Iterative Magnitude Pruning Reduces Weight-Space Coupling

Lucas Perez [1] **Mariana Ordones Oliveira Soares** [1] **Jackson de Faria** [2] **Fabricio Murai** [3] **Renato Assunção** [1 4]

Abstract

Neural networks contain many redundant and approximately equivalent parameterizations, yet it remains unclear how pruning affects this weight-space structure. We study iterative magnitude pruning through Fisher geometry and introduce a per-dimension log-determinant ratio that measures parameter coupling. Across several vision architectures and datasets, pruning consistently reduces this coupling over broad sparsity ranges. Controlled comparisons with smaller unpruned networks suggest that the effect is not explained by parameter count alone. We further show that as Fisher coupling decreases, Adam updates become more aligned with natural-gradient updates, while the same is not true for SGD. These results suggest that winning tickets are not only sparse and trainable, but also occupy locally simpler, more coordinate-separable regions of weight space.

1. Introduction

Modern neural networks are highly overparameterized, which introduces many degrees of freedom in weight space. Some of these degrees of freedom correspond to exact symmetries, such as permutations of hidden units, while others are approximate, data-dependent, or only visible locally through the geometry of the loss. Understanding how training and compression interact with these redundancies is central to the study of weight-space structure.

The Lottery Ticket Hypothesis (LTH) (Frankle & Carbin, 2019) states that dense randomly initialized networks contain sparse subnetworks, or winning tickets, that can be trained in isolation to match the performance of the original dense model. These subnetworks are commonly found through iterative magnitude pruning (IMP), which alternates between training, removing low-magnitude weights, and rewinding the surviving weights. While the empirical phenomenon is well established (Liu et al., 2024), it is less clear what kind of weight-space structure pruning selects for. Does pruning simply reduce the number of parameters, or does it also remove redundant and coupled directions in the local geometry of the model?

In this work, we study pruning as a mechanism for reducing local weight-space coupling. Our guiding hypothesis is that IMP does not merely shrink a model, but progressively removes parameters that participate in redundant curvature interactions. As a result, the surviving subnetwork occupies a region of weight space whose geometry is closer to coordinate-separable.

To test this hypothesis, we use the Fisher Information Matrix (FIM) as a local measure of parameter dependence. Off-diagonal FIM entries capture interactions between parameters in the local curvature of the predictive distribution, while a diagonal FIM corresponds to a setting in which curvature decomposes across coordinates. We introduce a log-determinant ratio derived from the Kullback-Leibler divergence between multivariate Gaussians that measures the gap between the full FIM geometry and its diagonal approximation. This metric is zero exactly when the FIM is diagonal and increases as parameter coupling grows.

Our empirical findings are threefold. First, across multiple architectures and datasets, IMP tends to reduce the FIM log-determinant ratio per dimension over a broad range of sparsities. Second, this reduction is not explained by parameter count alone: smaller unpruned networks with comparable numbers of parameters exhibit substantially larger coupling than lottery-ticket subnetworks. Third, the reduction in FIM coupling has an optimization consequence: as pruning progresses, Adam (Kingma & Ba, 2017) updates become increasingly aligned with natural-gradient updates. Together, these results support the view that lottery tickets are not only sparse and trainable, but also locally simplified in weight space. Source code and experimental artifacts are available at `https://github.com/PerezLucas2903/fisher_coupling`.

[1]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil [2]Simon Fraser University, Canada [3]Worcester Polytechnic Institute, Worcester, MA, USA [4]ESRI, Redlands, CA, USA. Correspondence to: Lucas Perez <perezlucas2903@gmail.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. Related Work

Weight-space symmetries and alignment: Neural networks admit many equivalent or near-equivalent parameterizations, including exact permutation symmetries and approximate data-dependent redundancies. Recent work studies these structures through model alignment and linear mode connectivity, showing that independently trained models can often be connected by low-loss paths after accounting for permutation symmetries (Ainsworth et al., 2023). Our work studies a complementary setting: rather than aligning distinct models post hoc, we ask how IMP changes the local weight-space geometry of a single model.

Fisher curvature and pruning: The Fisher Information Matrix defines a local geometry on parameter space and underlies natural-gradient methods (Amari, 1998; Calin, 2020). Prior Fisher-based pruning methods typically use diagonal or block-diagonal Fisher information as an importance score for removing parameters (Liu et al., 2021). In contrast, we do not use Fisher information to prune. Instead, we measure how pruning changes the Fisher structure itself, focusing on the reduction of off-diagonal curvature coupling.

Lottery tickets and loss geometry: The Lottery Ticket Hypothesis shows that dense networks contain sparse subnetworks that can be trained in isolation to match dense-model performance (Frankle & Carbin, 2019). Related work has connected winning tickets to loss-landscape properties such as linear mode connectivity (Frankle et al., 2019). Our contribution is a Fisher-geometric diagnostic: we show empirically that IMP tends to select subnetworks with more coordinate-separable local curvature, and that this simplification makes Adam updates more aligned with natural-gradient directions.

3. Measuring Weight-Space Coupling

Let $H \succ 0$ be positive definite and write $H = C^\top C$. By Hadamard's inequality (Różański et al., 2017), $\det(H) = \det(C)^2 \leq \prod_{i=1}^n \|v_i\|^2 = \prod_{i=1}^n h_{ii}$, where v_i is the ith column of C. Hence, the determinant of H is upper bounded by the product of its diagonal entries.

Now let $H(\theta) = \mathbb{E}[\nabla\mathcal{L}(\theta) \cdot \nabla\mathcal{L}(\theta)^\top] + \lambda I$, with $\lambda > 0$, denote the empirical FIM evaluated at a trained model with parameters $\theta \in \mathbb{R}^P$. The diagonal entries of H describe coordinate-wise curvature, while the off-diagonal entries describe local coupling between parameters. To quantify how far the curvature is from a coordinate-separable regime, we define

$$f(H) \;:=\; \frac{1}{d}\log\left(\frac{\det(\mathrm{diag}(H))}{\det(H)}\right) \tag{1}$$

where d is the number of unpruned dimensions of H. We refer to $f(H)$ as the *log-determinant ratio per dimension* (or LogDet, for short). This quantity has several useful properties. By Hadamard's inequality, LogDet ≥ 0, with equality if and only if H is diagonal. Thus, lower values correspond to weaker local parameter coupling. The normalization by dimension allows comparison across layers and architectures with different parameter counts, analogous to "log-likelihood per dimension" in diffusion models or "bits per dimension" in compression. The metric is also invariant to coordinate-wise rescaling of the curvature matrix, making it suitable for comparing FIM structure across pruning stages.

LogDet has a probabilistic interpretation. If the full FIM defines a local Gaussian approximation with precision H, and its diagonal approximation defines a Gaussian with precision $diag(H)$, then the unnormalized LogDet is proportional to the KL divergence between these two local Gaussian models, as shown in Appendix A. In this sense, LogDet measures the information lost by replacing the full local geometry with a diagonal approximation.

We interpret LogDet as a diagnostic for approximate weight-space redundancy. When many parameters participate in coupled curvature directions, the full FIM differs substantially from its diagonal approximation and LogDet is large. When pruning removes redundant or co-adapted parameters, the remaining coordinates become less coupled and LogDet decreases. This gives a simple empirical test of whether pruning simplifies the local geometry of weight space.

4. Experimental Results

We study iterative magnitude pruning across several architectures and datasets. For small-scale experiments, we use ConvSmall, a compact convolutional network with two convolutional blocks followed by a linear classifier. This model is small enough to allow full-network FIM computation. We train ConvSmall on MNIST (LECUN, 1998), KMNIST (Clanuwat et al., 2018), EMNIST (Cohen et al., 2017), and FashionMNIST (Xiao et al., 2017).

To test whether this behavior extends beyond small networks, we evaluate ResNet18 (He et al., 2015) and DenseNet (Huang et al., 2018) on CIFAR10 (Krizhevsky et al., 2014), SVHN (Netzer et al., 2011), and STL10 (Coates et al., 2011). For tractability, we estimate FIMs layer-wise on selected convolutional layers using parameter subsampling that preserves within-filter correlations; these estimates omit cross-layer interactions and serve as scalable local proxies. As a sanity check, Appendix Figure 4 shows that layer-wise LogDet follows the full-network trend despite differences in scale.

The pruning procedure follows the standard IMP pipeline. At each pruning round, the network is trained, low-

magnitude weights are removed, and the surviving weights are rewound to initialization according to the lottery-ticket protocol. We evaluate the FIM at the end of each training cycle, immediately before the next pruning step. This allows us to measure the geometry of the subnetwork that has been selected by the previous pruning rounds. All results are averaged across multiple random seeds, using 10 runs for ConvSmall and 5 for the larger architectures.

4.1. Pruning Reduces FIM Coupling

Across datasets and architectures, IMP consistently reduces the LogDet over broad sparsity ranges. For ConvSmall, where the full-network FIM can be computed, LogDet decreases steadily as pruning progresses from the dense model to moderately sparse subnetworks. This trend appears on all datasets, as shown on Table 1, suggesting that pruning removes curvature coupling in a way that is not specific to a single dataset.

Table 1. Diagonalization results for ConvSmall on small datasets

	LogDet (ConvSmall)			
Density	MNIST	KMNIST	EMNIST	FashionMNIST
100%	3.082±.21	1.933±.19	1.449±.15	2.214±.09
90%	2.876±.19	1.754±.15	1.413±.16	2.152±.12
80%	2.660±.28	1.575±.13	1.365±.18	2.076±.09
70%	2.363±.22	1.489±.09	1.352±.21	1.963±.11
60%	2.203±.20	1.443±.11	1.412±.18	1.970±.03
50%	1.928±.18	1.351±.08	1.366±.19	1.847±.07
40%	1.722±.12	1.293±.09	1.361±.15	1.866±.07
30%	1.493±.16	1.227±.12	1.216±.10	1.744±.14
20%	**1.251±.14**	**1.191±.10**	**1.108±.10**	1.707±.22
10%	1.393±.34	1.275±.26	1.243±.29	**1.565±.22**

The same qualitative behavior appears in larger architectures. For DenseNet, shown in Figure 1, and ResNet18, shown in Appendix B, layer-wise FIM measurements generally contract as sparsity increases. Although different layers have different absolute LogDet values, early, middle, and late layers tend to follow the same broad trend. This suggests that pruning induces a distributed simplification of the local geometry rather than affecting only a specific depth region.

The relationship between coupling and accuracy is not one-to-one. In many cases, LogDet decreases substantially before accuracy degrades, indicating that the network can lose curvature coupling while maintaining predictive performance. At extreme sparsities, however, some models exhibit a regime shift: LogDet may stop decreasing or even increase, while accuracy begins to fall. We interpret this as an over-pruning regime in which the remaining parameters must coordinate more strongly to preserve performance. In this regime, the model may still classify reasonably well, but its local geometry becomes more coupled and less coordinate-separable.

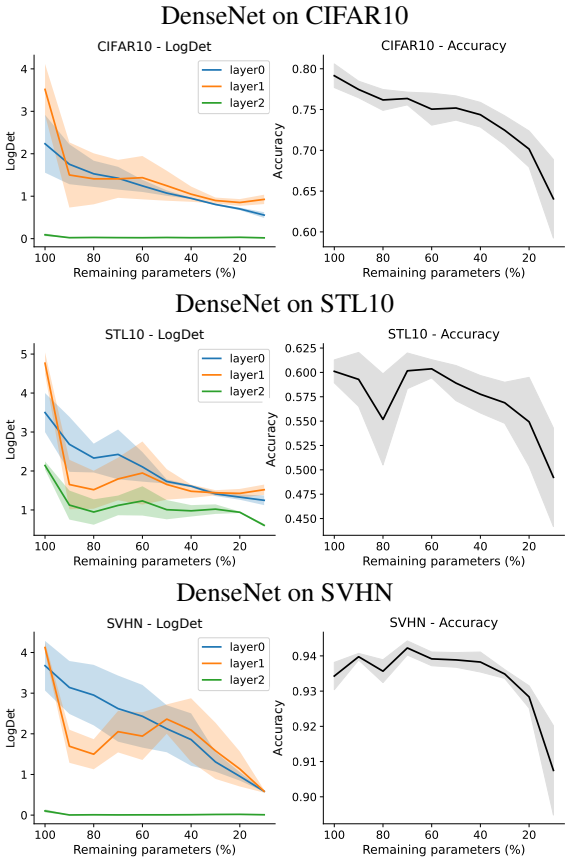


Figure 1. DenseNet results on CIFAR10, STL10 and SVHN

These findings support the view that IMP acts as a weight-space simplification mechanism. Rather than merely discarding parameters, pruning appears to select subnetworks whose remaining parameters interact less strongly in the FIM geometry.

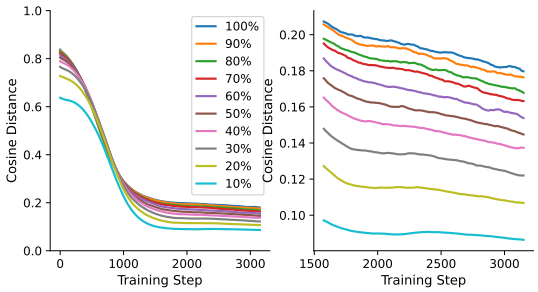
4.2. Reduced Coupling Aligns Adam with Natural Gradient

If pruning makes the FIM more diagonal, then diagonal adaptive optimizers should become better matched to the local geometry. Natural-gradient descent preconditions gradients by the inverse FIM, while Adam uses a diagonal preconditioner derived from gradient statistics. When the FIM has strong off-diagonal structure, Adam cannot capture the full natural-gradient direction. But when the FIM becomes nearly diagonal, the natural-gradient update becomes closer to a coordinate-wise rescaling, which Adam can approximate.

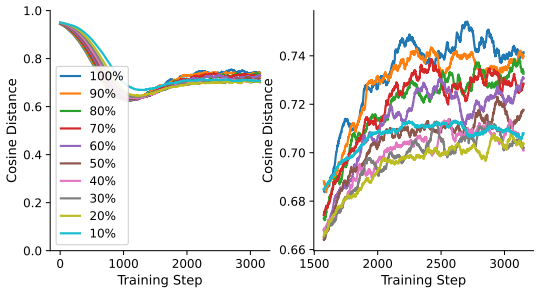
We test this prediction by comparing Adam and natural-gradient update directions across pruning densities. To avoid confounding the comparison with different optimization trajectories, we evaluate both updates along the same Adam trajectory. At each step, we compute the Adam update

and the natural-gradient update at the same parameter state using the same minibatch, then apply only the Adam update. We measure the cosine distance between the two update directions, excluding pruned dimensions.

The results show that Adam becomes more aligned with natural gradient as sparsity increases, while standard SGD optimization does not achieve the same results. Figure 2a shows that for DenseNet, the cosine distance between Adam and NGD decreases as sparsity increases: the dense network exhibits large discrepancies, while pruned subnetworks converge to much closer update directions. Additional results across architectures and datasets are provided in Appendix C.



(a) NGD vs Adam



(b) NGD vs SGD

Figure 2. Cosine distance between NGD and first-order optimizers under IMP of DenseNet trained on CIFAR-10. The right plots are a zoom at the second half of training.

In contrast, Figure 2b shows that SGD does not exhibit the same alignment with NGD, even at high sparsity. Unlike Adam, SGD applies uniform updates and cannot capture the coordinate-wise curvature scaling that emerges as the FIM becomes near-diagonal. As a result, it fails to achieve the tight alignment observed with Adam, indicating that this effect is specific to diagonal adaptive preconditioning rather than a generic consequence of pruning. Further comparisons between SGD and NGD are reported in Appendix D.

4.3. The Effect Is Not Explained by Parameter Count Alone

A natural alternative explanation is that smaller models simply have lower FIM coupling. If so, the observed LogDet reduction would not be specific to lottery-ticket pruning; it would merely reflect the reduced number of parameters.

To isolate the effect of IMP from that of simply reducing parameter count, we perform a controlled experiment using compact convolutional architectures that preserve the structure of ConvSmall while reducing its width. We construct models with approximately 60%, 40%, and 20% of the original parameters and train them from scratch without pruning. For each configuration, we compute LogDet and compare it to the values obtained from IMP-pruned subnetworks at matching parameter fractions.

Figure 3 contrasts these networks. The unpruned small models exhibit substantially higher LogDet values than the pruned networks at matching capacities, despite achieving similar accuracies. This gap indicates that the diagonalization effect is not merely a consequence of reduced model size, but is instead induced by the structured removal of parameters performed by IMP. Results for all datasets are provided in Appendix E.

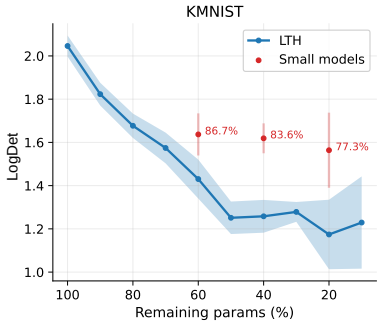


Figure 3. LogDet for KMNIST comparing small models trained from scratch vs. LTH-pruned models at matching parameter fractions. The numbers reflect each model's accuracy.

5. Conclusion

We studied iterative magnitude pruning through Fisher weight-space geometry. Using a log-determinant ratio that measures deviation from diagonal curvature, we found that pruning consistently reduces local parameter coupling across architectures and datasets. This effect is not explained by parameter count alone: compact unpruned networks remain more coupled than matched lottery-ticket subnetworks. We also found that reduced FIM coupling makes Adam updates more aligned with natural-gradient updates, while the same trend does not appear for SGD.

These results suggest that winning tickets are not merely smaller networks, but sparse parameterizations with simpler,

more coordinate-separable local geometry. Future work should test whether the same coupling reduction appears in larger architectures such as transformers, LoRA adapters, mixture-of-experts models, and alternative pruning methods beyond IMP.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries, 2023. URL https://arxiv.org/abs/2209.04836.

Amari, S.-i. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Calin, O. *Deep Learning Architectures: A Mathematical Approach*. Springer Publishing Company, Incorporated, 1st edition, 2020. ISBN 3030367207.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018. URL http://arxiv.org/abs/1812.01718.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. Emnist: an extension of mnist to handwritten letters, 2017. URL https://arxiv.org/abs/1702.05373.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks, 2019. URL https://arxiv.org/abs/1803.03635.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. *CoRR*, abs/1912.05671, 2019. URL http://arxiv.org/abs/1912.05671.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Krizhevsky, A., Nair, V., Hinton, G., et al. The cifar-10 dataset, 2014.

LECUN, Y. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998. URL https://cir.nii.ac.jp/crid/1571417126193283840.

Liu, B., Zhang, Z., He, P., Wang, Z., Xiao, Y., Ye, R., Zhou, Y., Ku, W.-S., and Hui, B. A survey of lottery ticket hypothesis, 2024. URL https://arxiv.org/abs/2403.04861.

Liu, L., Zhang, S., Kuang, Z., Zhou, A., Xue, J.-H., Wang, X., Chen, Y., Yang, W., Liao, Q., and Zhang, W. Group fisher pruning for practical network compression, 2021. URL https://arxiv.org/abs/2108.00708.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Różański, M., Witula, R., and Hetmaniok, E. More subtle versions of the hadamard inequality. *Linear Algebra and its Applications*, 532, 07 2017. doi: 10.1016/j.laa.2017.07.003.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

A. LogDet and KL Divergence

This subsection provides a probabilistic interpretation of the unnormalized LogDet:

$$\tilde{f}(H) = \log\left(\frac{\det(\mathrm{diag}(H))}{\det(H)}\right), \qquad H \succ 0.$$

Let $H \succ 0$ be any positive definite matrix and let $D := \mathrm{diag}(H)$. Consider the two Gaussians

$$X \sim \mathcal{N}(\mu,\, H^{-1}), \qquad Y \sim \mathcal{N}(\mu,\, D^{-1}),$$

which share the same mean μ but differ in covariance.

The KL divergence between Gaussians (with identical means) is

$$\mathrm{KL}(Y\|X) = \frac{1}{2}\Big(\mathrm{Tr}(HD^{-1}) - k + \log\frac{\det(H^{-1})}{\det(D^{-1})}\Big),$$

where k is the dimension.

Since D is the diagonal of H, we have

$$\mathrm{Tr}(HD^{-1}) = \sum_{i=1}^{k} H_{ii}\,(D^{-1})_{ii} = \sum_{i=1}^{k} H_{ii}\cdot\frac{1}{H_{ii}} = k.$$

Substituting yields

$$\mathrm{KL}(Y\|X) = \frac{1}{2}\log\frac{\det(H^{-1})}{\det(D^{-1})} = \frac{1}{2}\log\frac{\det(D)}{\det(H)} = \frac{1}{2}\tilde{f}(H).$$

Thus, up to a factor of 2, the unnormalized log-determinant ratio $\tilde{f}(H)$ equals the KL divergence between a Gaussian with diagonal precision D and the Gaussian with full precision H. The metric reported in the main experiments is the normalized quantity $\tilde{f}(H)/d$.

B. Extended LTH LogDet Results

This appendix provides extended LogDet measurements under iterative magnitude pruning across architectures and datasets. The results show that both layer-wise and full-network estimates follow similar decreasing trends in coupling. Figure 4 illustrates this comparison, where the absolute values differ but the overall trends are closely aligned.

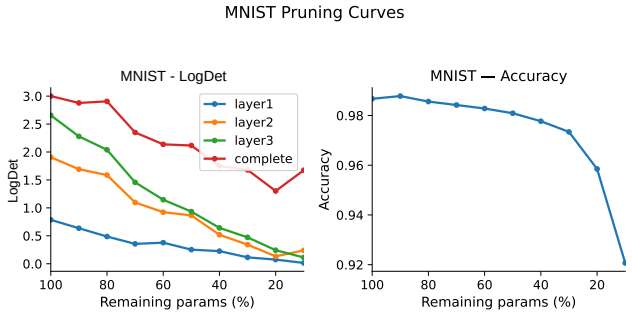


Figure 4. ConvSmall trained on MNIST

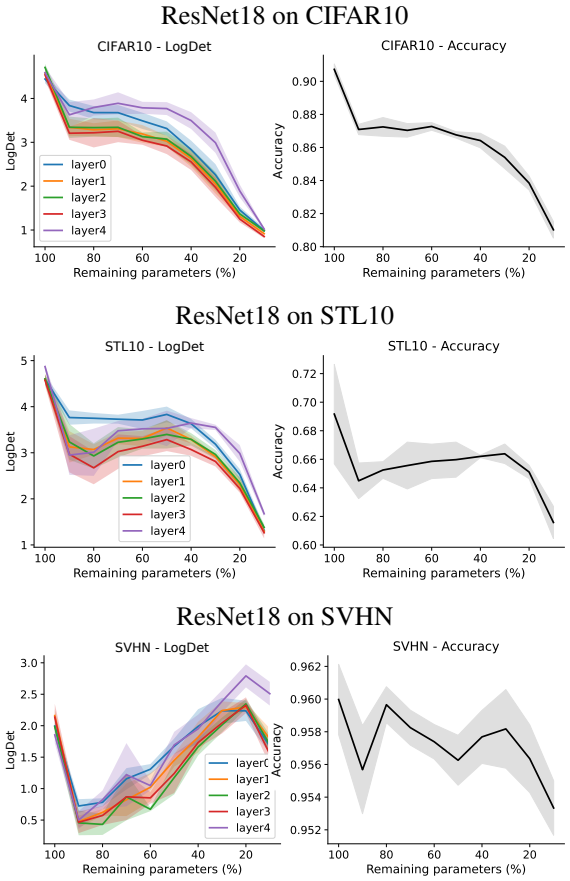


Figure 5. ResNet18 results demonstrating FIM diagonalization and accuracy across CIFAR10, STL10, and SVHN datasets.

C. Extended Adam vs NGD Results

This appendix presents additional comparisons between Adam and natural-gradient updates across datasets and architectures. The plots report cosine distances along the training trajectory, illustrating that the alignment between Adam and natural gradient improves as sparsity increases, consistent with the diagonalization of the FIM discussed in the main text.

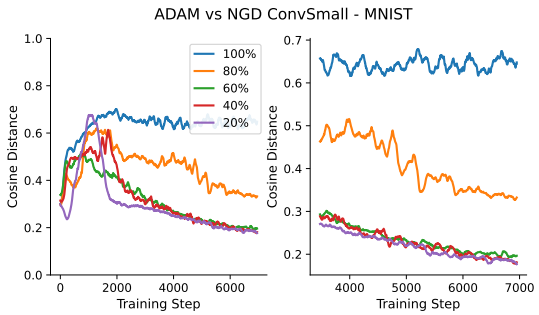


Figure 6. Adam vs NGD Cosine Distance for ConvSmall on MNIST.

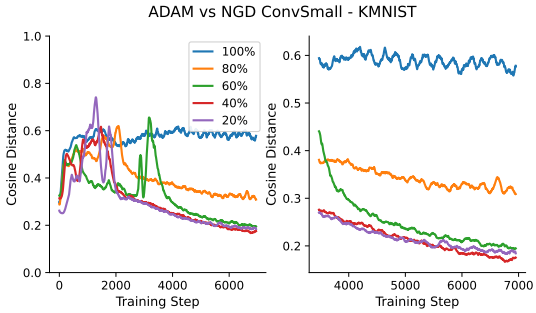


Figure 7. Adam vs NGD Cosine Distance for ConvSmall on KMNIST.

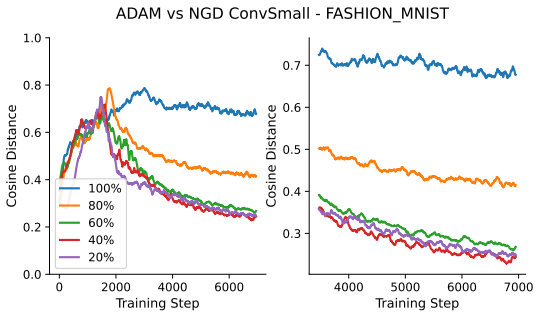


Figure 8. Adam vs NGD Cosine Distance for ConvSmall on FashionMNIST.

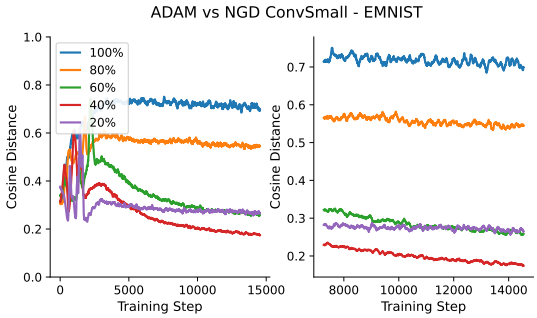


Figure 9. Adam vs NGD Cosine Distance for ConvSmall on EMNIST.

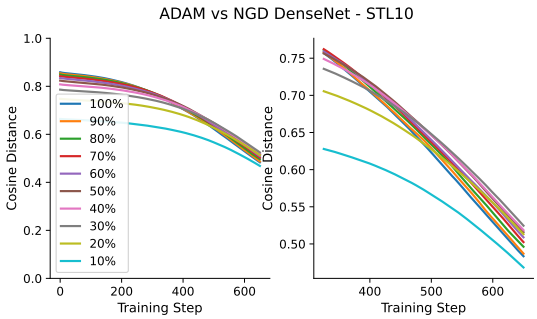


Figure 10. Adam vs NGD Cosine Distance for DenseNet on STL10.

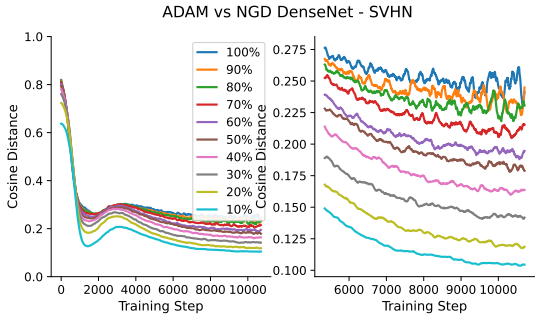


Figure 11. Adam vs NGD Cosine Distance for DenseNet on SVHN.

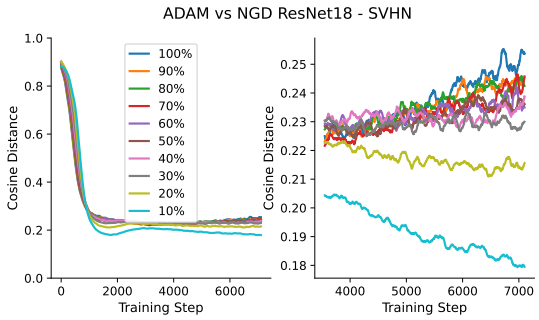


Figure 12. Adam vs NGD Cosine Distance for ResNet18 on SVHN.

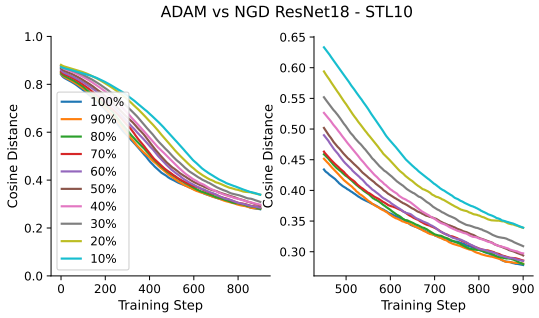


Figure 13. Adam vs NGD Cosine Distance for ResNet18 on STL10.

D. SGD vs NGD Baseline Comparisons

This appendix presents additional comparisons between SGD and natural-gradient updates. In contrast to Adam, SGD does not show the same clear trend toward improved alignment as sparsity increases, with cosine distances remaining high throughout training.

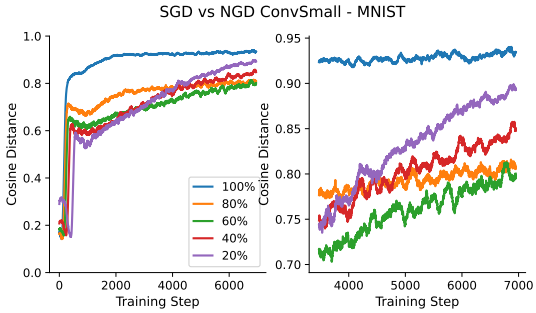


Figure 14. SGD vs NGD Cosine Distance for ConvSmall on MNIST.

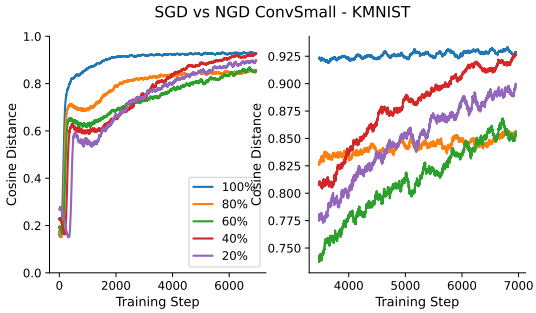


Figure 15. SGD vs NGD Cosine Distance for ConvSmall on KMNIST.

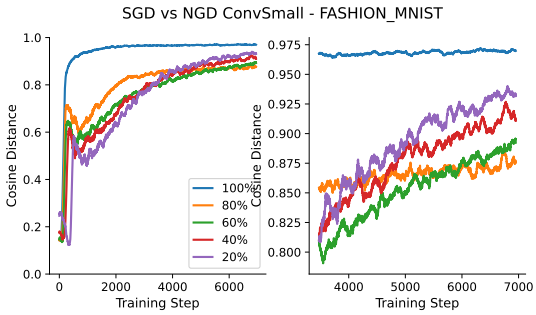


Figure 16. SGD vs NGD Cosine Distance for ConvSmall on FashionMNIST.

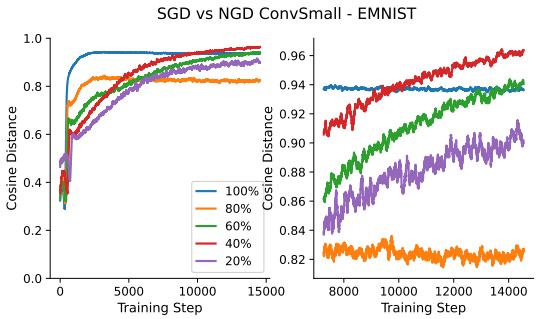


Figure 17. SGD vs NGD Cosine Distance for ConvSmall on EMNIST.

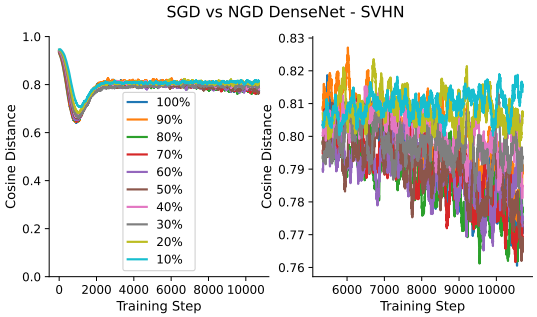


Figure 18. SGD vs NGD Cosine Distance for DenseNet on SVHN.

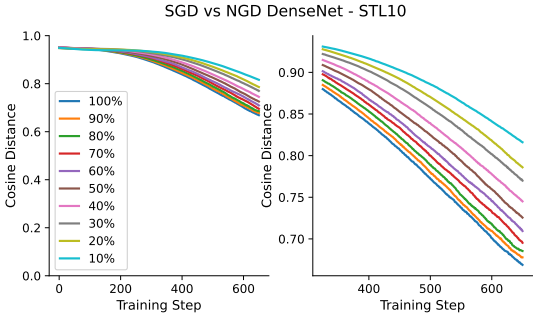


Figure 19. SGD vs NGD Cosine Distance for DenseNet on STL10.

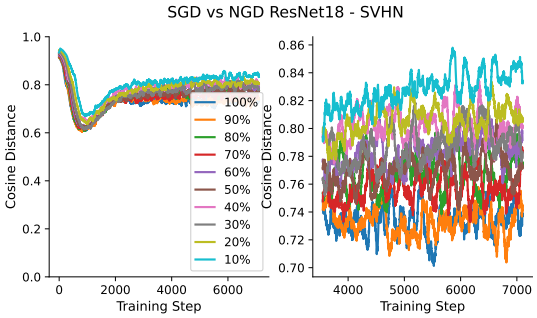


Figure 20. SGD vs NGD Cosine Distance for ResNet18 on SVHN.

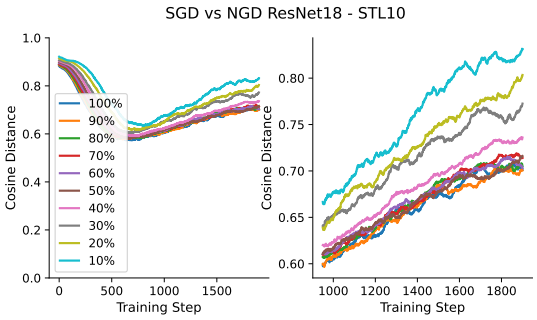


Figure 21. SGD vs NGD Cosine Distance for ResNet18 on STL10.

E. Diagonalization Is Induced by Pruning, Not by Model Size

This appendix provides additional comparisons between pruned subnetworks and smaller dense models trained from scratch at matched capacities. Figure 22 and Figure 23 show that pruned models tend to exhibit lower LogDet values even when parameter counts are similar. Values annotated next to the red points indicate mean classification accuracy.

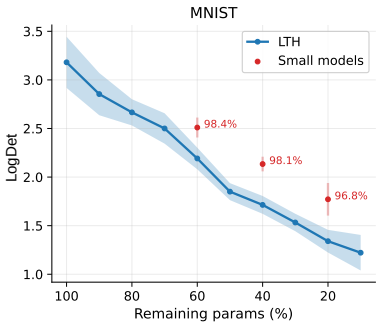


Figure 22. MNIST

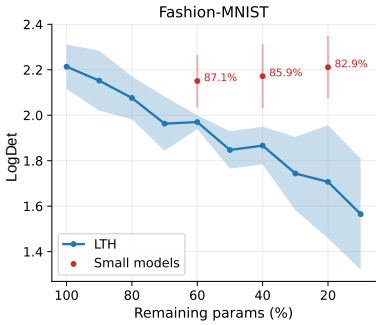


Figure 23. FashionMNIST