FISHER GEOMETRY OF ACTIVATIONS FOR DEEP REPRESENTATION LEARNING: DIAGNOSIS AND MECHANISM

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual connections and normalization are central to deep-network training, yet their effects on layer-wise activation spectra remain incompletely characterized. The primary contribution of this paper is a staged diagnosis of those spectra and a local account of the measured patterns. We measure effective rank, participation ratio, and condition number of activation covariance, together with the Fisher information of a local Gaussian activation model. Under that Gaussian model, Fisher information for the mean is the precision, and layer linear Fisher information (LFI) is the squared length of a mean shift in this metric. LFI is separate from parameter Fisher information. Covariance eigenvalues are reciprocals of Fisher eigenvalues. On the tested MLP, CNN, and attention encoders, plain stacks often concentrate covariance energy, residual variants often have broader spectra, and per-channel standardization changes diversity and conditioning. These are coordinate-dependent measurements: normalization rescales those coordinates, so broader covariance is not invariant evidence of preserved information. Locally, a residual identity path changes gain propagation, but the leading Jacobian singular value does not determine effective rank. FisherNorm learns a per-channel standardization exponent from the task loss. Under the shared placement protocol, it is comparable to the strongest baselines.

1 INTRODUCTION

Deep networks now span substantial depth, width, and architectural variety, including attention-based encoders (Vaswani et al., 2017). A basic question remains: *how is information organized and transformed across layers?* We examine how residual connections and normalization change layer-wise activation spectra, beyond what parameter count or training loss alone can reveal. The primary results are the spectral diagnosis and its local Jacobian interpretation. FisherNorm learns the exponent of that standardization map from the task loss.

Several lines of work address complementary aspects of this problem. Optimization landscapes, gradient flow, covariate shift, and residual architectures explain aspects of successful deep-network training (He et al., 2016a; Ioffe & Szegedy, 2015; Ba et al., 2016); information geometry, in turn, provides tools for analyzing parameter-space curvature (Amari, 1998). These perspectives do not jointly account for (i) why residuals help across MLPs, CNNs, and Attention encoders; (ii) why normalization can facilitate optimization while *reshaping* activation spectra; or (iii) why the same normalization behaves differently across architectures. A common measurement language is needed to separate measured geometry, local interpretation, and task-level evaluation.

Fisher geometry of the activations provides that language. The spectrum of Σ describes how activation energy is spread. LFI is the Fisher information of a local Gaussian model of the same activations for a shift of their mean.

The study is organized around three questions. **RQ1:** How do activation spectra evolve with depth across plain, residual, and normalized architectures? **RQ2:** What local mechanisms—from Jacobian propagation to covariance rescaling—account for the measured patterns? **RQ3:** How competitive is a learnable per-channel standardization exponent when it is trained only by task loss? The answers

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.

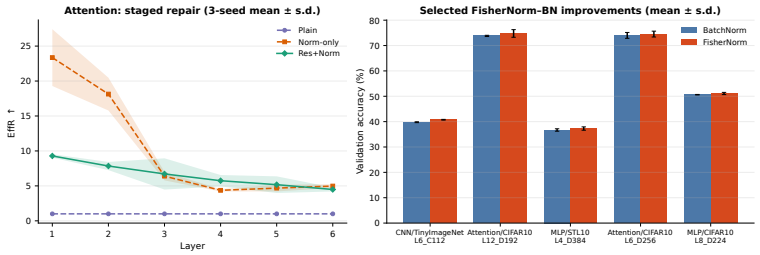

Figure 1: Diagnostic overview and a secondary task-level comparison. (a) Layer-wise EffR on the selected Attention/CIFAR-10 probe (L=6): seed means ± one sample s.d.; Plain concentrates to rank one, whereas Res+Norm has a broader measured spectrum and higher validation accuracy. (b) FisherNorm and BatchNorm on the five classification scenarios with the largest absolute validation gaps under Rule A.

to RQ1 and RQ2 are the primary contribution. RQ3 compares the learnable exponent with the other normalization families under the same placement protocol.

Our primary contributions are the diagnosis and the mechanism.

1. **Staged covariance-spectrum diagnosis.** Across MLP, CNN, and Attention encoders (13 datasets, 60 scenarios), plain chains often concentrate covariance energy with depth, residuals often broaden spectra, and fixed-strength normalization shifts the conditioning–diversity trade-off in a task-dependent way (Figs. 2–3; Fig. A.1).
2. **Local Jacobian interpretation.** Residuals introduce an identity term (J=$D_\phi W$+I), while normalization rescales coordinates. Layer norms $\|J^{(l)}\|_2$ check contraction versus expansion but cannot establish effective-rank broadening.
3. **Secondary parameterization study.** FisherNorm learns a per-channel exponent from the task loss and is compared with the other families under the same four placements.

Extended protocols and full tables appear in the appendix.

1.1 Related Work

Normalization and the standardization exponent. BatchNorm, LayerNorm, and later families—Batch Renorm, GroupNorm, decorrelated BN, Switchable Norm, RMSNorm, Filter Response Normalization, EvoNorm, PowerNorm, and ScaleNorm—specify how activations are recentered and rescaled (Ioffe & Szegedy, 2015; Ba et al., 2016; Ioffe, 2017; Wu & He, 2018; Huang et al., 2018; Luo et al., 2019; Zhang & Sennrich, 2019; Singh & Krishnan, 2020; Liu et al., 2020; Shen et al., 2020; Nguyen & Salazar, 2019). Their optimization benefit has been linked to smoother loss landscapes, while normalizer-free residual networks show that a normalization layer is not required for high accuracy (Santurkar et al., 2018; Brock et al., 2021). In Transformers, where the layer sits also affects stability (Xiong et al., 2020).

PowerNorm replaces BatchNorm's variance standardization by a fixed power computed at tensor level (Shen et al., 2020). FisherNorm trains a per-channel exponent $\lambda_i \in (0, 1)$ on the standardization map, with BatchNorm recovered in the limit $\lambda_i \to 1$. The experiments compare these two layers under the placement protocol in Section 3.

Residuals, rank collapse, and the spectrum measured here. Skip connections have been analyzed as identity maps and through shattered gradients, dynamical isometry, and signal propagation (He et al., 2016a;b; Balduzzi et al., 2017; Pennington et al., 2017; Schoenholz et al., 2017), including deep linear networks and normalization-free residual initialization (Saxe et al., 2014; Zhang et al., 2019; Bachlechner et al., 2021). Two rank results sit next to our plain-stack and attention findings. **?** prove that BatchNorm avoids rank collapse in randomly initialized deep *linear* networks. **?** prove that pure attention can lose rank doubly exponentially with depth, and that skip connections counteract that loss. Both are rank statements about random linear networks and about attention

maps, respectively. On trained nonlinear MLP, CNN, and Attention encoders, this paper records the eigenvalue spectrum of activation covariance together with LFI. $\|J\|_2$ is used only as a contraction check.

Parameter Fisher information. The parameter Fisher information matrix $F(\theta)$ induces a Riemannian metric on parameter space and underlies natural gradient, KFAC, and Fisher-based continual learning (Amari, 1998; Pascanu & Bengio, 2014; Ollivier, 2015; Martens & Grosse, 2015; Grosse & Martens, 2016; Kirkpatrick et al., 2017). Mean-field and representation analyses study that matrix at scale (Karakida et al., 2019; Achille et al., 2019); the neural tangent kernel is a different infinite-width linearization (Jacot et al., 2018). Empirical-Fisher products of gradients need not match this curvature (Kunstner et al., 2019). The encoders in this paper are Euclidean MLP, CNN, and Attention stacks with coordinate-wise normalization, and Section 2 defines the activation-space quantities computed from them.

2 Activation-Space Fisher Framework

This section defines the Fisher probes in activation space used throughout. Appendix A details the estimation protocol.

2.1 Activation Fisher Information

For a parametric model $p(y \mid x; \theta)$, the parameter Fisher information matrix is

$$F(\theta) = \mathbb{E}_{x,y}\left[\nabla_\theta \log p(y \mid x; \theta)\, \nabla_\theta \log p(y \mid x; \theta)^\top\right], \tag{1}$$

which induces a Riemannian metric on parameter space (Amari, 1998). We do not estimate $F(\theta)$. Let $a^{(l)}(x) \in \mathbb{R}^d$ be the post-activation at layer l, with empirical covariance $\Sigma^{(l)}$ and a mean displacement $\Delta\mu^{(l)}$. If these activations are modeled locally as Gaussian, the Fisher information for their mean is the precision $(\Sigma^{(l)})^{-1}$. The eigenvalues of $\Sigma^{(l)}$ are therefore not Fisher information. LFI, defined below, is the squared length of $\Delta\mu^{(l)}$ in that activation Fisher metric. It is not an entry of $F(\theta)$ and is not a natural-gradient preconditioner (Karakida et al., 2019; Achille et al., 2019; Raghu et al., 2017). EffR, PR_n, Top-10 energy, and κ remain functions of $\Sigma^{(l)}$. An invertible per-channel rescaling changes both this covariance spectrum and the activation Fisher metric, and Eqs. (5)–(7) apply exactly such a rescaling. A broader spectrum after normalization is not by itself evidence that information was preserved.

2.2 Linear Fisher Information

On the evaluation set, form the three input blocks (x^-, x^0, x^+) with $x^\pm = x^0 \pm \delta\mathbf{1}$ (δ=0.3). Let $\mu_\pm^{(l)}$ be the empirical mean of $a^{(l)}(x^\pm)$ over evaluation examples. The covariance, mean-sensitivity, and LFI estimators are Eqs. (8)–(10) in Appendix A.1, with a Moore–Penrose pseudoinverse (ε=10^{-9} truncation). LFI is directional: it measures perturbation response relative to internal variability and depends jointly on $\Delta\mu^{(l)}$ and the shape of $\Sigma^{(l)}$.

With $\Sigma^{(l)} = U^{(l)}\mathrm{diag}(s_1^{(l)}, \ldots, s_d^{(l)})\, U^{(l)\top}$ and $\alpha^{(l)} = U^{(l)\top}\Delta\mu^{(l)}$,

$$\mathrm{LFI}^{(l)} = \sum_{i:\, s_i^{(l)} > \varepsilon} \frac{\left(\alpha_i^{(l)}\right)^2}{s_i^{(l)}}, \tag{2}$$

matching that truncated pseudoinverse (directions with $s_i^{(l)} \le \varepsilon$ are dropped).

2.3 Fisher Distribution Metrics

Let $\bar{s}_i^{(l)} = s_i^{(l)} / \sum_j s_j^{(l)}$ and $\hat{\mu}^{(l)} = \Delta\mu^{(l)} / \|\Delta\mu^{(l)}\|_2$. Our entropy-based effective-rank statistic follows the continuous dimensionality measure of Roy & Vetterli (2007). Six layer-wise quantities are evaluated at each run's best validation epoch. Their definitions are Eqs. (11)–(14) in Appendix A.2. The main text emphasizes the primary triad $\{\mathrm{EffR}, \mathrm{PR}_n, E_{10}\}$; the appendix presents the complementary triad $\{\mathrm{LFI}, \log_{10}\kappa, \mathrm{Align}\}$ in matched layouts.

3 DIAGNOSIS: FISHER EVOLUTION IN PLAIN, RESIDUAL, AND NORMALIZED NETWORKS

To address RQ1, we add components one at a time: plain chains, then skips, then normalization. FisherNorm is held out here.

3.1 EXPERIMENTAL SETUP

The measurement and comparison rules are fixed as follows.

Scenario. One dataset at one depth and width. The suite has 13 datasets and 60 scenarios, with depth at most 12 and width at most 384 (Appendix B.1). Every encoder is a Euclidean MLP, CNN, or Attention stack.

Family. Each scenario trains Plain and nine normalizers: BatchNorm, LayerNorm, GroupNorm, RMSNorm, PowerNorm, EvoNorm, ScaleNorm, Filter Response Normalization, and FisherNorm. PowerNorm is the published layer of Shen et al. (2020), with one fixed power computed at tensor level. FisherNorm uses a separate learnable exponent on each channel. The tables compare these complete layers.

Placement. Each of the nine normalizers is trained at the same four placements:

$$\text{Pre: } h' = \phi(\mathrm{Norm}(Wh)), \qquad \text{Post: } h' = \mathrm{Norm}(\phi(Wh)),$$
$$\text{Res-Pre: } h' = \phi(\mathrm{Norm}(Wh)) + h, \qquad \text{Res-Post: } h' = \mathrm{Norm}(\phi(Wh) + h).$$

A variant is one family at one placement. With Plain, each scenario has $9 \times 4 + 1 = 37$ variants.

Which placement Fig. 2 uses. The staged diagnosis uses fixed placements. Norm-only is BatchNorm Pre, and Res+Norm is residual BatchNorm Pre, on CIFAR-10 (MLP L=8, CNN L=12, Attention L=6). Plain, Norm-only, and Res+Norm therefore differ by one architectural change.

Seed-trial. Seeds are $\{42, 123, 456\}$. A seed-trial is one scenario on one seed. Fifty-nine scenarios have three seeds and one scenario has seed 42 only, giving 178 trials.

Checkpoint. For every run, the spectrum and the validation score are read from the checkpoint with that run's best validation score.

Rule A, aggregate tables. In Table 1, each family is represented by the placement whose cross-seed mean validation score is highest. The table reports that same validation score.

Rule B, win counts. In Table 2, seeds are counted separately. Inside one seed, each family keeps the placement with the highest validation score on that seed, and the family with the highest score wins the trial. Every family chooses from the same four placements. This count gives FisherNorm $45/178$ wins.

Probe. Means and covariances are empirical moments over evaluation examples, grouped as the three blocks (x^-, x^0, x^+), not a population law for one input. The shift δ=0.3 is added to every stored input coordinate. Those coordinates are not clipped afterward, and tabular features are not standardized before the shift, so the same δ is a different relative perturbation on images and on heterogeneous tables (Appendix A.1). Training is Adam with learning rate 5×10^{-4}, batch size 256, and early stopping.

3.2 PLAIN NETWORKS: BASELINE FISHER DISTRIBUTION AND EVOLUTION

We begin with **plain deep networks**—networks with neither skip connections nor normalization layers. Across encoders, plain chains often become more anisotropic with depth, and covariance energy accumulates in fewer directions.

Distribution shape. On California Housing (MLP L=8), the Plain snapshot has MSE 0.1778 ± 0.0016, EffR 4.24 ± 0.47, and $\log_{10}\kappa = 11.82 \pm 0.18$, indicating a sharply concentrated spectrum. On MLP/CIFAR-10 (L=8), Plain reaches $45.72\pm0.35\%$ validation accuracy with EffR 15.42 ± 0.66. LFI can rise when the perturbation aligns with small-variance directions (Eq. (2)); it is interpreted jointly with spectral metrics rather than in isolation.

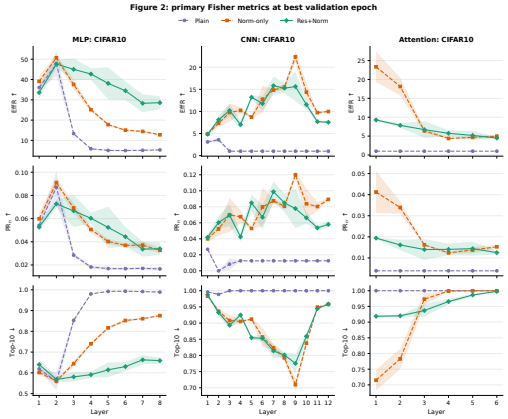

Figure 2: Layer-wise EffR (top), PR_n (middle), and Top-10 energy (bottom) on CIFAR-10. Norm-only is BN Pre; Res+Norm is Res-BN Pre. Plain (circles), Norm-only (squares), Res+Norm (diamonds): seed means $\pm$ one sample s.d. at the best validation epoch. MLP L=8, CNN L=12, Attention L=6.

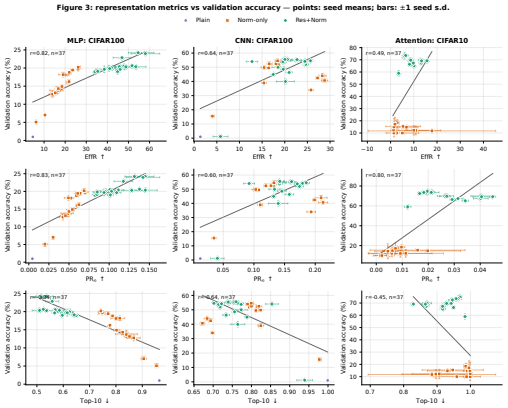

Figure 3: Primary triad versus validation accuracy. Columns: MLP, CNN, Attention. Rows: EffR, PR_n, Top-10. Bars are $\pm$ one seed s.d. Lines are descriptive fits and are not pooled across datasets.

Fisher evolution. Appendix B shows spectral concentration in many plain-network runs: EffR falls while κ and Top-10 energy rise. Its onset and magnitude vary by task and architecture; the pattern is an empirical baseline, not a universal monotonic law.

Layer-wise geometry. Paired 3D Fisher trajectories for the fixed Plain-network scenarios appear in Fig. A.1. Plain Attention/CIFAR-10 collapses consistently across seeds (EffR $= 1$, accuracy $= 10\%$), whereas MLP and CNN profiles are less uniform. This behavior motivates the residual and normalization interventions below.

3.3 Adding Residual Connections

We then ask how skip connections alter Fisher **distributions and their evolution**, holding the normalization family fixed whenever possible.

Spectral broadening. Across 144 paired MLP seed-trials, BN_Pre→Res_BN_Pre raises mean EffR by 9.36 ± 7.63, while LN_Pre→Res_LN_Pre raises it by 9.22 ± 7.99; average $\log_{10} \kappa$ decreases by 1.16 ± 2.72 and 1.23 ± 2.37, respectively. On MLP/CIFAR-10, Res_LN_Pre raises EffR from 25.73 ± 0.64 to 49.29 ± 0.39, while validation accuracy changes from $48.31 \pm 0.25\%$ to $47.33 \pm 1.03\%$. This probe thus exhibits residual spectral broadening, whereas the effect on task performance remains configuration-dependent.

Staged mitigation across depth. Figure 2 asks how the layer profile changes after normalization alone, and then after residuals on top of normalization. Figure 3 relates the same triad to validation accuracy in one scenario per encoder.

For the MLP diagnostic, Plain, Norm-only, and Res+Norm obtain $45.72 \pm 0.35\%$, $49.53 \pm 0.44\%$, and $47.45 \pm 0.31\%$ validation accuracy, respectively; the residual variant has the broader layer-wise spectrum. For CNN/CIFAR-10 (L=12), Plain, Norm-only, and Res+Norm obtain $10.00 \pm 0.00\%$, $85.84 \pm 0.36\%$, and $82.64 \pm 0.46\%$. For Attention/CIFAR-10, Plain remains at 10%, Norm-only reaches $52.44 \pm 1.82\%$, and Res+Norm reaches $73.62 \pm 0.40\%$. The Attention probe shows a strong residual effect inside this fixed BN-Pre comparison. Plain versus Norm-only also changes the coordinate system, because normalization rescales channels. Res+Norm versus Norm-only keeps BatchNorm Pre fixed. With skips, EffR is typically more stable than in plain chains (Appendix B); Fig. A.4 provides complementary LFI/κ/Align curves.

3.4 Adding Normalization

We lastly examine how **normalization layers** alter covariance spectra and LFI in nonlinear stacks, still excluding FisherNorm.

Per-channel standardization and spectral change. BN and LN standardize marginal channel or feature scales; they do not generally make the covariance matrix the identity. In the fixed MLP/CIFAR-10 snapshot, BN_Pre raises EffR from 15.42 ± 0.66 to 26.56 ± 0.98 and lowers $\log_{10} \kappa$ from 13.13 ± 0.02 to 3.91 ± 0.02. On California Housing, BN_Pre raises EffR from 4.24 ± 0.47 to 9.72 ± 1.67, while its $\log_{10} \kappa$ does not decrease in this particular snapshot. Standardization changes the spectrum in the coordinates it produces.

3.5 Association with Validation Performance

The associations in Fig. 3 are descriptive and protocol-specific: variants with broader spectra often attain higher validation accuracy, but the relationship is neither pooled across datasets nor causal. Plain, Norm-only, and Res+Norm occupy distinct regions in the selected probes, with the Attention case providing the clearest residual-associated recovery in accuracy. Fig. A.5 reports the complementary triad.

4 Local Account of Residual and Normalization Effects

We next interpret the staged diagnostic patterns (RQ2) by propagating input perturbations through layer Jacobians and tracking activation covariance $\Sigma^{(l)}$ together with directional sensitivity $\Delta\mu^{(l)}$. FisherNorm is not considered in this section.

4.1 Covariance and Sensitivity Propagation

Consider a differentiable encoder with layer-wise map $a^{(l)} = F^{(l)}(a^{(l-1)})$ and local Jacobian $J^{(l)} = \partial a^{(l)} / \partial a^{(l-1)}$. Linearizing about the evaluation distribution gives

$$\begin{aligned} \Delta\mu^{(l)} &\approx J^{(l)} \cdots J^{(1)} \, \Delta\mu^{(0)}, \\ \Sigma^{(l)} &\approx J^{(1:l)} \, \Sigma^{(0)} \, J^{(1:l)\top}, \end{aligned} \tag{3}$$

where $J^{(1:l)} := J^{(l)} J^{(l-1)} \cdots J^{(1)}$. The leading singular value does not fix EffR.

Let $s_1^{(l)} \geq \cdots \geq s_d^{(l)} \geq 0$ denote the eigenvalues of $\Sigma^{(l)}$. For plain layers $a^{(l)} = \phi(W^{(l)} a^{(l-1)})$ with $J^{(l)} = D_\phi^{(l)} W^{(l)}$ ($D_\phi^{(l)}$ the activation derivative), Eq. (17) (Appendix C) bounds $\|\Sigma^{(l)}\|_2$ by the product of squared layer operator norms, so covariance energy is compressed whenever that product is contractive. Smaller eigenvalues may vanish more rapidly than the leading mode, causing $\kappa^{(l)} = s_1^{(l)} / s_d^{(l)}$ to increase. Equation (17) is a *sufficient* condition for compression of the leading eigenvalue, not a statement that every trained plain layer has $\|J^{(j)}\|_2 < 1$, and not a determination of EffR, PR_n, or Top-10 energy. A contractive map can leave those spectral shapes unchanged. The broadening reported above is empirical.

4.2 Residuals Preserve Sensitivity Through an Identity Path

For a residual layer $a^{(l)} = \phi(W^{(l)} a^{(l-1)}) + a^{(l-1)}$,

$$J^{(l)} = D_\phi^{(l)} W^{(l)} + I. \tag{4}$$

Let $A^{(l)} = D_\phi^{(l)} W^{(l)}$. The singular values of $I + A^{(l)}$ cannot in general be obtained by adding one to the singular values of $A^{(l)}$: that simplification requires additional alignment conditions. The perturbation bounds in Eq. (20) (Appendix C) explain why a near-identity residual branch can locally preserve gain. Residual spectral broadening is an empirical phenomenon illuminated by this near-identity analysis, not a universal theorem.

At the covariance level, Eq. (19) (Appendix C) carries the identity term forward from the preceding representation, while cross terms may reinforce or cancel individual modes. The paired EffR increases reported above accord with this near-identity interpretation.

4.3 Per-Channel Standardization Reshapes Spectra

Batch normalization standardizes each channel across a mini-batch, whereas layer normalization standardizes features within each sample; neither operation generally decorrelates all channels. For the simplified coordinate-wise transformation $\hat{x}_i = (x_i - \mu_i)/\sqrt{\sigma_i^2 + \varepsilon}$, and omitting statistic-dependence terms, the local Jacobian is approximately

$$\frac{\partial \hat{x}}{\partial x} \approx \operatorname{diag}\big((\sigma_1^2 + \varepsilon)^{-1/2}, \ldots, (\sigma_d^2 + \varepsilon)^{-1/2}\big), \tag{5}$$

This controls marginal scales but does not imply $\mathrm{Cov}(\hat{x}) = I$. Its consequences for κ, EffR, and accuracy are consequently empirical and task-dependent.

A useful abstraction is **partial standardization**, parameterized by an exponent $\lambda_i \in [0, 1]$:

$$\hat{x}_i = \frac{x_i - \mu_i}{(\sigma_i^2 + \varepsilon)^{\lambda_i/2}} \quad \Rightarrow \quad \mathrm{Var}(\hat{x}_i) = \frac{\sigma_i^2}{(\sigma_i^2 + \varepsilon)^{\lambda_i}}. \tag{6}$$

At λ_i=1, this recovers the standard BN form $(x_i - \mu_i)/\sqrt{\sigma_i^2 + \varepsilon}$; at λ_i=0, it centers only. When $\varepsilon \ll \sigma_i^2$, the variance scales as $\sigma_i^{2(1-\lambda_i)}$. The exponent interpolates between centering and full per-channel variance standardization—the design space explored by FisherNorm.

4.4 Empirical Check: Layer Jacobian Spectral Norms

We inspect operator norms in Fashion-MNIST depth-6 probes under a factorial ablation (Plain / Residual-only / Norm-only), estimating $\|J^{(l)}\|_2$ by power iteration. Residual layers are expansive on this probe, whereas Plain MLP and Attention layers are predominantly contractive. That classification is all the norm supplies. It does not show that an expansive residual broadens EffR, nor that a contractive plain layer lowers EffR: a map with $\|J\|_2 < 1$ can leave EffR, PR_n, Top-10 energy, and κ unchanged. The EffR changes cited above are a separate paired measurement. The curves are in Appendix C.5.

4.5 Composed Blocks and Bounded Findings

In a Res-BN block, $J_{\mathrm{block}} = \big(D_\phi \operatorname{diag}(\gamma) \operatorname{diag}\big((\sigma_1^2+\varepsilon)^{-1/2}, \ldots, (\sigma_d^2+\varepsilon)^{-1/2}\big) W\big) + I$ combines per-channel scale standardization with identity-path preservation. Substituting Eq. (3) into Eq. (2) shows that LFI can rise when propagation amplifies the mean difference along retained low-variance modes. Both the identity path and per-channel standardization can change this quantity; the net effect stays empirical because covariance and sensitivity co-evolve during training.

The measurements support three bounded findings, not design principles or a training objective. **Finding 1 (concentration):** low EffR and high Top-10 energy describe a concentrated spectrum in the measured coordinates. $\|J\|_2<1$ is a separate contraction check and is neither necessary nor sufficient for that spectrum. **Finding 2 (diversity):** higher EffR or PR_n describes a broader covariance spectrum in those same coordinates. Residual variants often have broader spectra in the paired BN-Pre comparison, but the Jacobian norm does not imply that result. **Finding 3 (rescaling):** κ and the spectrum move together under per-channel rescaling (Eq. (5)). The exponent λ parameterizes the strength of that rescaling, and FisherNorm learns it from the task loss.

5 FisherNorm

FisherNorm trains one exponent per channel from the task loss, $\lambda_i = \sigma(\rho_i) \in (0, 1)$:

$$\hat{x}_i = \frac{x_i - \mu_i}{(\sigma_i^2 + \varepsilon)^{\lambda_i/2}}, \quad y_i = \gamma_i \hat{x}_i + \beta_i. \tag{7}$$

Table 1: Aggregate performance (mean±s.d. over seeds; placement by cross-seed validation mean).

Normalization	Mean clf. acc.	Reg. gain vs. Plain
FisherNorm	$62.02 \pm 0.11\%$	$+21.50 \pm 1.57\%$
PowerNorm	$61.99 \pm 0.27\%$	$+21.49 \pm 1.91\%$
BatchNorm	$61.94 \pm 0.11\%$	$+21.15 \pm 2.14\%$
EvoNorm	$61.37 \pm 0.26\%$	$+20.13 \pm 1.69\%$
LayerNorm	$60.79 \pm 0.10\%$	$+17.57 \pm 1.41\%$
GroupNorm	$60.19 \pm 0.15\%$	$+17.98 \pm 1.87\%$
RMSNorm	$60.08 \pm 0.06\%$	$+18.22 \pm 2.23\%$
ScaleNorm	$59.80 \pm 0.06\%$	$+18.30 \pm 1.92\%$
FilterResponseNorm	$59.58 \pm 0.08\%$	$+18.11 \pm 1.96\%$
Plain	$42.84 \pm 0.34\%$	—

Table 2: Seed-trial wins under Rule B (178 trials). Inside each seed, each family keeps its validation-best placement.

Normalization	Wins	Rate
FisherNorm	45	25.3%
PowerNorm	36	20.2%
BatchNorm	32	18.0%
EvoNorm	21	11.8%
FilterResponseNorm	13	7.3%
RMSNorm	13	7.3%
ScaleNorm	8	4.5%
GroupNorm	7	3.9%
Plain	2	1.1%
LayerNorm	1	0.6%

Training starts at $\rho_i = 3$. The task loss updates ρ_i together with γ_i and β_i. The forward pass uses the current batch statistics and this exponent; it does not form $F(\theta)$ or LFI (Appendix E.2). Each layer has $3d$ parameters, which is d more than BatchNorm. The BatchNorm entry in the tables is this same map with every λ_i held at 1. The PowerNorm entry is the published fixed per-tensor power (Shen et al., 2020). Both are trained at the four placements above and then summarized by Rule A or Rule B. The training forward pass is Algorithm 2 (Appendix E.2). As $\rho_i \to +\infty$, $\lambda_i \to 1$ and the map approaches BatchNorm. At finite ρ_i, $\lambda_i = \sigma(\rho_i) < 1$.

5.1 Illustrative Geometry Comparisons

Figs. A.3–A.6 compare FisherNorm with all nine families and Plain on three scenarios, complementing Fig. 2, using the checkpoints defined above.

5.2 Matched Comparison with BatchNorm (λ=1)

The paired numbers below use Eq. (7) twice: FisherNorm with $\lambda_i = \sigma(\rho_i)$ trained from $\rho_i = 3$, and BatchNorm with λ_i held at 1. Each side keeps the placement chosen by Rule A or Rule B.

Paired task outcomes. Across paired seed–scenario comparisons, FisherNorm wins/ties/losses versus BatchNorm are $93/8/77$ (seed 42: $29/4/27$; seed 123: $35/2/22$; seed 456: $29/2/28$).

5.3 Performance Comparison

Tables 1 and 2 apply Rule A and Rule B. Seeds $\{42, 123, 456\}$ give 178 trials: 59 scenarios contribute three seeds, and one scenario contributes seed 42 only.

Table 3: Largest absolute FisherNorm–BatchNorm gaps under Rule A (mean±s.d. over seeds); L denotes depth.

Scenario	L	**BN**	**FisherNorm**	Δ
Classification (Acc ↑)				
CNN/TinyImageNet	6	39.77 ± 0.17	$\mathbf{40.75 \pm 0.14}$	$+0.97$
Attention/CIFAR-10	12	73.80 ± 0.24	$\mathbf{74.77 \pm 1.51}$	$+0.97$
MLP/STL-10	4	36.70 ± 0.47	$\mathbf{37.27 \pm 0.66}$	$+0.57$
Attention/CIFAR-10	6	74.01 ± 1.16	$\mathbf{74.52 \pm 1.15}$	$+0.51$
MLP/CIFAR-10	8	50.62 ± 0.10	$\mathbf{51.10 \pm 0.44}$	$+0.47$
Regression (MSE ↓; Δ vs. BN)				
MLP/PowerPlant	12	0.0452 ± 0.0028	$\mathbf{0.0434 \pm 0.0009}$	-3.94%
MLP/PowerPlant	6	0.0415 ± 0.0004	$\mathbf{0.0403 \pm 0.0012}$	-2.90%
MLP/ETTh1	6	0.4394 ± 0.0001	$\mathbf{0.4267 \pm 0.0099}$	-2.89%
MLP/Parkinsons	8	0.1564 ± 0.0006	$\mathbf{0.1525 \pm 0.0020}$	-2.47%
MLP/PowerPlant	8	0.0433 ± 0.0009	$\mathbf{0.0426 \pm 0.0006}$	-1.69%

Table 3 lists, under Rule A, the five classification and five regression scenarios with the largest absolute FisherNorm–BatchNorm gaps.

6 Discussion

Plain-stack concentration is a statement about covariance energy. Depth often packs that energy into fewer directions, and plain Attention on CIFAR-10 reaches effective rank one. A skip on fixed BatchNorm Pre often spreads the energy again. With the normalizer held fixed, that broader spectrum is the contribution of the identity path. Accuracy still need not move with it. Attention pairs the broader spectrum with a higher validation score, while residual LayerNorm on MLP CIFAR-10 pairs a broader spectrum with a slightly lower score. Figure 3 keeps each scenario separate. The geometric claim concerns the measured coordinates and is sharpest when both sides use the same normalizer.

The Jacobian accounts for the identity path locally. The identity term changes how gain passes through the block, and the Fashion-MNIST spectral norm classifies layers by the leading singular value. Effective rank and conditioning remain properties of the full covariance, so the norm alone does not establish spectral broadening. Layer linear Fisher information is the squared length of a mean displacement in the local Gaussian precision, and a low-energy direction can dominate it. Per-channel standardization moves diversity and conditioning together in the coordinates it produces. FisherNorm trains the exponent of this map by the task loss within the same four-placement comparison. Task accuracy and the spectral diagnosis answer two different questions here.

7 Conclusion

On these Euclidean encoders, plain stacks often concentrate activation covariance, residual paths often broaden it relative to the same normalizer, and per-channel rescaling moves diversity and conditioning in the coordinates of the layer. The Jacobian norm records local contraction or expansion, and LFI measures a mean displacement in the activation precision metric. FisherNorm learns the exponent of that rescaling from the task loss and is compared with the other families under the same four placements, both by cross-seed validation means and by seed-wise wins. The primary result is this diagnosis together with the local account of residuals and normalization. Read this way, the spectrum, the Jacobian, and the task comparison answer different questions. A broader covariance spectrum describes the coordinates produced by the layer, and the residual comparison is informative because the normalizer is held fixed. Those measurements single out the standardization exponent as the parameter that moves diversity and conditioning together, and FisherNorm is that parameter trained by the task loss. The resulting accuracy is a summary of the shared placement protocol, kept separate from the spectral diagnosis.

AI USE STATEMENT

In this work, we used generative AI tools to clarify the conceptual framing of the existing activation-space Fisher quantities and to interpret the reported measurements in the discussion. We have not used generative AI tools to generate synthetic data, formulate new mathematical claims, assist with proofs, propose hypotheses, design or implement experiments, translate source material, clean datasets, or carry out qualitative data analysis, and survey or interview design is not applicable to this work. Additionally, we used generative AI tools to draft and edit passages for readability, to propose the title, and to adjust figure sizing and citation formatting in the manuscript source. We checked the revised wording against the existing experiments, equations, and tables, and did not introduce measurements that were not already computed. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

REFERENCES

Alessandro Achille, Giovanni Paolini, and Stefano Soatto. Where is the information in a deep neural network? *arXiv preprint arXiv:1905.12213*, 2019.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361, 2021.

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, pp. 342–350, 2017.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1059–1071, 2021.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 573–582, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, volume 9908 of *Lecture Notes in Computer Science*, pp. 630–645. Springer, 2016b. doi: 10.1007/978-3-319-46493-0_38.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 791–800, 2018.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1945–1953, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580, 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1032–1041, 2019.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4156–4167, 2019.

Hanxiao Liu, Andrew Brock, Karen Simonyan, and Quoc V. Le. Evolving normalization-activation layers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13539–13550, 2020.

Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. In *International Conference on Learning Representations*, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, 2019.

Yann Ollivier. Riemannian metrics for neural networks I: Feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 2015.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations*, 2014.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4785–4795, 2017.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pp. 606–610, 2007.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, volume 31, pp. 2488–2498, 2018.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Sheng Shen, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. PowerNorm: Rethinking batch normalization in transformers. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8741–8751, 2020.

Saurabh Singh and Shankar Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11237–11246, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019.

This appendix collects measurement details, extended tables, and the figures referenced from the main paper. Summaries use seeds $\{42, 123, 456\}$.

Main text	Here
Appendix A / FI protocol	Sec. A
Appendix B / diagnosis	Sec. B
Fig. A.1	Sec. B.2
Appendix C.5 / $\|J\|_2$	Sec. C.5
Appendix E.2 / FisherNorm forward	Sec. E.2
Figs. A.3–A.6	Secs. F & H
Family-best placement	Secs. B.1 & E.3
Boundary conditions	Sec. I

A. Measurement Protocol

We record the estimators used for every logged Fisher quantity so that the main-text numbers can be reproduced.

A.1 Perturbation and LFI

For evaluation inputs x^0, form $x^{\pm} = x^0 \pm \delta\mathbf{1}$ with δ=0.3 (all-ones finite difference). Let $a^{(l)}(x) \in \mathbb{R}^d$ be post-activation features at layer l. With $\mu_{\pm}^{(l)} = \mathbb{E}[a^{(l)}(x^{\pm})]$,

$$\Sigma^{(l)} = \tfrac{1}{2}\big(\mathrm{Cov}(a^{(l)}(x^+)) + \mathrm{Cov}(a^{(l)}(x^-))\big), \tag{8}$$

$$\Delta\mu^{(l)} = \frac{\mu_+^{(l)} - \mu_-^{(l)}}{2\delta}, \tag{9}$$

$$\mathrm{LFI}^{(l)} = \Delta\mu^{(l)\top}\Big(\Sigma^{(l)}\Big)^{\dagger}\Delta\mu^{(l)} = \sum_{i:\, s_i^{(l)} > \varepsilon} \frac{\big(\alpha_i^{(l)}\big)^2}{s_i^{(l)}}, \tag{10}$$

where $\Sigma^{(l)} = U^{(l)}\mathrm{diag}(s_i^{(l)})\,U^{(l)\top}$, $\alpha^{(l)} = U^{(l)\top}\Delta\mu^{(l)}$, and the pseudoinverse truncates singular values at ε=10^{-9}. The means and covariances are empirical moments over evaluation examples, not a population law for one input. Stored inputs are passed in three blocks corresponding to x^-, x^0, and x^+, and δ=0.3 is an absolute coordinate shift of those stored inputs. It is not clipped or re-scaled per feature, so the same numerical δ has a different effect on image tensors and on heterogeneous tabular coordinates. LFI is an activation-space directional probe under local linearization; it is not an entry of the parameter FIM.

A.2 Six Spectral Metrics

With $\Sigma^{(l)} = U^{(l)}\mathrm{diag}(s_i)\,U^{(l)\top}$, $\bar{s}_i = s_i / \sum_j s_j$, $\hat{\mu} = \Delta\mu/\|\Delta\mu\|_2$:

$$\mathrm{EffR}^{(l)} = \exp\left(-\sum_{i=1}^{d} \bar{s}_i^{(l)} \log\big(\bar{s}_i^{(l)} + \varepsilon\big)\right), \tag{11}$$

$$\mathrm{PR}^{(l)} = \frac{\big(\sum_i s_i^{(l)}\big)^2}{\sum_i \big(s_i^{(l)}\big)^2}, \quad \mathrm{PR}_n^{(l)} = \frac{\mathrm{PR}^{(l)}}{d}, \tag{12}$$

$$\kappa^{(l)} = \frac{s_1^{(l)}}{s_d^{(l)} + \varepsilon}, \quad E_{10}^{(l)} = \sum_{i=1}^{\min(10,d)} \bar{s}_i^{(l)}, \tag{13}$$

$$\mathrm{Align}^{(l)} = \max_i \big|(U_i^{(l)})^{\top}\hat{\mu}^{(l)}\big|. \tag{14}$$

Unless noted otherwise, metrics are averaged over logged hidden layers at the best validation epoch. Figs. 2–3 use the primary triad $\{\mathrm{EffR}, \mathrm{PR}_n, E_{10}\}$; Figs. A.4–A.6 use $\{\mathrm{LFI}, \log_{10}\kappa, \mathrm{Align}\}$. We

refer to low EffR, high E_{10}, and/or extreme κ as spectral concentration. This is a property of Σ, not of the parameter FIM.

A.3 Logging Algorithm

We log layer-wise Fisher probes at evaluation time as follows.

Algorithm 1 Fisher information logging

Require: model f_θ, eval set $\mathcal{D}$, layers $\mathcal{L}$, δ=0.3
1: eval mode; optionally load best-val checkpoint for summary metrics
2: **for** minibatch in $\mathcal{D}$ **do**
3: form triplets; forward; cache $a^{(l)}$ for $l \in \mathcal{L}$
4: **end for**
5: **for** $l \in \mathcal{L}$ **do**
6: estimate $\Delta\mu, \Sigma$; SVD; truncate; log LFI, EffR, PR$_n$, κ, E_{10}, Align
7: **end for**

Reported snapshots use the best-validation checkpoint unless a figure caption states otherwise.

B. Diagnosis: Extended Evidence

This section expands the staged diagnosis in the main paper (Plain, residual, and fixed-strength norms; no FisherNorm).

B.1 Experimental Grid and Family-Best Protocol

13 datasets: CIFAR-10/100, Fashion-MNIST, TinyImageNet, STL-10, Covtype, Higgs, California Housing, ETTh1, Electricity, Parkinsons, Power Plant, Superconductivity.

Encoders: MLP (L, D): $(4, 384)$, $(6, 256)$, $(8, 224)$, $(12, 192)$; CNN (L, C): $(6, 112)$, $(12, 80)$; Attention (L, d): $(6, 256)$, $(12, 192)$.

Training: Adam (lr=5×10^{-4}, weight decay 5×10^{-4}, batch size 256); ReduceLROnPlateau with early-stopping patience 20.

Grid: 9 norm families $\times$ {Pre, Post, Res-Pre, Res-Post} + Plain $= 37$ variants/scenario; 60 scenarios; 2,220+ Fisher-traced runs.

Family-best: for aggregate means, each family keeps the placement with the best cross-seed mean validation score; for seed-trial wins, each family instead picks its best placement within that seed.

B.2 Plain Trajectories

Figure A.1 shows how plain stacks can concentrate covariance energy with depth. The severity depends on architecture and task; this is the baseline that motivates the interventions in Fig. 2.

B.3 Diagnosis Snapshots (No FisherNorm)

The tables below support the staged diagnosis in the main text: Plain concentrates; residuals raise EffR; Norm+Res recovers usable accuracy on the hard Attention case.

B.4 Paired Residual Effects

Across these pairs, residuals raise mean EffR by about $+9$ while $\log_{10}\kappa$ decreases on average. LFI changes have larger dispersion than the EffR changes.

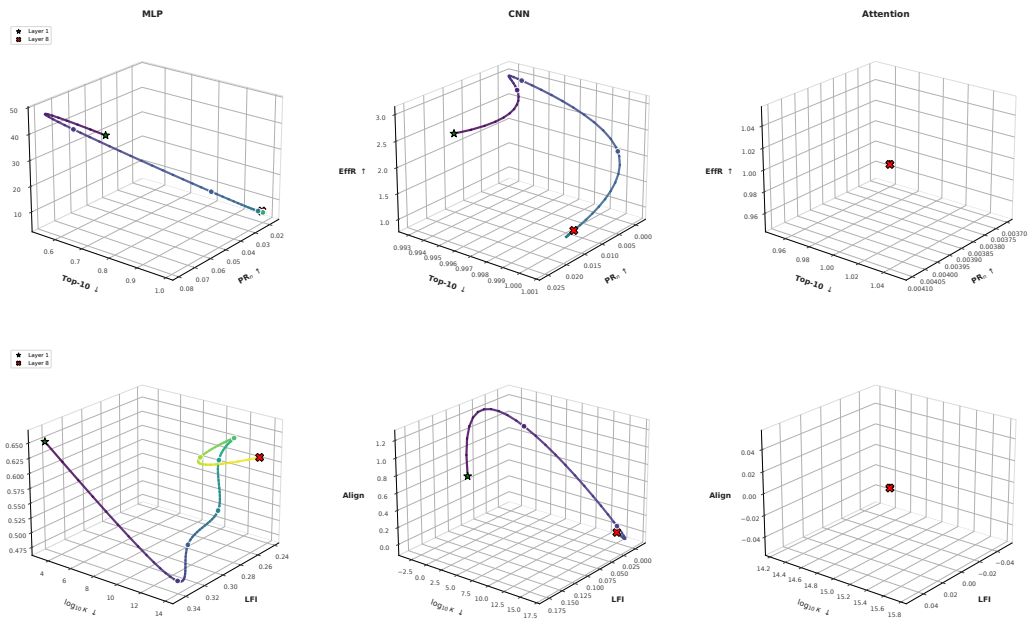

Figure A.1: Plain-network Fisher pathologies in dual 3D views. Each panel is one layer-wise trajectory (green star = layer 1; red cross = final layer; line color encodes depth). Top: $(\mathrm{PR}_n, E_{10}, \mathrm{EffR})$; bottom: $(\mathrm{LFI}, \log_{10}\kappa, \mathrm{Align})$. Columns: MLP, CNN, Attention.

Table A.1: Attention / CIFAR-10 (L=6): Plain collapse vs. staged repair (mean ± sample s.d.).

Variant	**Acc (%)**
Plain	10.00 ± 0.00
BN_Pre	52.44 ± 1.82
Res_BN_Pre	73.62 ± 0.40

C. Mechanism: Derivations and Jacobian Probe

The identities below are the linearized arguments used in the main text; the empirical $\|J\|_2$ probe follows.

C.1–C.4 Identities

Propagation

$$\Delta\mu^{(l)} \approx J^{(1:l)}\Delta\mu^{(0)}, \tag{15}$$

$$\Sigma^{(l)} \approx J^{(1:l)}\Sigma^{(0)}J^{(1:l)\top}, \tag{16}$$

where $J^{(1:l)} := J^{(l)}J^{(l-1)}\cdots J^{(1)}$.

Plain bound

$$\|\Sigma^{(l)}\|_2 \le \Big(\prod_j \|J^{(j)}\|_2^2\Big)\,\|\Sigma^{(0)}\|_2 \tag{17}$$

This is a sufficient condition for covariance-energy compression under contraction; smaller eigenvalues may vanish faster and worsen κ.

Table A.2: California Housing (MLP L=8): residuals broaden EffR relative to Plain/BN (mean ± sample s.d.).

Variant	MSE	EffR	$\log_{10}\kappa$
Plain	0.1778 ± 0.0016	4.24 ± 0.47	11.82 ± 0.18
BN_Pre	0.1770 ± 0.0049	9.72 ± 1.67	12.26 ± 0.08
Res_BN_Pre	0.1724 ± 0.0019	26.96 ± 0.79	11.32 ± 0.11
Res_LN_Pre	0.1820 ± 0.0023	19.59 ± 2.34	11.69 ± 0.50

Table A.3: Fashion-MNIST (MLP L=8): residuals raise EffR under matched norms (mean ± sample s.d.).

Variant	Acc	EffR	$\log_{10}\kappa$
Plain	85.64 ± 0.11	8.64 ± 0.46	13.72 ± 0.10
BN_Pre	85.67 ± 0.06	17.60 ± 0.46	10.16 ± 3.60
LN_Pre	85.32 ± 0.17	16.38 ± 0.38	6.35 ± 0.58
Res_BN_Pre	85.56 ± 0.20	20.40 ± 0.53	10.26 ± 3.45
Res_LN_Pre	85.53 ± 0.30	20.41 ± 0.16	5.27 ± 0.55

RESIDUAL

$$J = D_\phi W + I, \tag{18}$$

$$\Sigma_{\text{res}} \approx (D_\phi W + I)\Sigma(D_\phi W + I)^\top \tag{19}$$

The identity path contributes an additive mode, with possible cross-term reinforcement or cancellation. For $A = D_\phi W$,

$$\max\{0, 1 - \|A\|_2\} \le \sigma_{\min}(I + A) \le \sigma_{\max}(I + A) \le 1 + \|A\|_2. \tag{20}$$

These local bounds do not imply non-contraction at every depth.

PER-CHANNEL STANDARDIZATION / PARTIAL STANDARDIZATION

$$\frac{\partial \hat{x}}{\partial x} \approx \operatorname{diag}\Big((\sigma_i^2 + \varepsilon)^{-\lambda_i/2}\Big)_{i=1}^{d}, \tag{21}$$

$$\hat{x}_i = \frac{x_i - \mu_i}{(\sigma_i^2 + \varepsilon)^{\lambda_i/2}}, \tag{22}$$

$$\operatorname{Var}(\hat{x}_i) = \frac{\sigma_i^2}{(\sigma_i^2 + \varepsilon)^{\lambda_i}}. \tag{23}$$

The diagonal approximation controls marginal scales; it does not yield an identity covariance.

C.5 EMPIRICAL $\|J^{(l)}\|_2$

Protocol: Fashion-MNIST, L=6; factorial {Plain, Residual-only, Norm-only}×{MLP, CNN, Attention}; power iteration (30 iters, 32 probes) at the best-val checkpoint.

Residual layers are expansive in this probe. Plain MLP and Attention layers are mostly contractive; CNN Plain declines with depth (final 0.79).

D. MEASURED REGIMES AND PARAMETERIZATION BOUNDARY

For reference, the main text distinguishes three measured regimes: concentration (low EffR / high E_{10}), diversity (EffR/PR$_n$), and conditioning (κ). These quantities describe covariance in the coordinates produced by each layer; they are not invariant health criteria. BN and LN directly rescale

Table A.4: Paired residual effects on MLP seed-trials (norm family fixed; n=144; mean ± sample s.d.).

Pair	ΔEffR	$\Delta \log_{10} \kappa$	ΔLFI
BN_Pre→Res_BN_Pre	+9.36±7.63	−1.16±2.72	−5.05±58.46
LN_Pre→Res_LN_Pre	+9.22±7.99	−1.23±2.37	−8.09±56.57

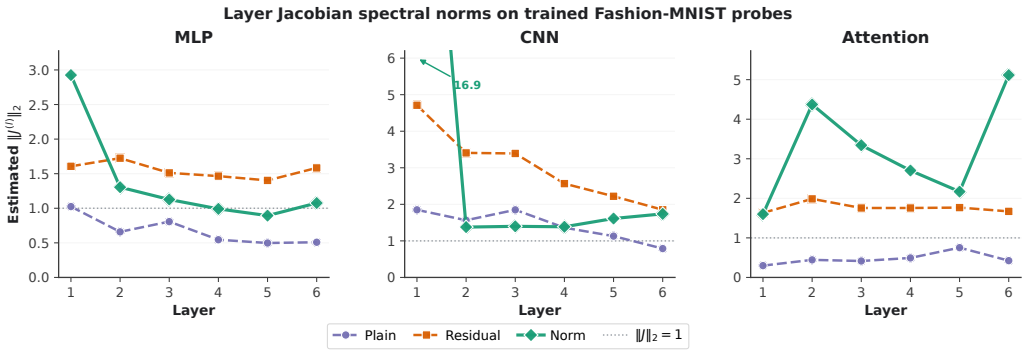

Figure A.2: Layer Jacobian spectral norms on Fashion-MNIST (L=6). Dashed line: $\|J\|_2$=1.

those coordinates, while a residual path changes the block Jacobian. FisherNorm parameterizes the strength of the former rescaling with a task-trained per-channel exponent. No result in this paper shows that this learned exponent improves the three statistics.

E. FisherNorm: Implementation

E.1 Definition and BN Limit

$$\hat{x}_i = \frac{x_i - \mu_i}{(\sigma_i^2 + \varepsilon)^{\lambda_i/2}}, \tag{24}$$

$$\lambda_i = \sigma(\rho_i), \tag{25}$$

$$y_i = \gamma_i \hat{x}_i + \beta_i. \tag{26}$$

As $\rho_i \to +\infty$, $\lambda_i \to 1$ and the map approaches BatchNorm: $(x_i - \mu_i)/\sqrt{\sigma_i^2 + \varepsilon}$ with the same $\mu, \sigma^2, \varepsilon, \gamma, \beta$. For finite ρ_i, $\lambda_i < 1$, so exact BatchNorm requires fixing λ_i=1 from outside. We initialize ρ_i=3.

E.2 Forward (Training)

Algorithm 2 FisherNorm forward (training)

Require: batch x; γ, ρ, β; EMA momentum
1: batch μ, σ^2; update running stats
2: $\lambda \leftarrow \sigma(\rho)$
3: $\hat{x} \leftarrow (x - \mu) \odot (\sigma^2 + \varepsilon)^{-\lambda/2}$
4: **return** $\gamma \odot \hat{x} + \beta$

No runtime Fisher estimation is required; task loss trains ρ. Relative to BN the overhead is $+d$ parameters and $O(d)$ element-wise ops. For CNN activations we use per-channel spatial statistics.

Table A.5: Estimated $\|J^{(l)}\|_2$ summary (Fashion-MNIST, L=6).

Arch	**Kind**	**Mean**	**Frac.** <1
MLP	Plain	0.67	$5/6$
MLP	Residual	1.55	$0/6$
MLP	Norm	1.39	$2/6$
CNN	Plain	1.42	$1/6$
CNN	Residual	3.02	$0/6$
CNN	Norm	4.08	$0/6$
Attention	Plain	0.47	$6/6$
Attention	Residual	1.76	$0/6$
Attention	Norm	3.22	$0/6$

Table A.6: FN–BN classification comparisons under family-best placement (mean±s.d.; same numbers as the main text).

Scenario	**BN**	**FN**	Δ
TinyImageNet/L=6	39.77±0.17	$\mathbf{40.75\pm0.14}$	+0.97
CIFAR-10/L=12	73.80±0.24	$\mathbf{74.77\pm1.51}$	+0.97
STL-10/L=4	36.70±0.47	$\mathbf{37.27\pm0.66}$	+0.57
CIFAR-10/L=6	74.01±1.16	$\mathbf{74.52\pm1.15}$	+0.51
CIFAR-10/L=8	50.62±0.10	$\mathbf{51.10\pm0.44}$	+0.47

E.3 Placements and Baselines

$$\text{Pre:} \quad h' = \phi(\mathrm{Norm}(Wh)), \tag{27}$$

$$\text{Post:} \quad h' = \mathrm{Norm}(\phi(Wh)), \tag{28}$$

$$\text{Res-Pre:} \quad h' = \phi(\mathrm{Norm}(Wh)) + h, \tag{29}$$

$$\text{Res-Post:} \quad h' = \mathrm{Norm}(\phi(Wh) + h). \tag{30}$$

Families: BN, LN, GN, RMS, Evo, Power, FRN, Scale, FisherNorm.

Externally fixing λ_i=1 gives BatchNorm; finite learned ρ_i only approaches that endpoint. Paired three-seed comparisons against family-best BN are $93/8/77$ (win/tie/loss).

F. Illustrative Geometry Cases

We compare all nine families plus Plain on three family-complete scenarios, complementary to Fig. 2. Curves are seed means at the best validation epoch; shaded bands (±1 seed s.d.) are shown only for FisherNorm and BatchNorm.

Scenarios (Figs. A.3 and A.6): MLP / Fashion-MNIST / L=8, CNN / TinyImageNet / L=6, and Attention / CIFAR-10 / L=6. Family-best curves on $\{\mathrm{EffR}, \mathrm{PR}_n, E_{10}\}$. Placement is chosen by the cross-seed mean validation score for each family.

On MLP/CIFAR-10 (L=8), FisherNorm Res-Post EffR is 53.48 ± 1.48 vs. BN Res-Post 52.91 ± 5.42. On California Housing, FisherNorm Res-Pre EffR is 28.82 ± 1.14 with MSE 0.1696 ± 0.0019.

G. Performance Extensions

The tables below restate the main-paper aggregates and extend the regression side of the FisherNorm–BatchNorm comparison. Numbers match the main paper.

Table A.7: Aggregate performance (mean±s.d. over seeds; placement by cross-seed validation mean).

Normalization	Mean clf. acc.	Reg. gain vs. Plain
FisherNorm	$62.02 \pm 0.11\%$	$+21.50 \pm 1.57\%$
PowerNorm	$61.99 \pm 0.27\%$	$+21.49 \pm 1.91\%$
BatchNorm	$61.94 \pm 0.11\%$	$+21.15 \pm 2.14\%$
EvoNorm	$61.37 \pm 0.26\%$	$+20.13 \pm 1.69\%$
LayerNorm	$60.79 \pm 0.10\%$	$+17.57 \pm 1.41\%$
GroupNorm	$60.19 \pm 0.15\%$	$+17.98 \pm 1.87\%$
RMSNorm	$60.08 \pm 0.06\%$	$+18.22 \pm 2.23\%$
ScaleNorm	$59.80 \pm 0.06\%$	$+18.30 \pm 1.92\%$
FilterResponseNorm	$59.58 \pm 0.08\%$	$+18.11 \pm 1.96\%$
Plain	$42.84 \pm 0.34\%$	—

Table A.8: Seed-trial wins under Rule B (178 trials). Inside each seed, each family keeps its validation-best placement.

Normalization	Wins	Rate
FisherNorm	45	25.3%
PowerNorm	36	20.2%
BatchNorm	32	18.0%
EvoNorm	21	11.8%
FilterResponseNorm	13	7.3%
RMSNorm	13	7.3%
ScaleNorm	8	4.5%
GroupNorm	7	3.9%
Plain	2	1.1%
LayerNorm	1	0.6%

Paired FisherNorm vs. family-best BatchNorm: $93/8/77$ (win/tie/loss). Each (scenario, seed) is one trial; placements are chosen within that seed.

H. Secondary Triad

Figures A.4–A.6 reuse the layouts of Fig. 2, Fig. 3, and Fig. A.3 with $\{\mathrm{LFI}, \log_{10}\kappa, \mathrm{Align}\}$.

H.1 Layer-wise Secondary Metrics

Norm+Res typically repair extreme κ. When spectra overconcentrate, LFI and Align can diverge from EffR.

H.2 Secondary Metrics vs. Accuracy

Accuracy associates more consistently with EffR/PR$_n$/Top-10 than with the secondary triad.

H.3 All-Family Secondary Triad

FisherNorm keeps conditioning moderate without large Align spikes. High LFI with low EffR in some fixed norms is consistent with over-concentration.

Table A.9: Largest FisherNorm (FN) vs. BatchNorm regression margins on MLP (MSE ↓; mean ± sample s.d.; same protocol as the main-text highlight table).

Scenario	**BN**	**FN**	**Δ**
PowerPlant/L=12	0.0452±0.0028	**0.0434±0.0009**	−3.94%
PowerPlant/L=6	0.0415±0.0004	**0.0403±0.0012**	−2.90%
ETTh1/L=6	0.4394±0.0001	**0.4267±0.0099**	−2.89%
Parkinsons/L=8	0.1564±0.0006	**0.1525±0.0020**	−2.47%
PowerPlant/L=8	0.0433±0.0009	**0.0426±0.0006**	−1.69%

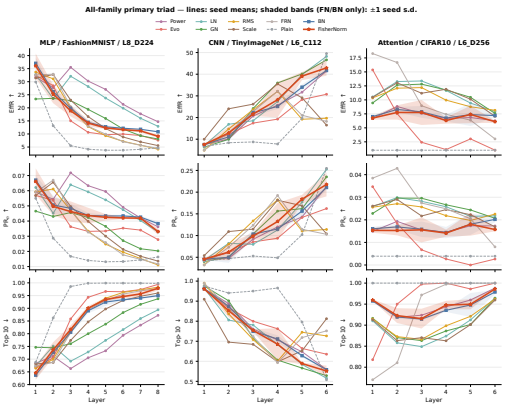

Figure A.3: All-family layer-wise primary triad. Seed means; ±1 seed s.d. shading only for FisherNorm and BatchNorm.

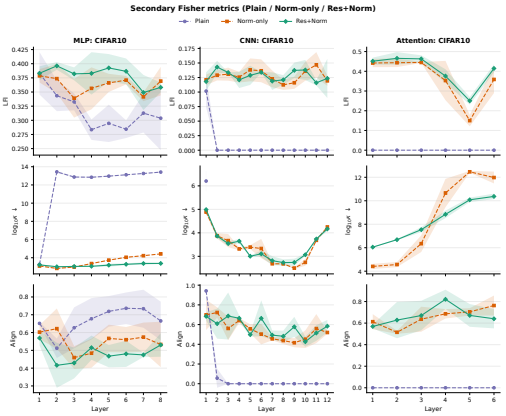

Figure A.4: LFI / $\log_{10}\kappa$ / Align for Plain, Norm-only, and Res+Norm.

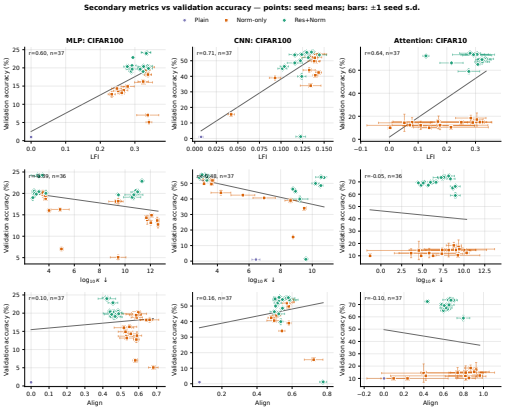

Figure A.5: Secondary metrics vs. accuracy within architecture. Each point is a family mean under Plain / Norm-only / Res+Norm; bars show ± one seed sample s.d.

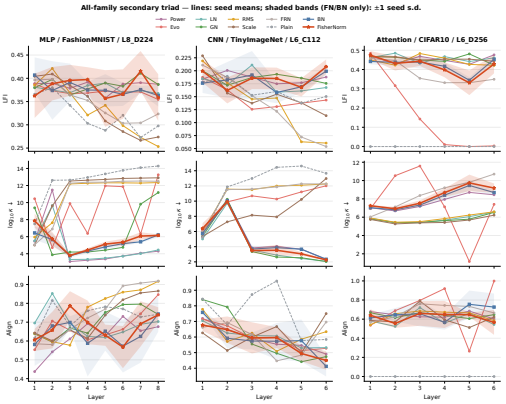

Figure A.6: All-family LFI / $\log_{10}\kappa$ / Align (same scenarios and shading as Fig. A.3).

I. Boundary Conditions

- On Covtype and ETTh1, BatchNorm and PowerNorm stay with FisherNorm.
- On some easy vision settings, the means of FisherNorm, BatchNorm, and EvoNorm remain close.

- On Superconductivity with a deep MLP, Plain wins where the tested normalization methods are not beneficial.
- Spectrum–accuracy associations in the main figures are descriptive.
- Rule-B seed-trial wins are FisherNorm 45, PowerNorm 36, and BatchNorm 32 of 178.

Reported means and win counts use seeds $\{42, 123, 456\}$.