labels.

4.4 Within-cluster Gradient Differentiation Is Associated with Generalization

The preceding results show that sample-wise gradients form structured groups that progressively reorganize during training. We next examine how gradient similarity within and between these groups evolves during training, and whether either aspect of this organization is associated with model generalization. To this end, we analyze 11 models spanning different architectures and training configurations.

Figure 5A shows the evolution of the mean within-cluster gradient cosine similarity during training. Within-cluster similarity progressively decreases at the beginning of training and then reaches a relatively stable level, indicating that examples assigned to the same gradient cluster become increasingly differentiated in their gradient directions. In contrast, the mean between-cluster gradient similarity increases gradually over training, but remains very low overall and close to zero (Figure 5B). Thus, the dominant change during training is the differentiation of gradients within clusters, rather than a substantial increase in similarity between different clusters.

We next examine the relationship between gradient similarity and generalization. For each model, we compute gradient similarities at the final training epoch and compare them with the model's final test accuracy. As shown in Figure 5C, lower within-cluster gradient similarity is associated with higher test accuracy across models with different architectures and training configurations. In contrast, between-cluster gradient similarity shows little relationship with final test accuracy (Figure 5D). These results suggest that generalization is more closely related to the internal differentiation of gradient clusters than to the similarity structure between different clusters.

Finally, we examine when this relationship emerges during training. Figures 5E and 5F show, for the training and test sets, respectively, the correlation between within-cluster gradient similarity at each checkpoint and the final test accuracy of the corresponding model. The negative association

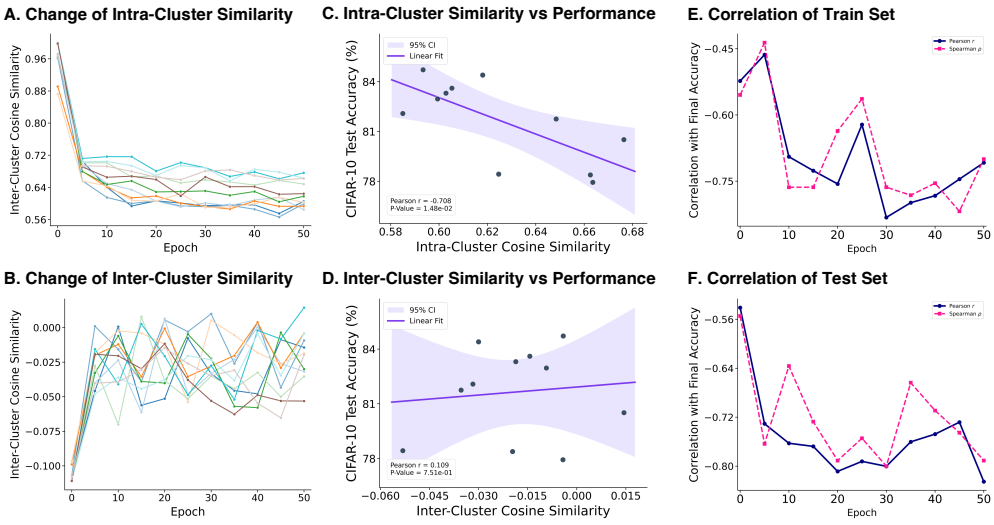

Figure 5: **Within-cluster gradient differentiation is associated with better generalization.** **(A)** Mean within-cluster gradient cosine similarity across training. Similarity progressively decreases and then reaches a relatively stable level. **(B)** Mean between-cluster gradient cosine similarity across training. Although it increases gradually, it remains low overall and close to zero. **(C)** Relationship between within-cluster gradient similarity at the final training epoch and final test accuracy across models with different architectures and training configurations. **(D)** Relationship between final-epoch between-cluster gradient similarity and final test accuracy. **(E–F)** Correlations between within-cluster gradient similarity at each training checkpoint and final test accuracy, computed on the training and test sets, respectively. A substantial negative association emerges after approximately 10 epochs, with correlation coefficients reaching approximately -0.7.

becomes evident after approximately 10 epochs and reaches a correlation coefficient of approximately -0.7. This relationship remains evident during subsequent training, indicating that models that ultimately generalize better tend to exhibit lower within-cluster gradient similarity relatively early in optimization.

Together, these results indicate that better generalization is associated with greater differentiation among the gradients of samples within the same cluster. This association is not explained by changes in between-cluster similarity, which remains close to zero throughout training. Moreover, within-cluster gradient similarity provides an early indicator of eventual generalization performance, with a substantial negative correlation emerging after only approximately 10 epochs of training.

4.5 Pretraining Changes the Initial Gradient Geometry

The association between within-cluster gradient geometry and generalization raises a further question: does pretraining already shape this geometry before downstream learning begins? To examine this, we compare pretrained and randomly initialized models at the start of downstream training.

Figure 6 shows that pretrained models consistently exhibit lower within-cluster gradient similarity than randomly initialized models at the initial downstream checkpoint. This difference is observed across network layers, indicating that pretraining systematically changes the organization of sample-wise optimization directions before substantial task-specific optimization takes place.

We test the robustness of this effect in two complementary settings. In Figure 6A, we fix the architecture (RoBERTa) and compare pretrained and randomly initialized models across multiple downstream tasks. In Figure 6B, we fix the downstream task (Yelp Dataset) and repeat the comparison across different architectures. In both settings, pretrained models show lower initial within-cluster gradient similarity, indicating more differentiated optimization directions from the onset of downstream training.

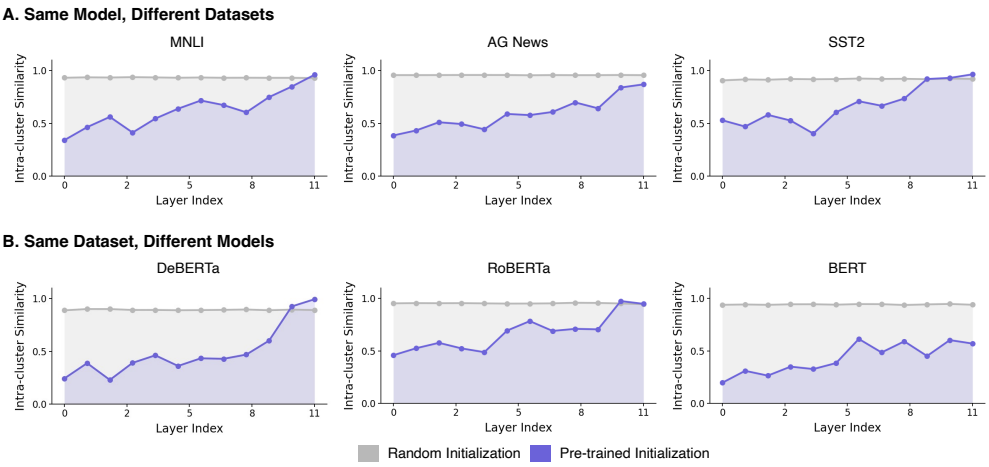

Figure 6: **Pretraining changes the initial sample-wise gradient geometry.** (A) Comparison of pretrained and randomly initialized models across multiple downstream tasks with a fixed architecture. (B) Comparison across model architectures on a fixed downstream task. Pretrained models consistently exhibit lower within-cluster gradient similarity at the start of downstream training, indicating more differentiated sample-wise optimization directions before task-specific fine-tuning.

Together with the generalization results in Section 4.4, these findings show that pretraining systematically shifts the initial gradient geometry toward the regime associated with stronger generalization. This result does not establish a causal role for gradient differentiation, but shows that the geometry of sample-wise optimization signals is already substantially shaped by the initialization inherited from pretraining.

5 Discussion and Conclusion

Our results show that sample-wise gradients form coherent and reproducible groups whose organization evolves throughout training. These groups are initially strongly aligned with class structure, but gradually reorganize as within-cluster gradient similarity decreases. This differentiation is associated with better generalization and is already evident in pretrained models before downstream fine-tuning, suggesting that initialization shapes the geometry of subsequent optimization signals.

More broadly, sample-wise gradient geometry provides a complementary view of optimization by characterizing how individual examples are organized according to their induced optimization directions. It connects class structure, training dynamics, initialization, and generalization within a common framework.

Limitations. Our analysis is computationally expensive and mainly focuses on supervised image classification and a limited range of architectures. The observed structure may also depend on choices of gradient representation and clustering methodology. Most importantly, although we characterize how gradient clusters reorganize during training, the meaning of the late-training clusters remains unclear. Whether they reflect learned features, finer-grained data structure, or optimization-specific effects requires further investigation. In addition, the relationship between within-cluster differentiation and generalization is correlational, and its underlying mechanism remains to be established.

Conclusion. We show that sample-wise gradients exhibit coherent, reproducible, and evolving geometry during neural network training. Their organization progressively differentiates beyond the initial class structure, with within-cluster differentiation associated with generalization. These findings highlight sample-wise gradient geometry as a useful perspective for understanding how individual examples contribute to learning.

REFERENCES

Nathaniel J Evans, Gordon B Mills, Guanming Wu, Xubo Song, and Shannon McWeeney. Data valuation with gradient similarity. *ArXiv*, pp. arXiv–2405, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. Pmlr, 2017.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in neural information processing systems*, 33: 19920–19930, 2020.

Vincent Roulet and Atish Agarwala. Per-example gradients: a new frontier for understanding and improving optimizers. *arXiv preprint arXiv:2510.00236*, 2025.

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *International conference on machine learning*, pp. 8469–8479. PMLR, 2020.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International conference on machine learning*, pp. 9058–9067. PMLR, 2020.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1998–2007. PMLR, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

AI USE DISCLOSURE

In this work, we used AI tools to improve the readability and clarity of the manuscript, including proofreading, grammatical correction, and refinement of phrasing. We reviewed and verified all AI-assisted text and take full responsibility for the final content, claims, and conclusions of this work.

A EXPERIMENTAL DETAILS

Dataset and architecture. For supervised classification experiments, we use CIFAR-10, which contains 50,000 training images from 10 classes. We train a custom HybridViT architecture consisting of a convolutional stem followed by three hierarchical stages. Each stage contains two HybridBlocks, and each HybridBlock combines a depthwise convolutional branch, multi-head self-attention, and an MLP branch with residual connections. The three stages use embedding dimensions of 64, 128, and 256, with 2, 4, and 8 attention heads, respectively. Linear projection layers are used to downsample the representations between consecutive stages. At the end of the backbone, the token representations are normalized and averaged across tokens, followed by a linear classification head.

Training configuration. The HybridViT models are trained for 50 epochs with a batch size of 64 using the Adam optimizer and a learning rate of 1×10^{-3}. We use cross-entropy loss and do not apply learning-rate scheduling or decay. For CUDA-based training, we use fused Adam and automatic mixed-precision computation. Unless otherwise specified, all models are trained under the same optimization and data-processing settings.

Sample-wise gradient computation. Sample-wise gradients are computed with respect to the MLP modules of all six HybridBlocks, spanning the three hierarchical stages. For each training checkpoint, we use the same fixed subset of 1,000 training examples, randomly sampled without replacement from the CIFAR-10 training set. The subset indices are shared across all independent runs and checkpoints, ensuring consistent sample correspondence throughout the analysis.

For each input sample, we first perform a forward pass and compute its classification loss. We then obtain the gradient of the loss with respect to the output of the target MLP module using automatic differentiation. Together with the corresponding MLP activations and output-layer weights, these quantities are used to construct a per-token gradient representation of the target MLP parameters. When a module produces multiple tokens, their gradient representations are summed over tokens to obtain a single gradient vector for each sample. The resulting sample-wise gradient vectors are subsequently ℓ_2-normalized before further analysis. Gradient extraction is performed with the model in evaluation mode while keeping gradient computation enabled, thereby avoiding stochastic training-time effects such as dropout and batch-normalization updates.

Clustering. We apply K-means clustering to PCA-reduced sample-wise gradient representations. Unless otherwise specified, the number of clusters is set to $K = 10$, corresponding to the number of classes in CIFAR-10. For the clustering sensitivity analysis in Appendix C, we additionally consider $K \in \{15, 20\}$.

ARI computation. We use the adjusted Rand index (ARI) implemented in scikit-learn to quantify the agreement between gradient-based cluster assignments. We evaluate clustering consistency from three complementary perspectives. First, for each layer and training checkpoint, we compute the pairwise ARI between cluster assignments obtained from three independently trained runs and report the mean across all valid run pairs. Second, within each run, we compute pairwise ARI between cluster assignments at different training checkpoints to characterize the temporal evolution of gradient structure. Third, we compare cluster assignments across both independent runs and training checkpoints to assess cross-run temporal consistency. In addition, we compute the ARI between the unsupervised gradient clusters and the ground-truth CIFAR-10 class labels. When multiple independent runs are available, the resulting ARI values are averaged across runs.

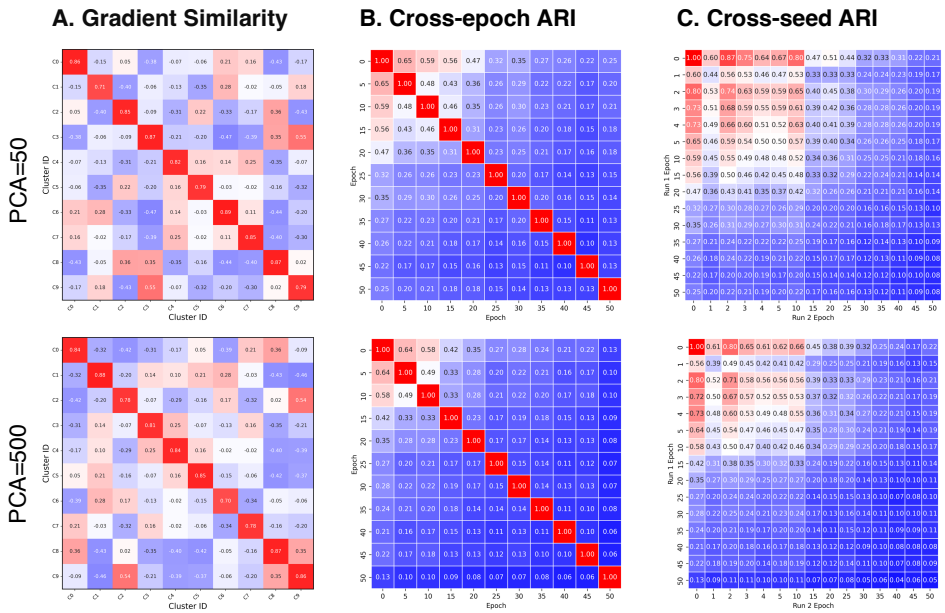

Figure 7: **Gradient organization is robust to PCA dimensionality.** Temporal ARI (left two columns) and cross-run ARI (right two columns) for ResNet-18 on CIFAR-10 using 50 and 500 principal components. The qualitative patterns remain consistent with the 150-PC setting used in the main text: gradient organization changes more strongly in shallow layers during training, while deeper layers show greater temporal stability and cross-run reproducibility.

B Gradient Organization is Robust to PC Dimension

In the main text, we reduce the sample-wise gradients to 150 principal components before clustering. Here, we examine whether the observed gradient organization depends on this choice by repeating the analysis with 50 and 500 principal components.

Figure 7 shows the temporal and cross-run ARI under different PCA dimensions. The same qualitative patterns are observed across all settings: (1) within-run temporal ARI is higher in early training and gradually decreases as training proceeds; (2) shallow layers exhibit greater changes in gradient organization, whereas deeper layers remain more stable; and (3) cross-run ARI shows the same depth-dependent pattern, with stronger reproducibility in deeper layers.

Although the absolute ARI values vary slightly with the number of retained components, the overall trends remain consistent across 50, 150, and 500 PCs. Thus, the observed evolution and reproducibility of gradient organization are robust to the choice of PCA dimension.

C Gradient Organization is Robust to Clustering Granularity

In the main text (Section 4.2), we analyzed gradient organization using $K = 10$ clusters, matching the number of classes in CIFAR-10. Here we verify that our findings are robust to the choice of clustering granularity by repeating the analysis with $K \in \{15, 20\}$.

Figure 8 shows temporal and cross-run ARI for different numbers of clusters. With larger K, we observe the same qualitative patterns: (1) within-run temporal ARI starts high in early epochs and gradually decreases as training proceeds; (2) this decrease is more pronounced in shallow layers while deep layers maintain higher consistency; (3) cross-run ARI between independently initialized models follows the same layer-dependent pattern, with deep layers showing sustained reproducibility even under finer-grained clustering.

The absolute ARI values are slightly lower with larger K, which is expected: finer partitions have more opportunities for disagreement, making perfect alignment less likely even for similar underlying

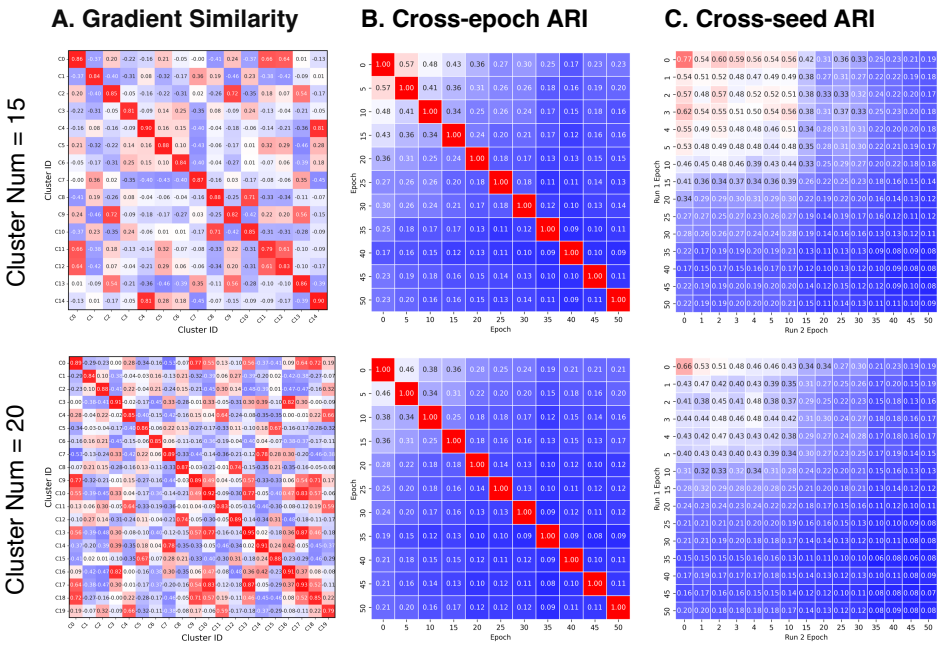

Figure 8: **Gradient organization patterns are robust to clustering granularity.** Temporal ARI (left two columns) and cross-run ARI (right two columns) for ResNet-18 on CIFAR-10 with $K \in \{15, 20\}$ clusters. Despite finer partitioning, deep layers maintain high temporal consistency and cross-run reproducibility, while shallow layers exhibit greater drift. The qualitative patterns match those observed with $K = 10$ in the main text.

structures. However, the relative trends—early stability followed by shallow-layer drift, and persistent deep-layer reproducibility—remain consistent across clustering granularities. This confirms that gradient organization structure is not an artifact of matching the cluster count to the number of classes, but reflects genuine geometric properties of the sample-wise gradient space.