THE GEOMETRY OF SPARSITY: FISHER INFORMATION ACROSS NEURAL NETWORK PRUNING METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern neural networks contain substantial parameter redundancy, yet the geometric consequences of removing parameters remain poorly understood. We study pruning through the Fisher Information Matrix (FIM), focusing on statistical coupling among surviving parameters. We introduce a log-determinant ratio that quantifies deviation from a diagonal Fisher structure and show that, under a fixed-Jacobian damped Gram representation, deleting any parameter monotonically reduces the unnormalized coupling measure, regardless of the pruning criterion. We then study how this structural effect manifests after training. A comparison of iterative magnitude pruning (IMP), random pruning, SNIP, GraSP, and RigL shows that reduced Fisher coupling is not specific to lottery-ticket pruning, although different methods produce substantially different geometric trajectories. Using IMP as a controlled setting, we further observe decreasing Fisher coupling across architectures, datasets, and sparsity levels. We characterize the sensitivity of the metric to its damping parameter and show, within IMP trajectories, that diagonalization is accompanied by greater alignment between Adam and natural-gradient updates. These results suggest that reduced Fisher coupling is a recurring consequence of sparsification, while the geometry of the resulting sparse model depends on how sparsity is imposed.

1 INTRODUCTION

Modern deep neural networks are often heavily overparameterized, which contributes to their expressive power but also makes them costly to train, store, and deploy. This has motivated a broad range of pruning methods that remove parameters while attempting to preserve predictive performance. Existing approaches differ substantially in how they identify parameters to remove: magnitude-based methods prune small weights after or during training, saliency-based methods use local gradient information (Lee et al., 2019; Wang et al., 2020), and dynamic sparse training methods continually modify the sparse connectivity pattern during optimization (Evci et al., 2021). Other approaches derive pruning criteria from properties of signal propagation or network structure (Tanaka et al., 2020). Despite these methodological differences, successful pruning algorithms often produce sparse subnetworks that retain a surprisingly large fraction of the performance of their dense counterparts.

A particularly influential example is the Lottery Ticket Hypothesis (LTH) (Frankle & Carbin, 2019), which proposes that randomly initialized dense networks contain sparse subnetworks, called winning tickets, that can be trained in isolation to match the performance of the original network. Winning tickets are commonly identified through iterative magnitude pruning (IMP), which alternates between training, pruning low-magnitude parameters, and rewinding the leftover parameters. The empirical success of LTH has motivated extensive work on the optimization and representational properties of sparse subnetworks (Liu et al., 2024; Chen et al., 2021). Yet, IMP represents only one particular sparsification mechanism. A broader question is whether successful pruning methods, despite using different criteria and optimization procedures, produce networks that share common geometric properties.

In this work, we investigate one such property: the structure of neural-network curvature after pruning. Our central hypothesis is that pruning generally reduces parameter co-dependence in curvature space, driving the Fisher Information Matrix (FIM) toward a more diagonal structure. Under this view, curvature diagonalization may extend beyond winning tickets and magnitude pruning, while its extent can depend on the sparsification procedure and subsequent optimization trajectory. Different pruning algorithms select different subnetworks and need not produce identical curvature after training.

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.

To quantify this change, we introduce a log-determinant–based measure of deviation from diagonality. For a positive-definite curvature matrix H, our metric compares $\det(H)$ with the determinant of its diagonal part and vanishes exactly when H is diagonal. The metric admits an information-theoretic interpretation through the Kullback–Leibler divergence between Gaussian distributions and provides a global measure of normalized parameter coupling. Importantly, under a fixed-Jacobian damped Gram-matrix model, we show that removing *any* parameter can only decrease this measure. This result is independent of the criterion used to select the removed parameter: magnitude pruning, gradient-based pruning, Fisher-based pruning, or any other coordinate-removal rule has the same one-step structural effect when the curvature representation is held fixed. Retraining can subsequently reshape the Jacobian, so we conduct a systematic empirical study to determine whether this diagonalization persists over a complete prune-and-retrain trajectory.

Our experiments indicate that this phenomenon holds true under retraining and extends beyond the LTH setting. Across multiple architectures, datasets, sparsity levels, and pruning procedures, we generally observe reduced off-diagonal Fisher structure as sparsity increases, although the trajectories differ substantially across methods and can become non-monotonic at extreme sparsity. IMP provides a particularly controlled setting in which to study this evolution, while complementary cross-method experiments show that reduced coupling is not unique to lottery-ticket pruning.

This perspective also connects pruning to the behavior of diagonal adaptive optimizers such as Adam (Kingma & Ba, 2017). Natural Gradient Descent (NGD) (Amari, 1998) preconditions gradients using the inverse Fisher matrix, accounting explicitly for interactions between parameters. When the FIM becomes approximately diagonal, however, natural-gradient preconditioning reduces approximately to independent coordinate-wise rescaling. Standard adaptive methods already perform such diagonal rescaling using running gradient statistics. We therefore hypothesize that pruning progressively places the model in a geometric regime in which diagonal adaptive optimization becomes a better approximation to natural-gradient optimization. Consistent with this prediction, we observe increased alignment between Adam and natural-gradient update directions as pruning progresses.

Our results suggest that pruning should be understood not only as parameter removal, but also as a transformation of the local geometry of the model. We make the following contributions: (i) we introduce a log-determinant–based metric for measuring off-diagonal coupling in Fisher/Gauss–Newton curvature, together with a per-dimension normalization for empirical comparisons; (ii) we prove that coordinate removal monotonically decreases the unnormalized metric under a fixed-Jacobian damped Gram-matrix model, establishing an algorithm-independent structural effect of pruning; (iii) we empirically study curvature diagonalization across multiple pruning procedures, architectures, datasets, and sparsity levels, finding evidence that increasingly diagonal Fisher structure is a general feature of pruning rather than one restricted to LTH; and (iv) we connect this diagonalization to natural-gradient optimization, showing theoretically and empirically that diagonal adaptive methods increasingly approximate natural-gradient updates as curvature becomes more diagonal.

2 Related Work

We review related work on neural network pruning, Fisher-based curvature analysis, and natural-gradient optimization, emphasizing how our approach differs from existing perspectives.

Neural Network Pruning encompasses several families of methods that differ in when sparsity is imposed and how the sparse connectivity pattern is selected. The Lottery Ticket Hypothesis (LTH) (Frankle & Carbin, 2019) proposes that dense, randomly initialized networks contain sparse subnetworks ("winning tickets") that can be trained in isolation to match the performance of the original model. Such subnetworks are commonly identified through iterative magnitude pruning (IMP), which alternates training, removal of low-magnitude weights, and rewinding of the surviving parameters. In contrast, pruning-at-initialization methods construct sparse networks before conventional training begins. SNIP (Lee et al., 2019) ranks connections according to their sensitivity to removal, while GraSP (Wang et al., 2020) selects parameters with the objective of preserving gradient flow. Dynamic sparse training methods instead modify the sparse topology throughout optimization; RigL (Evci et al., 2021) periodically removes low-magnitude connections and grows new ones according to gradient information. We consider LTH/IMP, SNIP, GraSP, and RigL as representative examples of iterative magnitude pruning, pruning at initialization, and dynamic sparse training, allowing us to

study whether common geometric behavior emerges across substantially different mechanisms for inducing sparsity.

Fisher Information provides a Riemannian metric on the parameter space of probabilistic models (Calin, 2020) and plays a central role in understanding both optimization and pruning. For deep networks, analyses of the Fisher Information Matrix (FIM) reveal how curvature scales with width and input statistics, helping to explain phenomena such as sharpness and optimization efficiency (Pennington & Worah, 2018). In model compression, Fisher-based criteria have been used to rank parameter importance: parameters with low Fisher information are considered less critical and thus more suitable for pruning (Tu et al., 2016). Moreover, under standard regularity conditions, the FIM equals the expected Hessian of the negative log-likelihood, equivalently the negative expected Hessian of the log-likelihood, and in practice it is widely used as a positive semi-definite approximation to the Hessian matrix (Martens, 2020). Unlike approaches that use Fisher information as a criterion for selecting parameters to prune, our work uses the FIM diagnostically to study how different sparsification procedures reshape the coupling structure of the surviving parameter space.

Natural Gradient Descent (Amari, 1998) replaces the standard Euclidean gradient with the natural gradient obtained by preconditioning with the inverse FIM. This update direction corresponds to the steepest descent in the space of model distributions rather than raw parameters, yielding invariance to reparameterizations and, under suitable conditions, asymptotically efficient learning. More recent work (Martens, 2020) provides a modern treatment of natural gradient methods, juxtaposing them with practical second-order approximations (e.g., K-FAC (Martens & Grosse, 2020)) and connecting them to popular adaptive optimizers such as RMSProp and Adam via the empirical Fisher.

In our experiments, we utilize Structured Inverse-Free Natural Gradient Descent (SINGD) (Lin et al., 2024) as a numerically stable and memory-efficient implementation of NGD. Our contribution differs from these lines of work by treating curvature as a property to be characterized after sparsification, rather than primarily as a pruning criterion or optimizer design tool.

3 A Method-Agnostic Effect of Coordinate Pruning

In this section, we analyze how pruning affects the curvature structure of neural networks. We first recall the Hessian decomposition for piecewise-linear networks and motivate the Gauss–Newton term as a Fisher surrogate. We then introduce a log-determinant–based metric that quantifies curvature coupling and prove that coordinate removal monotonically decreases its unnormalized form under a fixed-Jacobian Gram-matrix model. Finally, we connect pruning-induced diagonalization to the behavior of diagonal adaptive optimizers, showing how it leads to increasingly natural-gradient like dynamics.

ReLU Gauss–Newton structure of the Hessian Consider a ReLU network trained on samples $\{(x_m, y_m)\}_{m=1}^M$ with mean-squared error $L(\theta) = \sum_{m=1}^M \big(\hat{y}(x_m, \theta) - y_m\big)^2/(2M), \theta \in \mathbb{R}^P$. Let $\hat{y}_m := \hat{y}(x_m, \theta)$ and $e_m := \hat{y}_m - y_m$. Differentiating twice gives

$$\nabla_\theta^2 L(\theta) = \frac{1}{M}\sum_{m=1}^{M}\Big[(\nabla_\theta \hat{y}_m)(\nabla_\theta \hat{y}_m)^\top + e_m \nabla_\theta^2 \hat{y}_m\Big] = \frac{1}{M} J(\theta)^\top J(\theta) + \frac{1}{M}\sum_{m=1}^{M} e_m \nabla_\theta^2 \hat{y}_m,$$

where $J(\theta)$ is the Jacobian of the model outputs $(\hat{y}_1, \ldots, \hat{y}_M)$ with respect to θ, i.e., $J_{mi} = \partial \hat{y}_m / \partial \theta_i$. Note that this is not the Jacobian of the loss.

The first term is the Gauss–Newton (Jacobian outer-product) curvature term and is positive semi-definite. The second term is the residual-weighted second-derivative contribution, involving the parameter Hessian of the network output. For ReLU networks, within any region of parameter space where the pre-activation sign pattern on the training samples does not change, each layer admits a constant diagonal mask $D_\ell \in \{0,1\}^{d_\ell \times d_\ell}$ selecting the active neurons. In such a region, the network output is *multilinear across layerwise parameter blocks*, not affine in all parameters jointly. Consequently, second derivatives with respect to parameters in the same layer vanish, while mixed second derivatives across different layers generally do not.

To make this precise, consider a scalar-output ReLU MLP with L hidden layers and trainable matrices $W_1, \ldots, W_{L+1}$, where W_{L+1} is the output row vector, and let $D_1, \ldots, D_L$ denote the fixed diagonal

ReLU masks. Inside a fixed activation region, the output has the form

$$\hat{y}(x,\theta) = W_{L+1} D_L W_L D_{L-1} \cdots D_1 W_1 x.$$

For notational simplicity, we omit biases. The following proposition bounds the second term above. In particular, its magnitude is residual-weighted and scales with products of layer norms. The proof is given in Appendix A.2.

Proposition 3.1 (Bound on the second term in a fixed activation region). *Let $N := L+1$ be the number of trainable layers, and write $\theta = (\theta^{(1)}, \ldots, \theta^{(N)})$, $\theta^{(r)} := \mathrm{vec}(W_r)$. Inside a region with fixed ReLU masks, $\nabla^2_{\theta^{(r)}\theta^{(r)}} \hat{y}(x,\theta) = 0$ for all r, and for $r \neq s$, $\left\|\nabla^2_{\theta^{(r)}\theta^{(s)}} \hat{y}(x,\theta)\right\|_2 \leq \|x\|_2 \prod_{\ell \neq r,s} \|W_\ell\|_2$. Hence*

$$\left\| \frac{1}{M} \sum_{m=1}^{M} e_m \nabla^2_\theta \hat{y}_m \right\|_2 \leq \frac{2}{M} \sum_{m=1}^{M} |e_m| \, \|x_m\|_2 \sum_{1 \leq r < s \leq N} \prod_{\ell \neq r,s} \|W_\ell\|_2.$$

Proposition 3.1 shows that when residuals are small and layer norms are controlled, the Gauss–Newton/Fisher term provides a natural approximation to the exact Hessian. Motivated by this bound, together with the fact that it is positive semi-definite and standard in large-scale optimization, we adopt this term as the curvature matrix used throughout the paper. With ℓ_2 regularization the surrogate becomes $H(\theta) = \frac{1}{M} J(\theta)^\top J(\theta) + \lambda I$. Defining $C := J(\theta)/\sqrt{M}$, we obtain $H(\theta) = C^\top C + \lambda I$, so the resulting curvature matrix is a regularized Gram matrix of Jacobian columns.

Although the exposition above uses MSE for clarity, our experiments focus on classification models. In that setting, the empirical Fisher and generalized Gauss–Newton matrices play the analogous role, and the arguments in Sections 3.2 and 3.3 depend only on the resulting damped Gram-type curvature structure.

Main diagonalization result (log-determinant ratio) A diagonal Fisher matrix corresponds to curvature directions that are decoupled in the chosen parameter coordinates, while off-diagonal entries encode parameter coupling and redundancy. To study how pruning affects these couplings, we seek a global, scale-invariant measure that vanishes exactly when curvature is diagonal and increases as parameter interactions grow.

Let $S \succ 0$ be positive definite and write $S = Q^\top Q$. By Hadamard's inequality (Różański et al., 2017) $\det(S) = \det(Q)^2 \leq \prod_{i=1}^n \|v_i\|^2 = \prod_{i=1}^n s_{ii}$, where v_i is the ith column of Q. Hence, the determinant of S is upper bounded by the product of its diagonal entries. In particular, curvature matrices arising from Fisher/Gauss–Newton constructions are typically positive semi-definite; throughout we use a damped curvature matrix (e.g., $H = C^\top C + \lambda I$ with $\lambda > 0$) to ensure positive definiteness and well-defined log-determinants.

Now let $C = [c_1, \ldots, c_P] \in \mathbb{R}^{M \times P}$ and define $H = C^\top C + \lambda I$. For any positive definite $H \succ 0$, define the determinant-based diagonality metric

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

We refer to $f(H)$ as the *log-determinant ratio* (or LogDet, for short). By Hadamard's inequality, $f(H) \geq 0$, with equality if and only if H is diagonal. Moreover, up to a factor of 2, $f(H)$ equals the Kullback–Leibler divergence between the Gaussians $\mathcal{N}(\mu, \mathrm{diag}(H)^{-1})$ and $\mathcal{N}(\mu, H^{-1})$ (see Appendix A.1).

Removing a parameter corresponds to deleting a column c_i and replacing H with $H_{-i} := C_{-i}^\top C_{-i} + \lambda I$ where C_{-i} is C with column c_i removed.

Theorem 3.2 (Log-det Gram-matrix diagonalization theorem). *Let $C = [c_1, \ldots, c_P] \in \mathbb{R}^{M \times P}$ and $H = C^\top C + \lambda I$ with $\lambda > 0$. For any index i, let $H_{-i} := C_{-i}^\top C_{-i} + \lambda I$. Then $f(H_{-i}) \leq f(H)$ with strict inequality if $C_{-i}^\top c_i \neq 0$. Moreover, writing H in block form as*

$$H = \begin{pmatrix} B & a \\ a^\top & d \end{pmatrix}, \qquad B = C_{-i}^\top C_{-i} + \lambda I, \qquad a = C_{-i}^\top c_i, \;\; d = \|c_i\|^2 + \lambda,$$

the decrease is given by $f(H) - f(H_{-i}) \;=\; \log\!\left(\frac{d}{d - a^\top B^{-1} a}\right)$.

Proof: See Appendix A.3.

Theorem 3.2 does not depend on how parameter i was selected. Thus, under a fixed Jacobian, magnitude pruning, random pruning, sensitivity-based criteria, and other coordinate-removal rules share the same one-step structural effect: deleting a coupled coordinate cannot increase $f(H)$. This does not imply that different pruning algorithms produce the same curvature after training. Retraining, rewinding, and dynamic connectivity modify C, and may therefore amplify, attenuate, or partially reverse the one-step effect.

The monotonicity result applies to $f(H)$ itself. Because pruning also changes matrix dimensionality, monotonicity of $f(H)/P$ does not follow directly from Theorem 3.2. We use the normalized quantity empirically to compare curvature matrices of different sizes.

Implicit natural-gradient behavior under diagonal adaptive methods We formalize the connection between pruning-driven diagonalization and natural-gradient behavior in the presence of *diagonal adaptive optimizers* such as Adam or RMSProp. Let $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ denote the standard gradient, and $\hat{g}_t = \widehat{D}_t^{-1} g_t$ be the update direction produced by a diagonal adaptive optimizer, where $\widehat{D}_t$ is a positive diagonal matrix constructed from running gradient statistics (e.g., the second-moment accumulator in Adam).

Natural gradient instead uses $\tilde{g}_t = H(\theta_t)^{-1} g_t$ where $H(\theta_t)$ denotes the Gauss–Newton (or Fisher) curvature matrix. To quantify how close $H(\theta_t)$ is to diagonal, we use $f(H)$ from equation 1. Writing $H = D^{1/2}(I + E)D^{1/2}$ with $D = \mathrm{diag}(H)$ and E the normalized off-diagonal part (zero diagonal), we have $f(H) = -\log\det(I + E)$ and in the near-diagonal regime, $f(H) = -\mathrm{Tr}\Big(\log(I + E)\Big) \approx \frac{1}{2}\|E\|_F^2$ where the approximation follows from the series expansion $\log(I + E) = E - \frac{1}{2}E^2 + O(\|E\|^3)$ and $\mathrm{Tr}(E) = 0$ (Appendix A.5). Thus, small $f(H)$ implies small normalized coupling.

Definition 3.3 (Near-diagonal regime LOGDET)**.** Training is in the near-diagonal regime if $f(H(\theta_t)) \le \varepsilon$ for small ε, with $H(\theta_t) \succ 0$.

The contribution of the following theorem is conceptual: it links the empirically observed pruning-induced diagonalization of curvature to the behavior of diagonal adaptive optimizers. It shows that under pruning-driven diagonalization and accurate diagonal curvature tracking, diagonal adaptive optimizers behave increasingly like natural-gradient methods. Rather than independently measuring the preconditioner's tracking accuracy, our cosine-alignment experiments (Section 4) serve as an operational validation of the theorem's combined prediction.

Theorem 3.4 (Implicit natural-gradient convergence for diagonal methods)**.** *Let $H_t \succ 0$ and write $H_t = D_t^{1/2}(I + E_t)D_t^{1/2}, \qquad D_t = \mathrm{diag}(H_t)$ where E_t is symmetric with zero diagonal. Let $g_t = \nabla_\theta \mathcal{L}(\theta_t) \neq 0$ and $\tilde{g}_t = H_t^{-1} g_t$. Let a diagonal adaptive method produce $\hat{g}_t = \widehat{D}_t^{-1} g_t$ where $\widehat{D}_t$ is positive diagonal. Assume there exist scalars $\alpha_t > 0$ such that*

$$\max_{1 \le i \le P} \frac{|(\widehat{D}_t)_{ii} - \alpha_t (D_t)_{ii}|}{\alpha_t (D_t)_{ii}} \;\le\; \varepsilon_t, \qquad \varepsilon_t \to 0.$$

Assume further that the pruning trajectory enters the near-diagonal regime in the sense that $\|E_t\|_2 \le \delta_t, \qquad \delta_t \to 0$. Finally, assume the diagonal scalings are uniformly well-conditioned: there exist constants $0 < \lambda \le \Lambda < \infty$ such that for all t, $\lambda \le (D_t)_{ii} \le \Lambda, \qquad \forall i$. (For example, the lower bound is ensured if one uses a damped curvature $H_t = G_t + \lambda I$ with $\lambda > 0$.) Then the angle between $\hat{g}_t$ and $\tilde{g}_t$ vanishes: $\angle(\hat{g}_t, \tilde{g}_t) \to 0$.

Proof: See Appendix A.4.

When H_t is near-diagonal, H_t^{-1} is also near-diagonal: specifically, $H_t^{-1} = D_t^{-1/2}(I + E_t)^{-1}D_t^{-1/2}$ and $(I + E_t)^{-1} = I + O(\|E_t\|)$. Thus the natural-gradient direction $\tilde{g}_t$ is close (up to small coupling terms) to a per-coordinate rescaling by D_t^{-1}. If the diagonal optimizer produces $\widehat{D}_t^{-1} \approx (\alpha_t D_t)^{-1}$, then $\hat{g}_t$ matches this rescaling up to vanishing relative error, and the remaining discrepancy comes from the small off-diagonal component E_t. These two errors together imply asymptotic alignment.

Practical implications Iterative pruning does not merely shrink the parameter space; it also yields a *structural simplification of curvature*. Along pruning trajectories for which $f(H)$ decreases

toward 0, the dominant curvature corrections effectively decouple across coordinates. In this regime, curvature behaves like a collection of one-dimensional scaling factors rather than a fully coupled matrix. Consequently, diagonal adaptive optimizers such as Adam operate in a geometry that is increasingly aligned with the true curvature structure: their coordinate-wise normalization approaches the coordinate-wise rescaling implicit in natural-gradient updates. In this sense, a pruning trajectory that substantially diagonalizes the curvature can act as a form of structural preconditioning: without explicitly estimating the full curvature, it drives diagonal methods closer to natural gradient behavior in both update direction and conditioning.

4 Experimental Results

We use two complementary experimental designs. First, we use iterative magnitude pruning (IMP) as a controlled sparsification trajectory to study the evolution of Fisher coupling across architectures, datasets, and sparsity levels. Second, we compare qualitatively different pruning methods under common architecture/dataset settings to determine whether the behavior observed under IMP extends to other mechanisms for inducing sparsity. Additional experiments examine sensitivity to the damping parameter, the relationship between Fisher diagonalization and natural-gradient optimization, and whether the observed effect can be explained solely by reduced parameter count.

For the broader IMP experiments, we first evaluate a small convolutional model, referred to as "ConvSmall", which consists of two blocks of convolutional layers and max pooling followed by a linear layer for classification. This model is used for experiments on smaller, classical datasets, such as MNIST (LECUN, 1998), EMNIST (Cohen et al., 2017), FashionMNIST (Xiao et al., 2017), and KMNIST (Clanuwat et al., 2018), with the advantage of enabling the calculations for the complete FIM (space complexity $O(n^2)$) on our computational resources. Additionally, we also run ResNet18 (He et al., 2015) and DenseNet (Huang et al., 2018) trained on CIFAR10 (Krizhevsky et al., 2014), SVHN (Netzer et al., 2011), and STL10 (Coates et al., 2011) to obtain layer-wise approximations of the FIM. For these IMP experiments, the FIM is evaluated at the end of each prune–retrain cycle, immediately before the next pruning step, to characterize the geometry of the subnetwork obtained at each density.

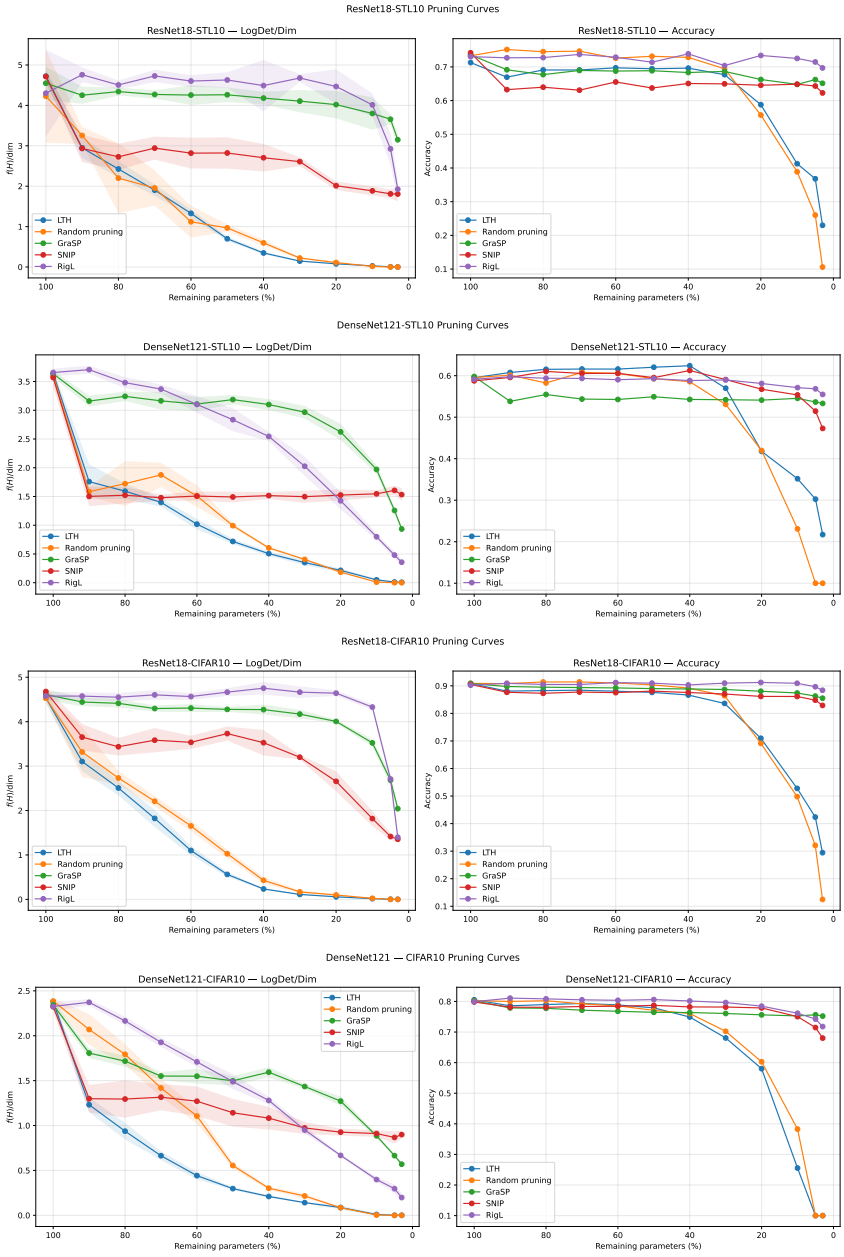
4.1 Fisher Coupling Under Network Pruning

We study Fisher coupling across both IMP and alternative sparsification procedures. For ConvSmall, the model size allows us to compute the full-network FIM. For ResNet18 and DenseNet121, memory constraints require a layer-wise approximation: we evaluate convolutional layers distributed throughout the network depth and subsample parameters within each selected layer, capping the largest FIM at 8000×8000. We report the log-determinant ratio per dimension (LogDet/dim) to compare matrices of different dimensionalities. As discussed in Section 3, Theorem 3.2 applies to the unnormalized LogDet, so the behavior of LogDet/dim along a complete train-and-prune trajectory remains an empirical question. Sensitivity to the damping parameter is analyzed in Appendix C.

We first use IMP as a controlled sparsification trajectory. Across ConvSmall, ResNet18, and DenseNet, Fisher coupling generally decreases as sparsity increases, and the contraction is observed across datasets and network depths. The trend is strongest throughout the moderate sparsity regime, while non-monotonic behavior can appear at extreme sparsity. For example, ConvSmall occasionally exhibits a rebound in LogDet/dim at the lowest densities. This is consistent with the distinction between the fixed-Jacobian result of Theorem 3.2 and a complete prune-and-retrain trajectory: subsequent optimization can reorganize the Jacobian and partially reverse the immediate effect of coordinate removal.

We next ask whether this behavior is specific to IMP. Figure 1 compares IMP/LTH with random pruning, SNIP, GraSP, and RigL at matched parameter densities for ResNet18 and DenseNet121 on STL10 and CIFAR10. The corresponding ConvSmall results on MNIST and KMNIST are reported in Appendix B.

Reduced Fisher coupling is observed beyond IMP, but the methods follow markedly different trajectories. IMP and random pruning generally produce the strongest contraction, whereas SNIP, GraSP, and RigL retain different amounts of off-diagonal structure depending on architecture, dataset, and



ResNet18 (a) STL10

DenseNet121 (b) STL10

ResNet18 (c) CIFAR10

DenseNet121 (d) CIFAR10

Figure 1: Fisher coupling and accuracy across pruning methods for ResNet18 and DenseNet121 on STL10 and CIFAR10. Reduced coupling extends beyond IMP, while the trajectories depend on the pruning method, architecture, and dataset.

sparsity. Thus, sparsification exhibits a common tendency toward reduced coupling, but the geometry of the trained sparse model is not determined by density alone.

The accuracy curves further separate this geometric effect from predictive performance. Models with similar accuracy can exhibit substantially different Fisher coupling, while at extreme sparsity LogDet/dim can continue to decrease even after accuracy begins to deteriorate. Consequently, LogDet/dim should be interpreted as a descriptor of sparse-network geometry rather than as a measure of pruning quality.

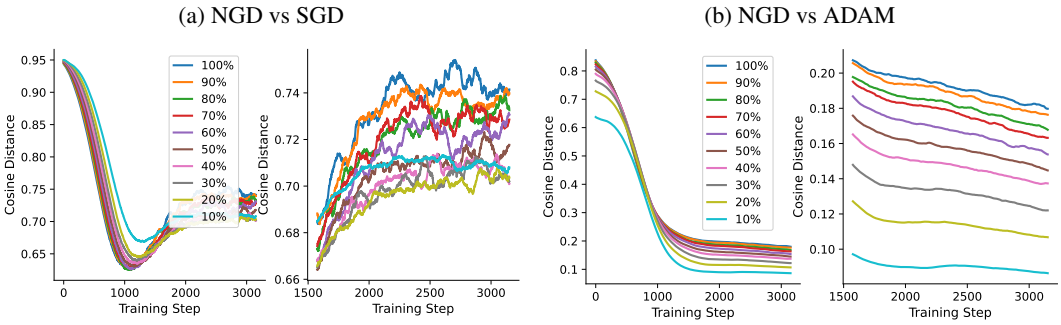


Figure 2: Cosine distance to NGD for SGD and Adam on DenseNet/CIFAR10 across IMP densities. Each panel shows the full trajectory and a late-training zoom.

4.2 Adam under iterative magnitude pruning as an approximation to NGD

Having characterized Fisher diagonalization across pruning methods, we next investigate one possible optimization consequence of this geometric change. Because the paired optimizer comparison is considerably more expensive, we study this question within the controlled IMP setting.

In order to assess the similarity between Adam and NGD on pruned networks, we compare the update step of both algorithms via cosine distance. The smaller the distance, the closer the two updates are. To control for the fact that pruned networks have many zeroed parameters and, consequently, would yield a smaller cosine distance by default, we exclude pruned dimensions from the calculation.

We use the SINGD (Lin et al., 2024) implementation of NGD. To measure the distance between Adam and NGD update steps, we compare both updates at the same parameters along a fixed optimization path. Concretely, we follow the Adam trajectory $\{\theta_t\}$: at iteration t, we compute Adam's update Δ_t^{Adam} and NGD's update Δ_t^{NGD} at θ_{t-1} (using the same mini-batch), but we apply only Adam's update to obtain θ_t. This keeps the trajectory fixed to the one produced by Adam while enabling a step-by-step comparison between Δ_t^{Adam} and Δ_t^{NGD}.

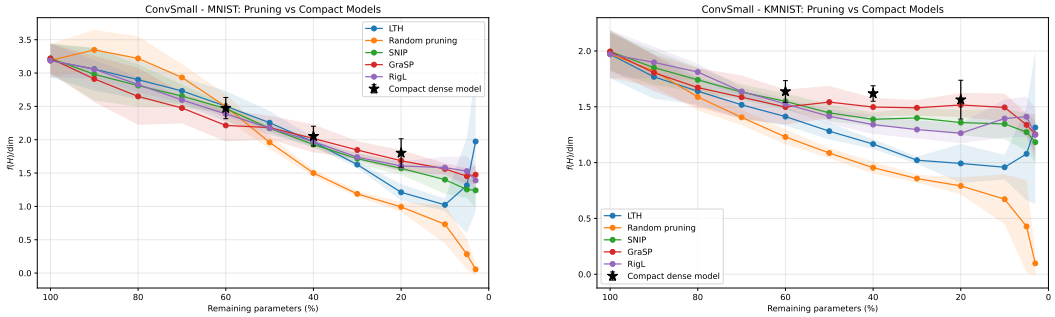
Figure 2 compares SGD–NGD and Adam–NGD cosine distances for DenseNet on CIFAR10 across IMP densities. As sparsity increases, Adam becomes systematically more aligned with NGD, with the separation between densities becoming particularly clear late in training. In contrast, SGD remains substantially farther from NGD and does not exhibit the same pruning-dependent alignment.

These observations are consistent with the prediction of Section 3: as Fisher coupling decreases, natural-gradient preconditioning becomes increasingly coordinate-wise, making a diagonal adaptive optimizer a better approximation to NGD. Additional Adam–NGD results are reported in Appendix D, with additional SGD–NGD baselines in Appendix E.

4.3 Reduced Fisher Coupling Is Not Explained by Parameter Count Alone

The reduction in LogDet observed under pruning could arise simply because the resulting sub-networks contain fewer parameters. To control for this possibility, we construct compact dense variants of ConvSmall with approximately 60%, 40%, and 20% of the parameters of the original architecture. These models preserve the main sequence of layer types while reducing their widths and are trained from scratch without pruning. We compare their normalized Fisher log-determinant ratio, $f(H)/\dim$, with those of IMP, random pruning, SNIP, GraSP, and RigL at matched parameter fractions on MNIST and KMNIST.

Figure 3 shows that compact dense models generally retain larger $f(H)/\dim$ than pruned networks with comparable parameter counts. The difference is particularly clear relative to random pruning and IMP at higher sparsity, while the compact models are closer to GraSP and some uncertainty intervals overlap. Thus, reducing model width does not reproduce the same reduction in Fisher coupling produced by sparsification. Parameter count alone therefore does not explain the observed



(a) MNIST (b) KMNIST

Figure 3: Fisher coupling for pruned and compact dense ConvSmall models. Curves show the mean $f(H)/\dim$ for five pruning methods, with shaded regions denoting one standard deviation. Black stars represent compact dense models trained from scratch at approximately 60%, 40%, and 20% of the original parameter count, with error bars denoting one standard deviation.

diagonalization; both the pruning mask and the method used to obtain it influence the curvature structure of the final network.

5 Conclusion

We studied how neural-network sparsification reshapes Fisher geometry. Under a fixed-Jacobian damped Gram model, coordinate removal monotonically decreases the unnormalized LogDet measure independently of the pruning rule. Empirically, IMP produces broad Fisher diagonalization across architectures and datasets, while random pruning, SNIP, GraSP, and RigL show that reduced coupling extends beyond lottery-ticket pruning but follows method-dependent trajectories. Parameter count alone does not explain this effect, and lower LogDet does not necessarily imply higher accuracy. Along IMP trajectories, increasing diagonality is also associated with closer Adam–NGD alignment. Overall, sparsification tends to simplify Fisher geometry, while the resulting structure depends on how sparsity is imposed and optimized.

6 AI Use Disclosure

Generative AI tools were used to assist with language refinement and the identification of potentially relevant related work. All scientific content, experimental results, analyses, and conclusions were produced and reviewed by the authors. References suggested by AI tools were independently verified by the authors before inclusion.

References

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

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

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models, 2021. URL https://arxiv.org/abs/2012.06908.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018. URL http://arxiv.org/abs/1812.01718.

Adam Coates, Andrew Ng, and Honglak Lee. 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.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: an extension of mnist to handwritten letters, 2017. URL https://arxiv.org/abs/1702.05373.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners, 2021. URL https://arxiv.org/abs/1911.11134.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

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

Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, et al. The cifar-10 dataset, 2014.

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

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: Single-shot network pruning based on connection sensitivity, 2019. URL https://arxiv.org/abs/1810.02340.

Wu Lin, Felix Dangel, Runa Eschenhagen, Kirill Neklyudov, Agustinus Kristiadi, Richard E. Turner, and Alireza Makhzani. Structured inverse-free natural gradient: Memory-efficient & numerically-stable kfac, 2024. URL https://arxiv.org/abs/2312.05705.

Bohan Liu, Zijie Zhang, Peixiong He, Zhensen Wang, Yang Xiao, Ruimeng Ye, Yang Zhou, Wei-Shinn Ku, and Bo Hui. A survey of lottery ticket hypothesis, 2024. URL https://arxiv.org/abs/2403.04861.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature, 2020. URL https://arxiv.org/abs/1503.05671.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/18bb68e2b38e4a8ce7cf4f6b2625768c-Paper.pdf.

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

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow, 2020. URL https://arxiv.org/abs/2006.05467.

Ming Tu, Visar Berisha, Yu Cao, and Jae sun Seo. Reducing the model order of deep neural networks using information theory, 2016. URL https://arxiv.org/abs/1605.04859.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow, 2020. URL https://arxiv.org/abs/2002.07376.

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

A PROOFS

A.1 LOGDET AND KL DIVERGENCE

This subsection provides a probabilistic interpretation of the LOGDET:

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

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

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

Thus, up to a factor of 2, the LOGDET equals the KL divergence between a Gaussian with diagonal precision D and the Gaussian with full precision H.

A.2 PROOF OF PROPOSITION 3.1

Proof. We consider a scalar-output ReLU MLP inside a region of parameter space where the ReLU sign pattern on the training samples is fixed. In such a region, each layer admits a constant diagonal mask D_ℓ, and the network output can be written as

$$\hat{y}(x,\theta) = W_N D_{N-1} W_{N-1} \cdots D_1 W_1 x, \qquad N := L+1,$$

where W_N is the output row vector. Biases are omitted for simplicity; the same argument applies after augmenting each layer input with a constant 1.

Let $\theta = (\theta^{(1)}, \ldots, \theta^{(N)})$, where $\theta^{(r)} := \mathrm{vec}(W_r)$.

Same-layer second derivatives. Fix all layers except W_r. Then the output is affine in W_r:

$$\hat{y}(x,\theta) = a_r^\top W_r b_{r-1},$$

where a_r and b_{r-1} depend only on the remaining layers and the input. Hence

$$\nabla^2_{\theta^{(r)}\theta^{(r)}}\hat{y}(x,\theta) = 0.$$

Cross-layer mixed blocks. Let $r \neq s$, and assume, without loss of generality, $r < s$. For perturbations Δ_r, Δ_s, multilinearity across layer blocks gives

$$d^2\hat{y}(x,\theta)[\Delta_r,\Delta_s] = W_N D_{N-1}\cdots D_s\Delta_s D_{s-1}\cdots D_r\Delta_r D_{r-1}\cdots D_1 W_1 x.$$

Using $\|D_\ell\|_2 \le 1$ and submultiplicativity,

$$|d^2\hat{y}(x,\theta)[\Delta_r,\Delta_s]| \le \|x\|_2 \Big(\prod_{\ell\neq r,s} \|W_\ell\|_2\Big)\|\Delta_r\|_2\|\Delta_s\|_2.$$

Therefore,

$$\big\|\nabla^2_{\theta^{(r)}\theta^{(s)}}\hat{y}(x,\theta)\big\|_2 \le \|x\|_2 \prod_{\ell\neq r,s} \|W_\ell\|_2.$$

Bound on the second term. Let

$$R(\theta) := \frac{1}{M}\sum_{m=1}^{M} e_m \nabla^2_\theta \hat{y}_m.$$

Since the block diagonal terms vanish, summing the block norms and using symmetry, we have

$$\|\nabla^2_\theta \hat{y}(x_m,\theta)\|_2 \le 2\|x_m\|_2 \sum_{1\le r<s\le N} \prod_{\ell\neq r,s} \|W_\ell\|_2.$$

Since

$$\|R(\theta)\|_2 \le \frac{1}{M}\sum_{m=1}^{M} |e_m|\,\|\nabla^2_\theta \hat{y}_m\|_2,$$

we have

$$\|R(\theta)\|_2 \le \frac{2}{M}\sum_{m=1}^{M} |e_m|\,\|x_m\|_2 \sum_{1\le r<s\le N} \prod_{\ell\neq r,s} \|W_\ell\|_2.$$

This proves the proposition.

□

A.3 Proof of Theorem 3.2

Proof. Fix an index i and write

$$C = [C_{-i},\, c_i], \qquad H = C^\top C + \lambda I.$$

Then H admits the block form

$$H = \begin{pmatrix} B & a \\ a^\top & d \end{pmatrix}, \qquad \begin{aligned} B &:= C_{-i}^\top C_{-i} + \lambda I, \\ a &:= C_{-i}^\top c_i, \\ d &:= \|c_i\|^2 + \lambda. \end{aligned}$$

Because $\lambda > 0$, we have $B \succ 0$ and $H \succ 0$.

Let $D := \mathrm{diag}(H)$ and $D_B := \mathrm{diag}(B)$. Noting that the diagonal of H consists of the diagonal of B together with the scalar entry d, we have

$$\det(D) = \det(D_B)\cdot d.$$

Moreover, by the Schur complement formula for block determinants,

$$\det(H) = \det(B)\cdot\big(d - a^\top B^{-1} a\big).$$

(Positive definiteness implies $d - a^\top B^{-1}a > 0$.)

Therefore

$$\begin{aligned} f(H) = \log\frac{\det(D)}{\det(H)} &= \log\frac{\det(D_B)\,d}{\det(B)\,(d - a^\top B^{-1}a)} \\ &= \underbrace{\log\frac{\det(D_B)}{\det(B)}}_{=f(B)} + \log\Big(\frac{d}{d - a^\top B^{-1}a}\Big). \end{aligned}$$

But pruning parameter i replaces H by

$$H_{-i} = C_{-i}^\top C_{-i} + \lambda I = B,$$

so $f(H_{-i}) = f(B)$. Hence

$$f(H) - f(H_{-i}) = \log\Big(\frac{d}{d - a^\top B^{-1} a}\Big) \;\geq\; 0.$$

Moreover, $a^\top B^{-1} a > 0$ iff $a \neq 0$, in which case the logarithm is strictly positive. Thus $f(H_{-i}) \leq f(H)$, with strict inequality whenever $C_{-i}^\top c_i \neq 0$. □

A.4 Proof of Theorem 3.4

Proof. Write $H_t = D_t^{1/2}(I + E_t)D_t^{1/2}$ as in the theorem statement and note that

$$H_t^{-1} = D_t^{-1/2}(I + E_t)^{-1} D_t^{-1/2}.$$

Assume $\|E_t\|_2 \leq \delta_t$ with $\delta_t \to 0$. For all sufficiently large t, $\delta_t < 1$ and the Neumann expansion gives

$$(I + E_t)^{-1} = I - E_t + R_t, \qquad \|R_t\|_2 \leq \frac{\|E_t\|_2^2}{1 - \|E_t\|_2} \leq C\,\delta_t^2$$

for some constant C (independent of t) and all large t.

Define the "diagonal-preconditioned" direction

$$v_t := D_t^{-1} g_t.$$

Also define the whitened gradient $u_t := D_t^{-1/2} g_t$, so that $v_t = D_t^{-1/2} u_t$ and $u_t = D_t^{1/2} v_t$.

The natural-gradient direction is

$$\tilde{g}_t = H_t^{-1} g_t = D_t^{-1/2}(I + E_t)^{-1} u_t.$$

Thus

$$\tilde{g}_t - v_t = D_t^{-1/2}\Big((I + E_t)^{-1} - I\Big)u_t = D_t^{-1/2}\Big(-E_t + R_t\Big)u_t.$$

Taking norms and using submultiplicativity,

$$\|\tilde{g}_t - v_t\| \leq \|D_t^{-1/2}\|_2\,(\|E_t\|_2 + \|R_t\|_2)\,\|u_t\|.$$

By the uniform conditioning assumption, $\|D_t^{-1/2}\|_2 \leq 1/\sqrt{\lambda}$ and $\|u_t\| = \|D_t^{1/2} v_t\| \leq \|D_t^{1/2}\|_2 \|v_t\| \leq \sqrt{\Lambda}\,\|v_t\|$. Therefore

$$\|\tilde{g}_t - v_t\| \leq \sqrt{\frac{\Lambda}{\lambda}}\,(\delta_t + C\delta_t^2)\,\|v_t\| \leq C_1\,\delta_t\,\|v_t\|$$

for a constant C_1 and all large t.

Now compare v_t to the diagonal method update $\hat{g}_t = \widehat{D}_t^{-1} g_t$. Let $w_t := \alpha_t \hat{g}_t$. Componentwise,

$$w_{t,i} = \alpha_t \frac{g_{t,i}}{(\widehat{D}_t)_{ii}}, \qquad v_{t,i} = \frac{g_{t,i}}{(D_t)_{ii}}.$$

The assumption

$$\max_i \frac{|(\widehat{D}_t)_{ii} - \alpha_t (D_t)_{ii}|}{\alpha_t (D_t)_{ii}} \leq \varepsilon_t$$

implies (for all large t) that w_t and v_t are close: there exists $C_2 > 0$ such that

$$\|w_t - v_t\| \leq C_2\,\varepsilon_t\,\|v_t\|.$$

Combining the two comparisons,

$$\|w_t - \tilde{g}_t\| \leq \|w_t - v_t\| + \|v_t - \tilde{g}_t\| \leq (C_2\varepsilon_t + C_1\delta_t)\,\|v_t\|.$$

Since $\varepsilon_t \to 0$ and $\delta_t \to 0$, the relative error vanishes. Moreover, under $\|E_t\|_2 \leq \delta_t < 1$, $(I + E_t)^{-1}$ has bounded condition number, and together with $\lambda I \preceq D_t \preceq \Lambda I$ this yields $\|\tilde{g}_t\| \asymp \|v_t\|$ for large t. Therefore w_t and $\tilde{g}_t$ become asymptotically aligned, and since w_t is a scalar multiple of $\hat{g}_t$, we conclude

$$\angle(\hat{g}_t, \tilde{g}_t) \to 0.$$

□

A.5 Connection Between $f(H)$ and Normalized Off–Diagonal Energy

We use the LogDet:

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

which satisfies $f(H) \geq 0$ with equality iff H is diagonal, by Hadamard's inequality. This subsection provides a local expansion of $f(H)$ around the diagonal regime and clarifies its relationship with normalized off–diagonal curvature.

Let $D = \mathrm{diag}(H)$ and define

$$E = D^{-1/2}(H - D)D^{-1/2},$$

so that E is symmetric with zero diagonal and

$$H = D^{1/2}(I + E)\, D^{1/2}.$$

By multiplicativity of the determinant,

$$\det(H) = \det(D)\ \det(I + E),$$

and therefore

$$f(H) = -\log\det(I + E).$$

We now assume we are in the *near–diagonal regime*, meaning that $\|E\|_2 < 1$ so that the matrix logarithm admits its classical power series representation. Then

$$\log\det(I + E) = \mathrm{tr}\big(\log(I + E)\big) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}\,\mathrm{tr}(E^k).$$

Since E is symmetric with zero diagonal, $\mathrm{tr}(E) = 0$, so the linear term vanishes and the expansion begins at second order:

$$\log\det(I + E) = -\frac{1}{2}\mathrm{tr}(E^2)\ +\ O(\|E\|^3).$$

Because $\mathrm{tr}(E^2) = \|E\|_F^2$, we obtain

$$f(H) = \frac{1}{2}\|E\|_F^2 + O(\|E\|^3).$$

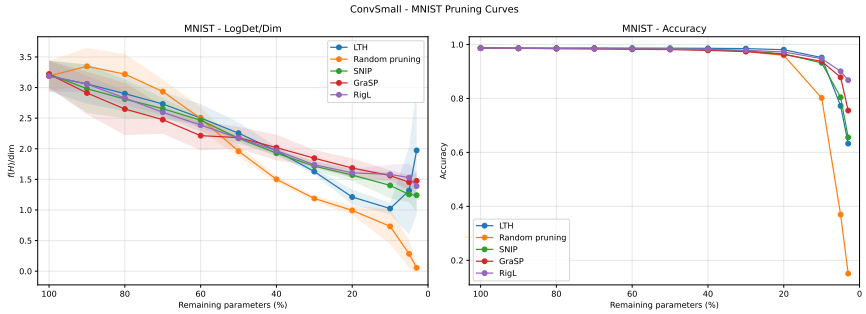
Thus, in the near–diagonal regime, $f(H)$ is (to second order) a positive quadratic function of the normalized off–diagonal energy of H. In particular, $f(H) \to 0$ implies $\|E\|_F \to 0$ and hence $\|E\|_2 \to 0$.

A.6 Curvature Estimators in Practice (Fisher vs. Gauss–Newton)

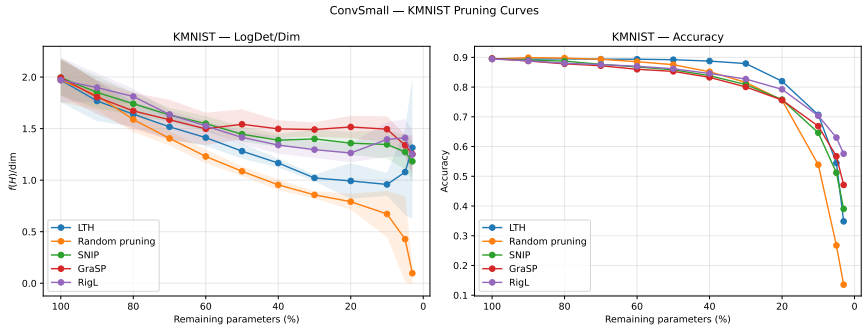
Throughout the paper, we use $H(\theta)$ to denote a positive semi-definite curvature matrix that captures local sensitivity. For squared-error regression on piecewise-linear networks, the Hessian decomposes into the Gauss–Newton term $J(\theta)^\top J(\theta)/M$ and a residual-weighted second-derivative term. Proposition 3.1 bounds this remainder within a fixed activation region, motivating the Gauss–Newton/Fisher surrogate when residuals are small. For classification objectives, the Fisher information matrix and the generalized Gauss–Newton matrix provide closely related PSD surrogates. Our theoretical results apply to any PSD curvature that can be written (after damping) in the regularized Gram form $H = C^\top C + \lambda I$.

B ConvSmall Across Pruning Methods

Figure 4 reports the corresponding side-by-side Fisher-coupling and classification-accuracy curves for ConvSmall on MNIST and KMNIST.



(a) MNIST



(b) KMNIST

Figure 4: Fisher coupling and classification accuracy across pruning methods for ConvSmall on MNIST and KMNIST. In each subfigure, the left panel reports $f(H)/\dim$ and the right panel reports test accuracy. Curves show the mean, and shaded regions denote one standard deviation.

C Sensitivity to Fisher Damping

Our curvature metric is evaluated on the damped matrix

$$H_\lambda = F + \lambda I.$$

The main experiments use $\lambda = 10^{-8}$. Because the diagonal shift affects the absolute magnitude of LogDet/dim, we evaluate whether the observed diagonalization could result from this particular choice of damping. We recompute the metric for $\lambda \in \{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}\}$ using ConvSmall on MNIST under IMP, random pruning, SNIP, GraSP, and RigL. Each curve reports the mean across ten runs, with shaded regions indicating one standard deviation.

Figure 5 shows that the diagonalization observed in the main experiments is not explained by $\lambda = 10^{-8}$ being excessively large. Reducing the damping to $\lambda = 10^{-10}$ preserves the sparsity-dependent contraction across pruning methods, indicating that the same broad geometric behavior remains when the damping contribution is substantially weaker. As λ increases relative to the empirical Fisher, the uniform diagonal shift reduces the relative influence of off-diagonal coupling on the log-determinant ratio, producing smaller initial values and flatter curves, particularly at $\lambda = 10^{-4}$.

Despite these changes in scale and dynamic range, the principal differences among pruning methods remain visible across the tested values. Random pruning and IMP generally produce stronger contractions, whereas SNIP, GraSP, and RigL retain comparatively higher coupling over much of the sparsity range. Overall, the ablation indicates that damping controls the magnitude and sensitivity of LogDet/dim, but the pruning-associated diagonalization reported in the main text is not an artifact of the main choice $\lambda = 10^{-8}$.

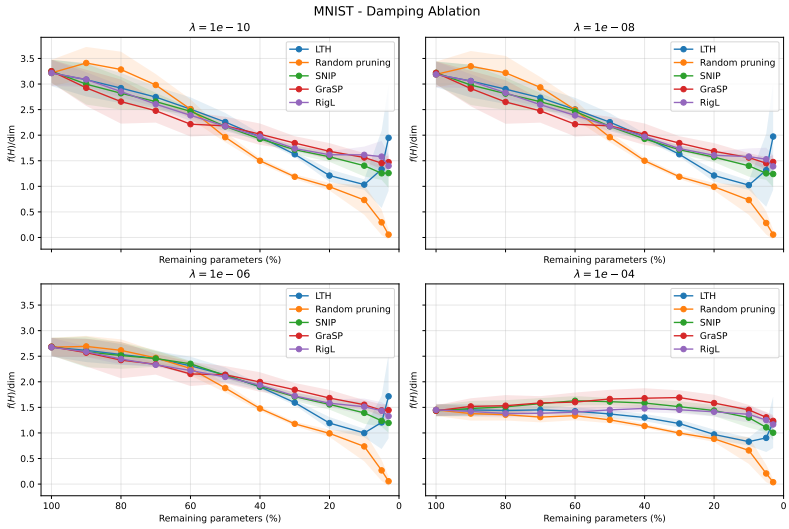
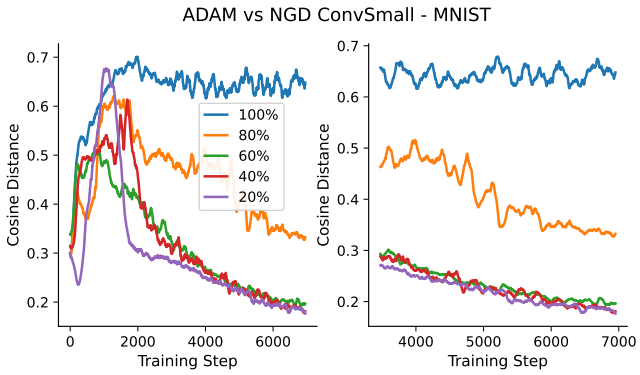
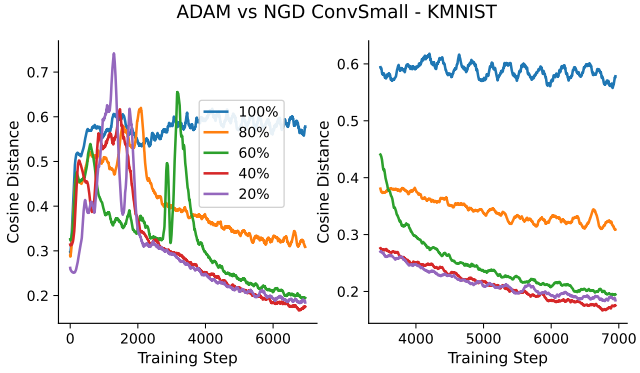


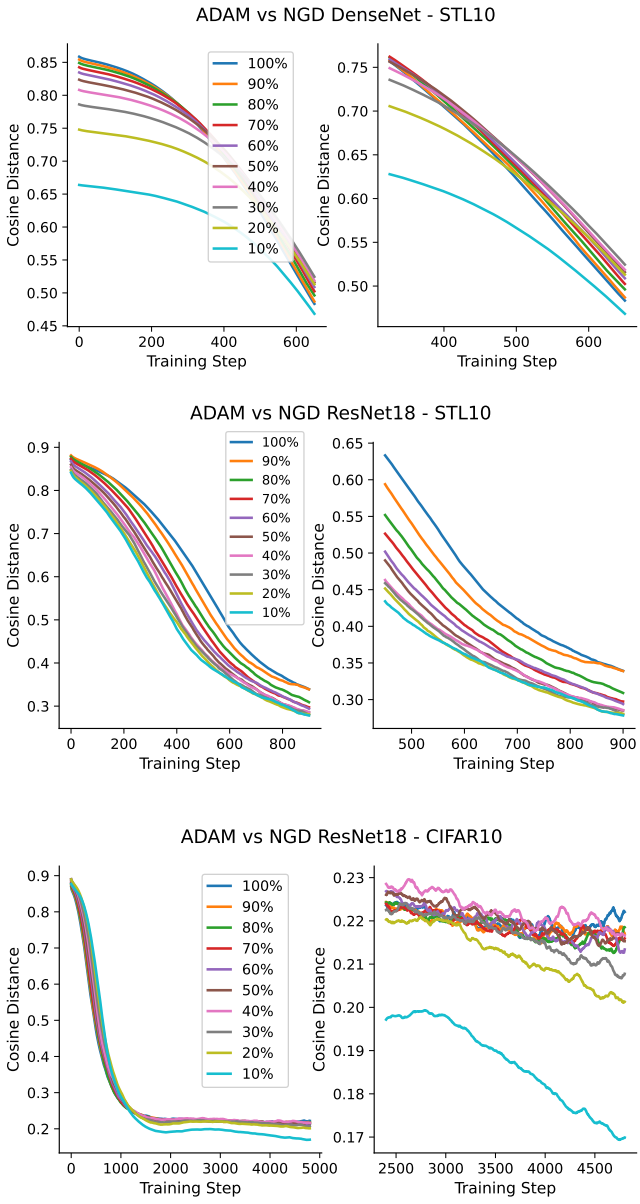
Figure 5: Sensitivity of Fisher coupling to the damping parameter on ConvSmall trained on MNIST. Each panel reports $f(H_\lambda)/\dim$ across pruning densities for a different value of λ. The main experiments use $\lambda = 10^{-8}$. Solid lines show the mean across ten runs, and shaded regions denote one standard deviation.

D Additional Adam–NGD Alignment Results

The figures below report additional Adam–NGD cosine-distance trajectories across architectures, datasets, and IMP densities. Consistent with the main results, increasing sparsity generally leads to stronger directional alignment.

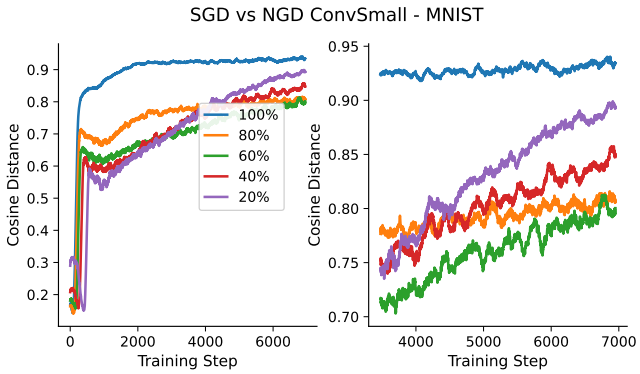


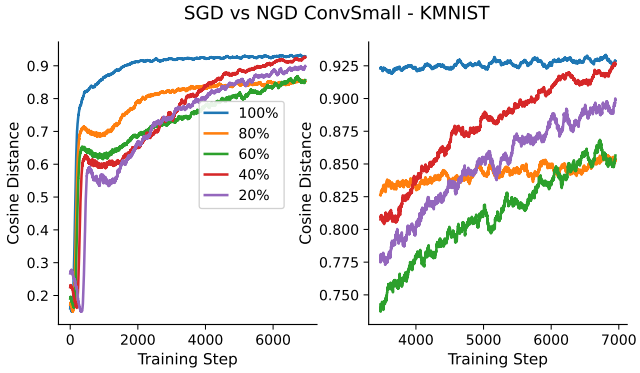


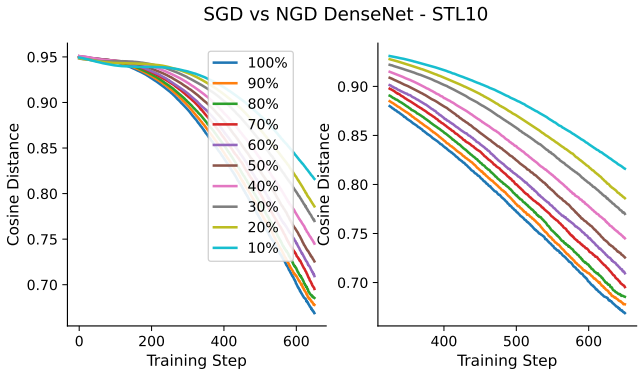


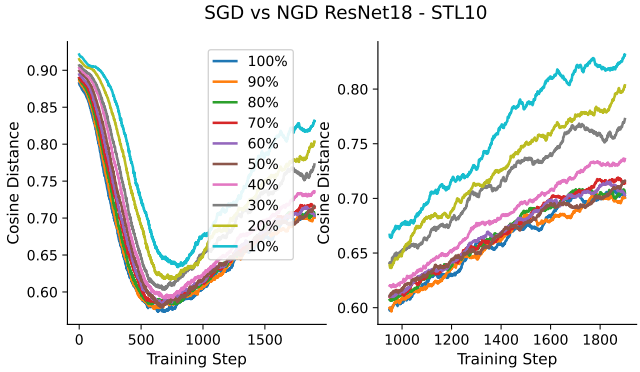
E SGD–NGD Cosine Distances

The figures below report additional SGD–NGD cosine-distance trajectories across architectures, datasets, and IMP densities. Unlike Adam, SGD does not show a consistent increase in directional alignment with sparsity.


SGD vs NGD ConvSmall - MNIST
Cosine Distance
Training Step
100%
80%
60%
40%
20%


SGD vs NGD ConvSmall - KMNIST
Cosine Distance
Training Step
100%
80%
60%
40%
20%


SGD vs NGD DenseNet - STL10
Cosine Distance
Training Step
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%


SGD vs NGD ResNet18 - STL10
Cosine Distance
Training Step
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%