THE GEOMETRY BEHIND THE BARRIER: FISHER GEOMETRY AND LINEAR MODE CONNECTIVITY IN WIDE NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

After permutation alignment, independently trained minima of wide neural networks often exhibit Linear Mode Connectivity (LMC), yet the geometric mechanism governing the interpolation loss barrier remains unclear. We show that Fisher information geometry explains the barrier through two decoupled axes: Shape, governing manifold curvature and geodesic alignment, and Length, governing the Fisher energy along the parameter displacement. Along the Shape axis, we show that the local variation of the Fisher metric vanishes in operator norm with width, driving Christoffel symbols and sectional curvature to zero with dimension-free bounds, simultaneously flattening the space and straightening Fisher geodesics into linear interpolations. However, this geometric flattening controls only the path's shape and is insufficient for barrier collapse. Along the Length axis, we establish a Fisher-barrier inequality showing that the barrier is quantitatively governed by the Fisher quadratic form and predictive uncertainty. Experiments across architectures and training regimes confirm that while shape flattening is universal, barrier collapse requires Fisher length reduction in the presence of controlled predictive uncertainty.

1 INTRODUCTION

Linear mode connectivity (LMC) reveals a striking property of over-parameterized neural networks: after resolving permutation symmetries, independently trained solutions can often be connected by a nearly lossless straight line in parameter space (Draxler et al., 2018; Garipov et al., 2018; Frankle et al., 2020; Vlaar & Frankle, 2022). This phenomenon has become increasingly relevant to model merging, weight averaging, and the study of connected low-loss regions (Entezari et al., 2022; Ainsworth et al., 2023; Jordan et al., 2023; Ferbach et al., 2024; Ito et al., 2025a;b; Sharma et al., 2024a; Zhao et al., 2025; Bürgin et al., 2026; Matena & Raffel, 2022; Tam et al., 2024; Wang et al., 2026; Qu & Horváth, 2025).

Yet the interpolation barrier does not behave uniformly as networks become wider. Depending on the training regime and model configuration, it may collapse, persist, or even grow (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Arora et al., 2019; Liu et al., 2020; Yang & Hu, 2021; Yang et al., 2022). These contrasting behaviors show that increasing width alone is not enough to explain linear mode connectivity. A geometric account must therefore distinguish the regimes in which the barrier collapses from those in which it persists or grows.

A natural intuition is that wider networks become geometrically flatter. However, Euclidean notions of flatness depend directly on parameter coordinates and need not reflect functional similarity between models. We therefore study the problem through Fisher information geometry (Chentsov, 1982; Amari, 1998; Amari & Nagaoka, 2000; Martens, 2020), which measures parameter changes according to their effect on the model's predictive distribution.

Our key observation is that two different geometric effects must be separated. We call them Shape and Length. Shape describes how the Fisher geometry bends and whether the connecting path becomes straighter as width increases. Length measures the full Fisher energy of the displacement between the two trained solutions.

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.

Along the Shape axis, we show that the Fisher geometry becomes locally flatter with width. However, this flattening occurs across all regimes we study, including regimes in which the interpolation barrier remains large. Geometric flattening alone therefore cannot explain barrier collapse: the connecting path may become nearly straight while still passing through a high-loss region.

Along the Length axis, we ask whether the displacement between the two trained solutions becomes smaller when both its direction and its magnitude are considered. We find that alignment with low-curvature directions alone does not explain the barrier. The displacement may point increasingly along low-curvature directions while still spanning a large distance in parameter space. Barrier collapse occurs when both effects combine to reduce the Fisher energy of the displacement, particularly when predictive uncertainty is controlled.

In summary, our main contributions are:

- **Shape–Length decomposition.** We separate Fisher geometry into Shape, which describes local curvature and path geometry, and Length, which measures the Fisher energy of the displacement between trained solutions.
- **Flattening is not enough.** We prove that the Fisher geometry becomes locally flatter with width, yet show that this flattening can occur even when the interpolation barrier persists.
- **Fisher length controls the barrier bound.** We derive a Fisher-barrier inequality linking barrier height to the full Fisher quadratic form and predictive uncertainty, and show empirically that Fisher length closely tracks barrier behavior across architectures and training regimes.

2 Related Work

Linear Mode Connectivity and Permutation Symmetries. The observation that independently trained networks can be connected by low-loss paths has motivated extensive study of loss surface topology (Draxler et al., 2018; Garipov et al., 2018; Frankle et al., 2020; Vlaar & Frankle, 2022). A key development was identifying discrete permutation symmetries as the primary obstruction to linear connectivity (Entezari et al., 2022), leading to algorithmic remedies including weight-matching heuristics (Ainsworth et al., 2023; Jordan et al., 2023; Ito et al., 2025a;b), optimal transport formulations (Ferbach et al., 2024), and structural identifiability constraints (Bürgin et al., 2026; Zhao et al., 2025; Sharma et al., 2024b). These methods resolve the combinatorial alignment problem but do not characterize the continuous geometric mechanisms governing the residual barrier height.

Loss Landscape Geometry. A common approach quantifies landscape flatness through the Euclidean Hessian spectrum (Dinh et al., 2017; Sagun et al., 2018; Papyan, 2020; Simsek et al., 2021; Singh et al., 2024). While informative, this perspective measures curvature in a coordinate system that is not invariant to reparameterization, and therefore does not directly capture functional equivalence between models. The resulting flatness assessments may not align with the actual barrier behavior observed across training regimes.

Fisher Information Geometry. The Fisher-Rao metric provides a reparameterization-invariant measure of statistical divergence (Chentsov, 1982; Amari, 1998; Amari & Nagaoka, 2000). The Fisher Information Matrix has been used to precondition optimization via the Natural Gradient (Martens, 2020; Karakida et al., 2019; Kunstner et al., 2019) and as a weighting scheme for model merging (Matena & Raffel, 2022; Wang et al., 2026). These applications treat the Fisher matrix as a local distance measure or static preconditioner, without exploiting the global Riemannian structure it induces or connecting it quantitatively to the interpolation barrier.

Lazy Training and Feature Learning. The distinction between lazy and feature-learning regimes is well studied in the context of training dynamics and generalization (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Arora et al., 2019; Liu et al., 2020; Yang & Hu, 2021; Yang et al., 2022). The implications of this distinction for mode connectivity, however, remain largely phenomenological: each combination of parameterization regime, architecture, and activation exhibits a distinct barrier trajectory, and no existing framework provides a geometric account of these differences. Our work contributes to this direction by decomposing Fisher geometry into Shape and Length axes, and establishing a quantitative link between Fisher length, predictive uncertainty, and the interpolation barrier within this specific setting.

3 BACKGROUND

A model $f(x; w) \in \mathbb{R}^K$ with parameters $w \in \mathbb{R}^P$ is trained by minimizing the cross-entropy loss $L(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log p_w(x)_y]$, where $p_w(x) = \mathrm{softmax}(f(x; w))$; a dot marks the argument left free, so $f(x; \cdot)$ is the map $w \mapsto f(x; w)$. Two independently trained minima w_A, w_B define the linear path $\gamma_{\mathrm{lin}}(t) = (1-t)w_A + tw_B$ and its barrier:

$$B(\gamma_{\mathrm{lin}}) = \max_{t\in[0,1]} L(\gamma_{\mathrm{lin}}(t)) - \frac{1}{2}\big(L(w_A) + L(w_B)\big). \tag{1}$$

Appendix A tabulates every symbol used below.

3.1 FISHER INFORMATION METRIC

Definition 3.1 (Fisher information matrix)**.** With the logit Jacobian $J(x; w) = \partial f(x; w)/\partial w \in \mathbb{R}^{K\times P}$ and $S(p) = \mathrm{diag}(p) - pp^\top \in \mathbb{R}^{K\times K}$, the Fisher matrix is

$$F(w) = \mathbb{E}_{x\sim\mathcal{D}_X}\left[J(x; w)^\top S(p_w(x))J(x; w)\right] \succeq 0, \tag{2}$$

averaged over the input marginal $\mathcal{D}_X$ of $\mathcal{D}$.

This is the Fisher matrix in the usual sense: the score covariance under a class $k \sim p_w(x)$ drawn from the model, so no label enters it. Here $S(p)$ is the softmax family's Fisher information in the logits, pulled back by J.

$F(w)$ is rank-deficient in over-parameterized networks, so we use the damped metric $G_F(w) = F(w) + \lambda I_P \succ 0$, which makes $(\mathbb{R}^P, G_F)$ Riemannian. We set $\lambda = 10^{-2}\lambda_{\max}(F)$, a fixed fraction of the metric's own scale, so that G_F is comparably conditioned across regimes whose Fisher scales differ by orders of magnitude (Appendix D). Only the geodesic–linear deviation depends on λ: $\partial G_F = \partial F$ leaves the Shape axis untouched, and the Fisher length, the Rayleigh quotient, the barrier and ρ^* are built from F alone.

3.2 THE FISHER-HESSIAN IDENTITY

Definition 3.2 (Predictive uncertainty)**.** The predictive uncertainty of a model w is

$$\rho^*(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\|p_w(x) - e_y\|_2\big] \in [0, \sqrt{2}], \tag{3}$$

with $e_y \in \mathbb{R}^K$ the one-hot label.

Definition 3.3 (Uniform derivative bounds)**.** Assume $f(x; \cdot) \in C^2$ for every input x on an open set $U \subset \mathbb{R}^P$, and let M_f bound its second derivative on U (the subscript names the function bounded, not the order of the derivative, as for M_F and M_σ),

$$M_f := \sqrt{K} \sup_{w\in U,\, x,\, k} \big\|\nabla_w^2 f_k(x; w)\big\|_{\mathrm{op}}. \tag{4}$$

If it is in addition C^3 on U, let M_L bound the third derivative of the *loss* on U,

$$M_L := \sup_{w\in U} \sup_{\|u\|=\|v\|=\|z\|=1} \big|\mathrm{D}^3 L(w)[u, v, z]\big|. \tag{5}$$

Theorem 3.4 (Fisher-Hessian Identity)**.** *Under dominated convergence conditions,* $\nabla^2 L(w) = F(w) + E(w)$*, where*

$$E(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\sum_{k=1}^{K}\big(p_w(x)_k - (e_y)_k\big)\nabla_w^2 f_k(x; w)\Big],$$

and, for all $v \in \mathbb{R}^P$*,*

$$\big|v^\top E(w)v\big| \le \rho^*(w)\, M_f\, \|v\|_2^2.$$

(Proof in Appendix B.1.)

So ρ^* controls how far $\nabla^2 L$ departs from F, and along γ_{lin} that residual is how predictive uncertainty enters the barrier.

3.3 Geodesics and Curvature of the Fisher Manifold

The geodesic γ_g of G_F satisfies, in terms of the Christoffel symbols Γ^k_{ij} of G_F and with repeated indices summed over $1, \dots, P$,

$$\ddot{\gamma}^k_g + \Gamma^k_{ij}(\gamma_g)\,\dot{\gamma}^i_g\,\dot{\gamma}^j_g = 0, \qquad \Gamma^k_{ij} = \frac{1}{2}(G_F^{-1})^{kl}\big[\partial_i(G_F)_{jl} + \partial_j(G_F)_{il} - \partial_l(G_F)_{ij}\big].$$

Departure from Euclidean flatness is measured by the sectional curvature $\mathrm{Sec}_{(\mathbb{R}^P, G_F)}(u, v)$ (do Carmo, 1992). Section 4 shows that once ∂F decays with width, both Γ and Sec inherit that decay with constants independent of P, while the barrier is governed by the Length axis instead. That decay is Assumption 4.2, which we posit rather than derive and verify only for a frozen-softmax surrogate under NTK.

4 Geometric Analysis

We decompose Fisher geometry into two axes, separated by what each is built from: Shape, from the derivatives of the metric F, and Length, from F itself contracted with the displacement. Each is one measurable scalar. Section 4.1 derives the barrier bound in which both appear; Sections 4.2–4.3 show that Shape governs the geometry of the interpolation path, and Section 4.4 that Length governs the barrier.

Throughout this section, let $\Delta = w_B - w_A$ and $\gamma_{\mathrm{lin}}(t) = w_A + t\Delta$, $t \in [0, 1]$, let n denote the hidden width of the network, and write

$$\mathcal{L}_F(w; \Delta) := \Delta^\top F(w)\Delta$$

for the Fisher length of the displacement measured at w.

4.1 The Barrier Is an Energetic Quantity

We now derive the main upper bound, which identifies the Fisher-geometric quantities governing the interpolation barrier.

Theorem 4.1 (Fisher-barrier inequality)**.** *Assume $f(x; \cdot) \in C^3$ for every input x on an open convex neighbourhood $U \supset [w_A, w_B]$, so that the bounds M_f and M_L of Definition 3.3 are finite there. Define the* anchor bound *at an endpoint w_0,*

$$R(w_0) := \|\Delta\|_2 \|\nabla L(w_0)\|_2 + \frac{1}{2}\mathcal{L}_F(w_0; \Delta) + \frac{1}{2}\rho^*(w_0) M_f \|\Delta\|_2^2 + \frac{M_L}{6}\|\Delta\|_2^3. \tag{6}$$

Then

$$B(\gamma_{\mathrm{lin}}) \le \min\{R(w_A), R(w_B)\} + \frac{1}{2}\left|L(w_A) - L(w_B)\right|. \tag{7}$$

(Proof in Appendix B.2.)

The only Fisher-geometric term in the anchor bound is the Fisher length $\mathcal{L}_F(w_0; \Delta)$; the remaining terms arise from endpoint gradients, predictive uncertainty, and higher-order smoothness. At exact minima the gradient term vanishes, and the inequality is one-sided, so no sharpness is claimed.

4.2 Shape: Width-Dependent Flattening

The relevant manifold is $(\mathbb{R}^P, G_F)$ with $G_F = F + \lambda I_P$. Since $\partial G_F = \partial F$, the damping does not alter derivative-flattening behavior. The Shape axis is carried by one scalar, the operator norm of the metric's own derivative,

$$\mathcal{S}_F(w) := \sup_{\|z\|_2 = 1} \|\partial_z F(w)\|_{\mathrm{op}},$$

written $\|\partial F\|_{\mathrm{op}}$ in the figures and measured in Figure 1b. Everything else on this axis follows from it: Theorem 4.4 bounds Γ and Sec through $\mathcal{S}_F$, and the geodesic–linear deviation of Section 5 is its path-level reading.

Assumption 4.2 (Lazy derivative flattening)**.** There exist $c_1, c_2 > 0$, an activation-dependent rate $\alpha(\sigma) > 0$, and a neighborhood U of the trained parameters such that, for all $w \in U$,

$$\mathcal{S}_F(w) \le \frac{c_1}{n^{\alpha(\sigma)}}, \qquad \sup_{\|h\|_2=\|z\|_2=1} \|\partial_h \partial_z F(w)\|_{\mathrm{op}} \le \frac{c_2}{n^{\alpha(\sigma)}}. \tag{8}$$

Assumption 4.3 (Uniform Fisher bound)**.** The quantity $M_F := \sup_{w\in U} \|F(w)\|_{\mathrm{op}}$ is finite, and hence $\sup_{w\in U} \|G_F(w)\|_{\mathrm{op}} \le M_F + \lambda$. Appendix C verifies this under NTK parameterization with the explicit value $M_F = C_J^2/2$, where C_J bounds the logit Jacobian.

Theorem 4.4 (Operator-norm metric flattening)**.** *Write $\beta := \sup_{w\in U} \mathcal{S}_F(w)$. Then, for all $u, v \in \mathbb{R}^P$, and under Assumption 4.2 for the second inequality,*

$$\|\Gamma(u,v)\|_2 \le \frac{3\beta}{2\lambda}\|u\|_2\|v\|_2 \le \frac{K_\Gamma^{op}}{n^{\alpha(\sigma)}}\|u\|_2\|v\|_2, \qquad K_\Gamma^{op} = \frac{3c_1}{2\lambda}, \tag{9}$$

with a constant independent of P. Furthermore, under Assumptions 4.2 and 4.3, for G_F-orthonormal u, v with $C_E = \sqrt{M_F + \lambda}/\lambda$,

$$\left|\mathrm{Sec}_{(\mathbb{R}^P, G_F)}(u,v)\right| \le \frac{K_{\mathrm{Sec}}^{op}}{n^{\alpha(\sigma)}} + O\Big(n^{-2\alpha(\sigma)}\Big), \qquad K_{\mathrm{Sec}}^{op} = \frac{3c_2\, C_E^2}{\lambda}, \tag{10}$$

also with constants independent of P. (Proof in Appendix B.3.)

Under standard NTK parameterization (Jacot et al., 2018; Lee et al., 2019), a frozen-softmax Fisher surrogate $\widetilde{F}$ satisfies Assumption 4.2 with $\alpha = 1/2$ (proof in Appendix C). For the true Fisher matrix, the derivative contains an additional softmax correction term $3C_J^3\rho_S(w)$, where

$$\rho_S(w) = \mathbb{E}_x\|S(p_w(x))\|_{\mathrm{op}} \in [0, 1/2].$$

Since $S(e_y) = 0$ and S is Lipschitz in p, one has $\rho_S(w) \le (3+\sqrt{2})\,\rho^*(w)$, so this correction is controlled when predictive uncertainty is sufficiently small. Thus, for the true Fisher matrix, derivative flattening additionally depends on the behavior of the predictive distribution.

4.3 Shape Controls Path Geometry, Not Barrier Height

The Christoffel bound in Theorem 4.4 shows that the Fisher connection becomes asymptotically Euclidean. Since $\ddot{\gamma}_{\mathrm{lin}} = 0$, the failure of γ_{lin} to satisfy the geodesic equation is measured at first order by $\Gamma_{\gamma_{\mathrm{lin}}}(\Delta, \Delta)$. Letting $\xi = \gamma_g - \gamma_{\mathrm{lin}}$ denote the deviation between the Fisher geodesic and the linear interpolation, we have:

Lemma 4.5 (Vector-valued Poincaré–Sobolev)**.** *If $\xi \in C^2([0,1]; \mathbb{R}^P)$ with $\xi(0) = \xi(1) = 0$, then*

$$\sup_t \|\xi(t)\|_2 \le \frac{1}{8}\sup_t \|\ddot{\xi}(t)\|_2.$$

(Proof in Appendix B.4.)

Applying Lemma 4.5 requires bounding the exact geodesic acceleration $\ddot{\xi} = -\Gamma_{\gamma_g}(\dot{\gamma}_g, \dot{\gamma}_g)$, which in turn requires control of the geodesic speed. Assuming the speed bound $\|\dot{\gamma}_g(t)\|_2 \le 2\|\Delta\|_2$ for all $t \in [0,1]$, the bilinear Christoffel estimate of Theorem 4.4 gives $\sup_t \|\ddot{\xi}(t)\|_2 \le 4\kappa\|\Delta\|_2^2$, where $\kappa := K_\Gamma^{\mathrm{op}}\, n^{-\alpha(\sigma)} = 3c_1/(2\lambda n^{\alpha(\sigma)})$, and Lemma 4.5 yields the estimate

$$\sup_{t\in[0,1]} \|\xi(t)\|_2 \;\le\; \tfrac{1}{8}\cdot 4\kappa\|\Delta\|_2^2 \;=\; \tfrac{1}{2}\,\kappa\,\|\Delta\|_2^2 \;=\; \frac{3c_1\,\|\Delta\|_2^2}{4\lambda\, n^{\alpha(\sigma)}}, \tag{11}$$

which is of order $n^{2s-\alpha(\sigma)}$ when the displacement scales as $\|\Delta\|_2 = \Theta(n^s)$ (details in Appendix B.4). Here s is a property of the regime rather than a constant: Appendix D measures $s \approx 0.94$ under NTK-lazy against $s \approx 0.26$ and $s \approx 0.39$ under Standard and μP. At first order, the deviation itself admits the representation

$$\sup_t \|\gamma_g(t) - \gamma_{\mathrm{lin}}(t)\|_2 \;\approx\; \sup_t \left\| \int_0^1 G(t,s)\,\Gamma_{\gamma_{\mathrm{lin}}(s)}(\Delta,\Delta)\,ds \right\|_2,$$

with $G(t,s) = s(1-t)$ for $s \le t$ and $G(t,s) = t(1-s)$ for $s \ge t$, which involves only Γ evaluated along the linear path. This representation serves as the basis for the direct empirical measurement of the geodesic–linear deviation in Section 5.

This Shape-axis flattening concerns the local variation of the Fisher manifold, not the values of the loss: $\mathcal{S}_F$ and the quantities it controls are built from derivatives of the metric, whereas the barrier depends on the loss itself. Empirically, the measured geodesic–linear deviation decreases with width in some regimes but not in others (Section 5), so path straightening is regime-dependent. Yet even when the manifold is locally flat and the path is nearly straight, the loss curvature can depart from F in regions of high predictive uncertainty (ρ^* large), as the Fisher-Hessian identity (Theorem 3.4) shows. Shape flattening therefore characterizes the geometry of the interpolation path, and nothing beyond it: a geodesic may be nearly straight and still carry a high barrier.

4.4 Length: Fisher Energy and Barrier Control

Since Shape flattening does not constrain the loss values along the path, we turn to the Length axis. The one Fisher-geometric term of the anchor bound is the Fisher length $\mathcal{L}_F$. Decomposing $\mathcal{L}_F = \|\Delta\|_2^2 \cdot \mathcal{R}_F$, where $\mathcal{R}_F = \widehat{\Delta}^\top F \widehat{\Delta}$ is the Rayleigh quotient, $\mathcal{R}_F \to 0$ says that the displacement increasingly concentrates in the low-eigenvalue subspace of F. However, $\mathcal{R}_F \to 0$ does not imply $\mathcal{L}_F \to 0$ unless $\|\Delta\|_2^2$ is controlled. The barrier bound depends on the full Fisher energy $\mathcal{L}_F$, not its normalized version.

Furthermore, the Length axis is informative precisely when predictive uncertainty is controlled. By the Fisher-Hessian identity, $\nabla^2 L \approx F$ only when ρ^* is small. When ρ^* is large, the uncertainty-dependent remainder $\frac{1}{2}\rho^*(w_0) M_f \|\Delta\|_2^2$ in equation 6 can dominate the Fisher energy, so the bound no longer certifies a small barrier even when $\mathcal{L}_F$ is small. Thus, Fisher length provides the geometric contribution to the barrier, while predictive uncertainty determines how tightly this geometric picture constrains it. This is consistent with the experiments in Section 5: metric flattening occurs in every setting, while barrier collapse is driven by Fisher length reduction and is strongest when predictive uncertainty is controlled.

5 Experiments

We validate the theoretical predictions of Section 4 through systematic experiments across architectures, activation functions, and parameterization regimes, testing whether the Shape–Length decomposition predicts when and why interpolation barriers vanish.

5.1 Experimental Setup

We evaluate MLPs, Teacher–Student (TS) networks, and CNNs across four activations (GELU, $\tanh$, Swish, Softplus) with widths 64–4096, trained under NTK-lazy, Standard, and μP (Yang & Hu, 2021). The MLP is trained on MNIST, the TS student on synthetic Gaussian inputs labeled by a frozen width-32 ReLU teacher, and the CNN on FashionMNIST, whose widest layer spans $n = 64$–512. The three parameterizations differ in how initialization and updates scale with width: NTK-lazy uses $O(1)$ initialization variance with a $1/\sqrt{n}$ forward rescaling per layer, Standard uses fan-in initialization, and μP transfers per-tensor initialization and optimizer scaling from an anchor width of 64. The base learning rate is not retuned per width, so width effects are not confounded with hyperparameter adaptation. Each width is trained from five seeds and all ten seed pairs are interpolated, 2160 pairs over the full grid (Appendix D).

Permutation symmetries are resolved via weight-matching alignment (Entezari et al., 2022; Ainsworth et al., 2023). For each aligned pair (w_A, w_B), we measure the barrier $B(\gamma_{\mathrm{lin}})$, metric-derivative norm $\|\partial_z F\|_{\mathrm{op}}$, geodesic–linear deviation, Rayleigh quotient $\mathcal{R}_F$, Fisher length $\mathcal{L}_F = \Delta^\top F \Delta$, and predictive uncertainty ρ^*. Figures 1–4 show medians over the four activations and the seed pairs (the seeds, for the per-network $\|\partial_z F\|_{\mathrm{op}}$), with interquartile bands, each curve normalized by its value at the smallest width; quantities evaluated at an endpoint are read at the anchor w_A throughout.

5.2 Shape-Axis Geometry and Barrier Decoupling

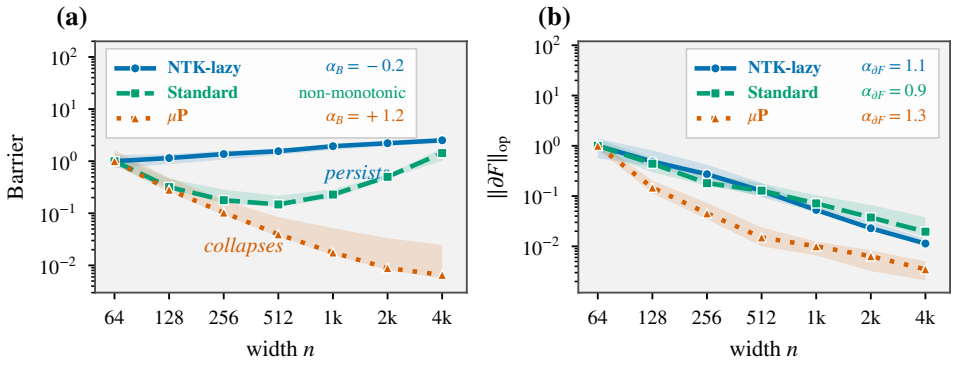

Figure 1: Geometry flattens in every regime; the barrier does not. Barrier (a) and the Shape scalar $\|\partial F\|_{\mathrm{op}}$ (b) against width, MLP/MNIST.

The Fisher metric flattens with width in every regime we study. Figure 1b shows $\|\partial F\|_{\mathrm{op}}$ decaying as a power law under all three parameterizations, at rates faster than the $\alpha = 1/2$ proved for the frozen-softmax surrogate. However, the interpolation barrier does not follow this geometric flattening. Instead, its behavior is three-way: NTK-lazy grows, μP collapses, and Standard falls with width up to a point and reverses beyond it. Metric flattening therefore occurs in all three regimes, yet is insufficient on its own to drive barrier collapse.

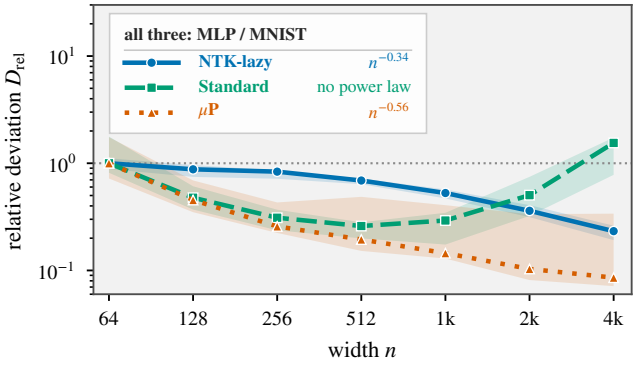

Figure 2: Path straightening is regime-dependent, so it cannot be what separates the regimes either. Relative geodesic–linear deviation $D_{\mathrm{rel}} = \sup_t \|\gamma_g - \gamma_{\mathrm{lin}}\|_2 / \|\Delta\|_2$ against width, MLP/MNIST.

The same split appears one level up, at the path rather than the metric. Figure 2 measures the geodesic–linear deviation through the first-order Green representation; what the figure reports is a direct measurement rather than a consequence of Theorem 4.4. Under NTK-lazy and μP the deviation decays as a power law, so the Fisher geodesic straightens into the linear interpolation. Under Standard it decays up to $n = 512$ and rebounds, mirroring the barrier reversal of Figure 1a. Neither flattening nor straightening decides the barrier: under NTK-lazy the geodesic is closest to the straight line, yet the barrier grows in that regime. Appendix E.5 repeats the panel on the other two families.

5.3 Decomposing Fisher Length: Spectral Alignment and Uncertainty

Having established that the Shape axis does not govern barrier collapse, we turn to the Length axis. Theorem 4.1 identifies the Fisher quadratic form $\mathcal{L}_F = \Delta^\top F \Delta$ as the energetic term controlling the barrier, which decomposes into an Euclidean scale and a directional Rayleigh quotient: $\mathcal{L}_F = \|\Delta\|_2^2 \cdot \mathcal{R}_F$, where $\mathcal{R}_F = \widehat{\Delta}^\top F \widehat{\Delta}$. We first test whether this directional curvature component $\mathcal{R}_F$ alone can explain barrier behavior.

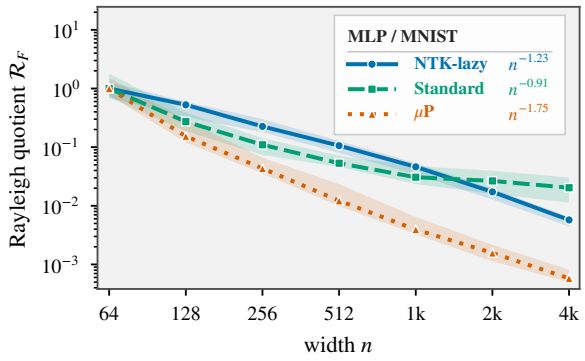

Figure 3: Directional alignment improves in every regime, so it separates nothing. Rayleigh quotient $\mathcal{R}_F = \widehat{\Delta}^\top F(w_A)\widehat{\Delta}$ against width, MLP/MNIST.

Directional alignment improves in every regime, so it cannot be what separates them. Figure 3 shows $\mathcal{R}_F$ decaying steeply with width in all three regimes, though the Standard curve saturates beyond $n = 1024$; the displacement concentrates in the low-eigenvalue subspace of F whether the barrier collapses or not. Pooled across the grid, $\mathcal{R}_F$ explains little of the barrier ($R^2 = 0.20$, against 0.63 for the full Fisher length at the same anchor), and what it does explain does not survive being refitted one regime or one architecture at a time. This is what Theorem 4.1 would lead one to expect: the bound contains the full Fisher length, not its normalized part, and it is gated by predictive uncertainty.

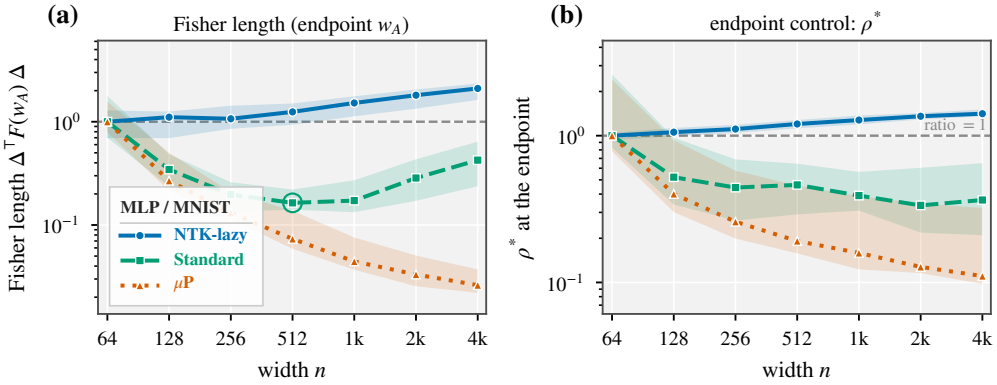

Figure 4: Uncertainty gates the bound and Fisher energy decides the trajectory. Endpoint Fisher length $\mathcal{L}_F$ (a) and endpoint uncertainty ρ^* (b) against width, MLP/MNIST.

The regimes are separated by the two conditions the bound requires. The first acts as a gate. Figure 4b shows that NTK-lazy never suppresses its endpoint uncertainty: ρ^* stays at or above its smallest-width value, so the remainder term of Theorem 4.1 stays active and spectral alignment alone cannot close the bound. Standard and μP suppress ρ^* steadily, and only for them does the energetic term take over. The second condition is the energy itself. Figure 4a shows $\mathcal{L}_F$ collapsing monotonically under μP, which is the regime whose barrier collapses, and bottoming at $n = 512$ before rebounding under Standard, which is the regime whose barrier turns at the same width. Regime divergence follows from the two together, not from either alone.

5.4 Fisher Length Governs Barrier

We next evaluate whether $\mathcal{L}_F$, read at an endpoint anchor w_0, quantitatively governs barrier scaling across the full experimental grid, as established by the upper bound in Theorem 4.1. Evaluating at trained minima anchors $w_0 \in \{w_A, w_B\}$ where predictive uncertainty is controlled ($\rho^*(w_0) \approx 0$) ensures that this energetic term dominates the bound.

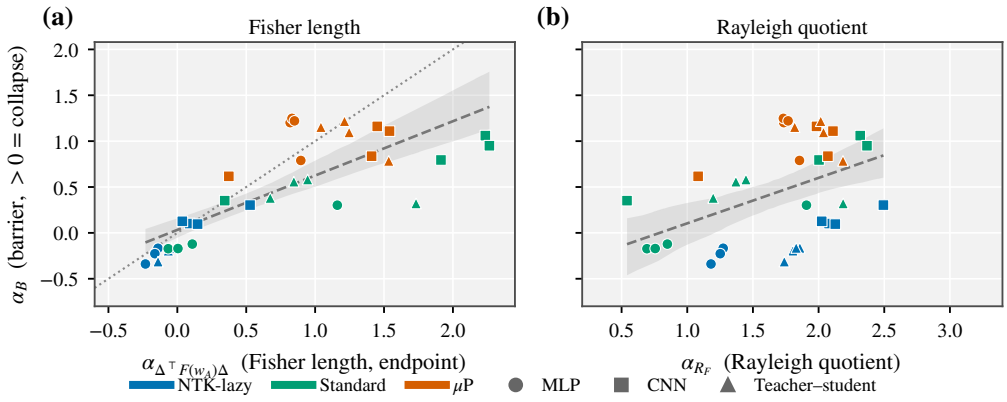

Figure 5: Fisher length governs barrier decay; its directional part alone does not. Barrier exponent α_B against the Fisher-length (a) and Rayleigh-quotient (b) exponents, over every cell of the grid.

Across the grid, the barrier exponent tracks the Fisher-length exponent and not the Rayleigh one. Figure 5a gives $R^2 = 0.63$ for the least-squares fit, with the diagonal $y = x$ drawn for reference only, while the directional part, read at the same anchor so that the two panels differ in what they measure and not in where, explains three times less ($R^2 = 0.20$). This coupling is not an artifact of clustering. Refitted within each regime and each architecture separately, the Fisher-length slope stays positive on all six splits and is significant on five; the exception is μP, where every cell has already collapsed and little spread is left to explain. The Rayleigh fit does not survive the same splits: it falls to $R^2 = 0.09$ under μP and changes slope ninefold between architectures. Because both axes are fitted exponents, errors-in-variables attenuate the least-squares slope (acting as a lower bound). Appendix E.4 repeats this comparison on CIFAR-10, with the architecture, the three parameterizations, the four activations and the channel multipliers held at the released CNN setting so that the dataset is the only thing that changes: the Fisher-length coupling is as tight there as here, and the Rayleigh quotient explains as little. These findings confirm that the primary geometric driver of linear mode connectivity is the full Fisher length $\mathcal{L}_F$, rather than directional spectral alignment.

6 Conclusion

We have shown that the geometric structure underlying Linear Mode Connectivity decomposes into two decoupled axes of Fisher information geometry. Along the Shape axis, we prove that once the metric's own derivative $\mathcal{S}_F$ decays with width, the Christoffel symbols and the sectional curvature inherit that decay with constants independent of P. The geodesic–linear deviation falls with width under the lazy and μP regimes but not under Standard, so path rectification is regime-dependent. Along the Length axis, we derive a Fisher-barrier inequality whose only Fisher-geometric term is the Fisher length $\mathcal{L}_F$, gated by predictive uncertainty ρ^* rather than by curvature.

The central finding is that these axes are logically independent: $\mathcal{S}_F$ decays across architectures, activations, and parameterization regimes, yet barrier collapse occurs only when Fisher length reduction is accompanied by controlled predictive uncertainty. This decoupling accounts for the persistent barriers observed in geometrically flattened lazy regimes and for the three-way barrier behavior—persistent, collapsing, and non-monotonic—that no single geometric scalar can capture. The Shape–Length decomposition provides a geometric framework for interpreting regime-dependent mode connectivity and motivates Fisher-geometric criteria for model merging, weight averaging, and connectivity-aware parameterization.

AI use statement

In this work we used generative AI tools to help with programming tasks such as code drafting, debugging, refactoring and implementation support, as well as language editing and clarifying the manuscript. We did not produce experimental results, mathematical proofs, or final scientific statements with generative AI tools. All AI-assisted code has been reviewed, tested and verified by the

authors. All text edited with the help of generative AI was carefully reviewed and revised by the authors. We are responsible for the final contents of this work, including the correctness of the code, experiments, theoretical results and conclusions.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, Providence, RI, 2000. ISBN 978-0-8218-0531-2. doi: 10.1090/mmono/191.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8139–8148, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dbc4d84bfcfe2284ba11beffb853a8c4-Abstract.html.

Vincent Bürgin, Daniel Herbst, Ya-Wei Eileen Lin, and Stefanie Jegelka. Beyond structural symmetries: Linear mode connectivity via neuron identifiability. *CoRR*, abs/2606.04754, 2026. doi: 10.48550/ARXIV.2606.04754. URL https://doi.org/10.48550/arXiv.2606.04754.

N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982. URL https://books.google.com/books?id=kTJJtQAACAAJ.

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2933–2943, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL http://proceedings.mlr.press/v70/dinh17b.html.

M.P. do Carmo. *Riemannian Geometry*. Mathematics (Boston, Mass.). Birkhäuser, 1992. URL https://books.google.com/books?id=uXJQQgAACAAJ.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1308–1317. PMLR, 2018. URL http://proceedings.mlr.press/v80/draxler18a.html.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=dNigytemkL.

Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, and Aymeric Dieuleveut. Proving linear mode connectivity of neural networks via optimal transport. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3853–3861. PMLR, 2024. URL https://proceedings.mlr.press/v238/ferbach24a.html.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3259–3269. PMLR, 2020. URL http://proceedings.mlr.press/v119/frankle20a.html.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf.

Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Analysis of linear mode connectivity via permutation-based weight matching: With insights into other permutation search methods. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=lYRkGZZi9D.

Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Linear mode connectivity between multiple models modulo permutation symmetries. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/ito25a.html.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: renormalizing permuted activations for interpolation repair. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=gU5sJ6ZggcX.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1032–1041. PMLR, 2019. URL http://proceedings.mlr.press/v89/karakida19a.html.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4158–4169, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8570–8581, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b7ae8fecf15b8b6c3c69eceae636d203-Abstract.html.

James Martens. New insights and perspectives on the natural gradient method. *J. Mach. Learn. Res.*, 21:146:1–146:76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *CoRR*, abs/2008.11865, 2020. URL https://arxiv.org/abs/2008.11865.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica (Cluj)*, 9:129–145, 1935.

Xingyu Qu and Samuel Horváth. Vanishing feature: Diagnosing model merging and beyond. In *Conference on Parsimony and Learning, Stanford University, USA, 24-27 March 2025*, volume 280 of *Proceedings of Machine Learning Research*, pp. 1051–1086. PMLR, 2025. URL https://proceedings.mlr.press/v280/qu25a.html.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJO1_M0Lf.

Ekansh Sharma, Devin Kwok, Tom Denton, Daniel M. Roy, David Rolnick, and Gintare Karolina Dziugaite. Simultaneous linear connectivity of neural networks modulo permutation. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9-13, 2024, Proceedings, Part VII*, volume 14947 of *Lecture Notes in Computer Science*, pp. 262–279. Springer, 2024a. doi: 10.1007/978-3-031-70368-3_16. URL https://doi.org/10.1007/978-3-031-70368-3_16.

Ekansh Sharma, Daniel M. Roy, and Gintare Karolina Dziugaite. The non-local model merging problem: Permutation symmetries and variance collapse. *CoRR*, abs/2410.12766, 2024b. doi: 10.48550/ARXIV.2410.12766. URL https://doi.org/10.48550/arXiv.2410.12766.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732. PMLR, 2021. URL http://proceedings.mlr.press/v139/simsek21a.html.

Sidak Pal Singh, Linara Adilova, Michael Kamp, Asja Fischer, Bernhard Schölkopf, and Thomas Hofmann. Landscaping linear mode connectivity. *CoRR*, abs/2406.16300, 2024. doi: 10.48550/ARXIV.2406.16300. URL https://doi.org/10.48550/arXiv.2406.16300.

Derek Tam, Mohit Bansal, and Colin Raffel. Merging by matching models in task parameter subspaces. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=qNGo6ghWFB.

Tiffany J. Vlaar and Jonathan Frankle. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22325–22341. PMLR, 2022. URL https://proceedings.mlr.press/v162/vlaar22a.html.

Jiayu Wang, Zuojun Ye, and Wenpeng Yin. Functionality-oriented LLM merging on the fisher-rao manifold. *CoRR*, abs/2603.04972, 2026. doi: 10.48550/ARXIV.2603.04972. URL https://doi.org/10.48550/arXiv.2603.04972.

Greg Yang and Edward J. Hu. Tensor programs IV: feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL http://proceedings.mlr.press/v139/yang21c.html.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: tuning large neural networks via zero-shot hyperparameter transfer. *CoRR*, abs/2203.03466, 2022. doi: 10.48550/ARXIV.2203.03466. URL https://doi.org/10.48550/arXiv.2203.03466.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding mode connectivity via parameter space symmetry. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhao25i.html.

APPENDIX

The appendix is organized as follows. Appendix A is a table of the notation used in Section 3. Appendix B proves the four results stated in the main text: the Fisher–Hessian identity (Theorem 3.4), the Fisher–barrier inequality (Theorem 4.1), the operator-norm metric-flattening bounds (Theorem 4.4), and the vector-valued Poincaré–Sobolev inequality (Lemma 4.5). Appendix C verifies Assumption 4.2 in the NTK regime for the frozen-softmax Fisher surrogate, with the rate $\alpha = 1/2$, and isolates the softmax correction that separates the surrogate from the true Fisher matrix. Appendix D gives the experimental setup — models, training, and the matrix-free estimators behind every measured quantity — and Appendix E reports the full numerical results underlying the figures of Section 5.

A NOTATION

Every symbol used in Sections 3–5 is listed below.

Symbol	Meaning
K	the number of classes, and hence of logits
P	the number of trainable parameters
n	the hidden width of the network
$\mathcal{D}$	the data distribution, over *labeled pairs* (x, y)
$\mathcal{D}_X$	the marginal of $\mathcal{D}$ on inputs alone, labels discarded
$f(x; w)$, $f(x; \cdot)$	the logits at x and w; a dot marks the argument left free
$J(x; w)$	the Jacobian $\partial f(x; w)/\partial w \in \mathbb{R}^{K \times P}$ of the logits in w
$H_k(x; w)$	the Hessian $\nabla_w^2 f_k(x; w)$ of the k-th logit in w
$p_w(x)$	$\mathrm{softmax}(f(x; w))$, a probability vector in $\mathbb{R}^K$
$S(p)$	$\mathrm{diag}(p) - pp^\top$, the covariance of a one-hot draw from p
$\rho^*(w)$	predictive uncertainty, $\mathbb{E}\|p_w(x) - e_y\|_2 \in [0, \sqrt{2}]$
U	an open subset of $\mathbb{R}^P$ containing the parameters of interest
$\widehat{v}$	the unit vector $v/\|v\|_2$
$\mathrm{D}^3 L(w)[u, v, z]$	the third directional derivative of L at w along u, v, z
Uniform bounds, indexed by the object bounded rather than by derivative order:	
M_f	bounds the second derivative of the logits (Definition 3.3)
M_L	bounds the third derivative of the loss (Definition 3.3)
M_F	bounds $\|F\|_{\mathrm{op}}$ on U (Assumption 4.3); $C_J^2/2$ under NTK
M_σ	bounds the first three derivatives of the activation (Appendix C)
C_J, C_H	bound the logit Jacobian and logit Hessian in Appendix C
The interpolation path, and the two geometric axes measured along it:	
Δ	the displacement $w_B - w_A$ between the two aligned solutions
γ_{lin}	the linear path $t \mapsto w_A + t\Delta$; γ_g is its Fisher geodesic
$B(\gamma_{\mathrm{lin}})$	the interpolation barrier along that path
F, G_F	the Fisher matrix, and the damped metric $F + \lambda I_P$ with $\lambda > 0$
$\alpha(\sigma)$	the activation-dependent rate of Assumption 4.2
$\mathcal{S}_F(w)$	*Shape*: $\sup_{\|z\|_2=1} \|\partial_z F(w)\|_{\mathrm{op}}$, written $\|\partial F\|_{\mathrm{op}}$ in the figures
D_{rel}	the path-level reading of Shape, $\sup_t \|\gamma_g - \gamma_{\mathrm{lin}}\|_2 / \|\Delta\|_2$
$\mathcal{L}_F(w; \Delta)$	*Length*: the Fisher quadratic form $\Delta^\top F(w) \Delta$
$\mathcal{R}_F$	the Rayleigh quotient $\widehat{\Delta}^\top F \widehat{\Delta}$, so $\mathcal{L}_F = \|\Delta\|_2^2 \, \mathcal{R}_F$
Riemannian conventions, with a repeated index summed over $1, \ldots, P$*:*	
Γ_{ij}^k	the Christoffel symbols of the Levi-Civita connection of G_F
$\Gamma(u, v)$	the same object as a bilinear map, with k-th component $\Gamma_{ij}^k u^i v^j$
$\mathrm{Sec}(u, v)$	the sectional curvature of the section spanned by u and v

Two averages must be kept apart, and the subscript is the only thing separating them: the loss L and the uncertainty ρ^* average over $\mathcal{D}$, because each compares a prediction against a label, whereas the Fisher matrix F averages over $\mathcal{D}_X$, because it is a property of the model's own predictive distribution

and no label enters it. The two geometric axes are told apart the same way, by which object they are built from: $\mathcal{S}_F$ from the *derivatives* of F, and $\mathcal{L}_F$ from F *itself*, contracted with the displacement.

B PROOFS

B.1 PROOF OF THEOREM 3.4

Proof. Write $\ell(w; x, y) := -\log p_w(x)_y$, so that $L(w) = \mathbb{E}_{(x,y)}[\ell(w; x, y)]$. From the chain rule through the logits and the softmax differential $\nabla_f(-\log p_y) = p - e_y$ (from $\log p_k = f_k - \log \sum_j e^{f_j}$), we have

$$\nabla_w \ell = J(x; w)^\top \big(p_w(x) - e_y\big), \qquad \text{i.e.} \qquad (\nabla_w \ell)_i = \sum_{k=1}^{K} J_{ki} \big(p_k - (e_y)_k\big). \tag{12}$$

Differentiating equation 12 once more with respect to w_j, the product rule splits $\partial_j \partial_i \ell$ into two terms,

$$\begin{aligned} \partial_j \partial_i \ell &= \sum_{k=1}^{K} (\partial_j J_{ki}) \big(p_k - (e_y)_k\big) + \sum_{k=1}^{K} J_{ki}\, \partial_j p_k \\ &= \sum_{k=1}^{K} \big(p_k - (e_y)_k\big)(H_k)_{ij} + \big(J(x; w)^\top S(p_w(x))\, J(x; w)\big)_{ij}, \end{aligned} \tag{13}$$

the first term by $\partial_j J_{ki} = (H_k)_{ij}$ and the second by

$$\partial_j p_k = \sum_{m=1}^{K} \frac{\partial p_k}{\partial f_m} J_{mj} = \big(S(p_w(x))\, J(x; w)\big)_{kj}, \qquad \sum_{k=1}^{K} J_{ki} \big(S(p) J\big)_{kj} = \big(J^\top S(p) J\big)_{ij}. \tag{14}$$

In matrix form, equation 13 reads

$$\operatorname{Hess}_w \ell = J(x; w)^\top S(p_w(x)) J(x; w) + \sum_{k=1}^{K} \big(p_w(x)_k - (e_y)_k\big) H_k(x; w). \tag{15}$$

Both sides of equation 15 are jointly continuous in (w, x, y) (y ranges over a finite set) and $\operatorname{supp}(\mathcal{D}_X)$ is compact, so exchanging the expectation and the differentiation is justified by dominated convergence; the first term being independent of y, we obtain

$$\begin{aligned} \operatorname{Hess} L(w) &= \mathbb{E}_x\big[J(x; w)^\top S(p_w(x)) J(x; w)\big] + \mathbb{E}_{(x,y)}\Big[\sum_{k=1}^{K} \big(p_w(x)_k - (e_y)_k\big) H_k(x; w)\Big] \\ &= F(w) + E(w), \end{aligned} \tag{16}$$

which is the identity, and is the Gauss–Newton splitting of the loss Hessian: F is the term that survives when the residual $p_w(x) - e_y$ vanishes.

For the residual bound, fix $v \in \mathbb{R}^P$ and set, for each (x, y),

$$r := p_w(x) - e_y, \qquad q_k := v^\top H_k(x; w) v, \qquad \text{so that} \qquad v^\top E(w) v = \mathbb{E}_{(x,y)}\big[\langle r, q\rangle\big]. \tag{17}$$

The factor $\|q\|_2$ is bounded pointwise in (x, y): from $|q_k| \le \|H_k\|_{\mathrm{op}} \|v\|_2^2$ and $\|q\|_2 \le \sqrt{K} \max_k |q_k|$,

$$\|q\|_2 \le \sqrt{K} \Big(\sup_{w \in U,\, x,\, k} \|H_k(x; w)\|_{\mathrm{op}} \Big) \|v\|_2^2 = M_f \, \|v\|_2^2, \tag{18}$$

which is where the factor $\sqrt{K}$ of Definition 3.3 is consumed, exactly once. Since equation 18 is a constant, it may be pulled out of the expectation; then $\|\mathbb{E} X\|_2 \le \mathbb{E}\|X\|_2$ and Cauchy–Schwarz on $\mathbb{R}^K$ pointwise in (x, y) give

$$\begin{aligned} \big|v^\top E(w) v\big| &\le \mathbb{E}\big[\|r\|_2 \, \|q\|_2\big] \\ &\le M_f \, \|v\|_2^2 \, \mathbb{E}_{(x,y)}\big[\|p_w(x) - e_y\|_2\big] \\ &= \rho^*(w)\, M_f \, \|v\|_2^2, \end{aligned} \tag{19}$$

the last line by Definition 3.2. Since $E(w)$ is a real-weighted sum of the symmetric matrices H_k, equation 19 also gives $\|E(w)\|_{\mathrm{op}} \le \rho^*(w) M_f$; contracting equation 16 with v and inserting equation 19 in both directions gives

$$v^\top F(w) v - \rho^*(w) M_f \|v\|_2^2 \;\le\; v^\top \mathrm{Hess}\, L(w)\, v \;\le\; v^\top F(w) v + \rho^*(w) M_f \|v\|_2^2. \tag{20}$$

□

B.2 Proof of Theorem 4.1

Proof. Write $\delta := \frac{1}{2}|L(w_A) - L(w_B)|$ and $\phi(t) := L(w_A + t\Delta)$, and recall the anchor quantity of Theorem 4.1,

$$R(w_0) := \|\Delta\|_2 \, \|\nabla L(w_0)\|_2 + \frac{1}{2}\, \Delta^\top F(w_0) \Delta + \frac{1}{2}\, \rho^*(w_0) M_f \, \|\Delta\|_2^2 + \frac{M_L}{6}\, \|\Delta\|_2^3. \tag{21}$$

Iterating the chain rule contracts one index against Δ per differentiation, so at $w_A + t\Delta$

$$\phi'(t) = \Delta^\top \nabla L, \qquad \phi''(t) = \Delta^\top \mathrm{Hess}\, L\, \Delta, \qquad \phi'''(t) = D^3 L[\Delta, \Delta, \Delta], \qquad \big|\phi'''(t)\big| \le M_L \|\Delta\|_2^3, \tag{22}$$

the last bound by trilinearity of $D^3 L$ and $w_A + t\Delta \in U$. Taylor's theorem with Lagrange remainder centered at w_A gives, for each $t \in [0, 1]$, a point $w_\xi = w_A + \xi_t \Delta \in [w_A, w_B]$ with

$$L(w_A + t\Delta) = L(w_A) + t\, \Delta^\top \nabla L(w_A) + \frac{t^2}{2}\, \Delta^\top \mathrm{Hess}\, L(w_A) \Delta + \frac{t^3}{6}\, D^3 L(w_\xi)[\Delta, \Delta, \Delta]. \tag{23}$$

Since $0 \le t \le 1$, the three t-dependent terms satisfy

$$\begin{aligned} t\, \Delta^\top \nabla L(w_A) &\le \|\Delta\|_2 \, \|\nabla L(w_A)\|_2, \\ \frac{t^2}{2}\, \Delta^\top \mathrm{Hess}\, L(w_A) \Delta &\le \max\Big\{0,\ \tfrac{1}{2}\, \Delta^\top \mathrm{Hess}\, L(w_A) \Delta\Big\} \le \tfrac{1}{2}\, \Delta^\top F(w_A) \Delta + \tfrac{1}{2}\, \rho^*(w_A) M_f \, \|\Delta\|_2^2, \\ \frac{t^3}{6}\, D^3 L(w_\xi)[\Delta, \Delta, \Delta] &\le \frac{M_L}{6}\, \|\Delta\|_2^3. \end{aligned} \tag{24}$$

The lines are Cauchy–Schwarz on $\mathbb{R}^P$; $t^2 \le 1$ with the upper half of equation 20, whose nonnegative right-hand side ($F \succeq 0$) also covers the branch $\max\{0, \cdot\} = 0$; and equation 22 at $w_\xi \in U$. Inserting these t-free bounds into equation 23 and taking $\max_t$, attained since $L \circ \gamma_{\mathrm{lin}}$ is continuous on $[0, 1]$, gives

$$B(\gamma_{\mathrm{lin}}) \le R(w_A) + \tfrac{1}{2}\big(L(w_A) - L(w_B)\big). \tag{25}$$

Reparametrising by $s = 1 - t$ writes $\gamma_{\mathrm{lin}}(t) = w_B + s(-\Delta)$; repeating the argument with $(w_B, -\Delta)$ produces the same bounds at w_B, each invariant under $\Delta \mapsto -\Delta$,

$$\|-\Delta\|_2 = \|\Delta\|_2, \qquad (-\Delta)^\top F(w_B)(-\Delta) = \Delta^\top F(w_B) \Delta, \tag{26}$$

so the right-hand side is exactly $R(w_B)$:

$$B(\gamma_{\mathrm{lin}}) \le R(w_B) + \tfrac{1}{2}\big(L(w_B) - L(w_A)\big). \tag{27}$$

Relabelling if necessary so that $L(w_A) \ge L(w_B)$, we have $\delta = \frac{1}{2}(L(w_A) - L(w_B)) \ge 0$, so equation 25 and equation 27 read $B \le R(w_A) + \delta$ and $B \le R(w_B) - \delta$; therefore

$$\begin{aligned} B(\gamma_{\mathrm{lin}}) &\le \min\big\{R(w_A) + \delta,\ \ R(w_B) - \delta\big\} \\ &\le \min\big\{R(w_A) + \delta,\ \ R(w_B) + \delta\big\} = \min\{R(w_A), R(w_B)\} + \delta, \end{aligned} \tag{28}$$

the last step by $\delta \ge 0$, which is the stated bound of Theorem 4.1. □

B.3 Proof of Theorem 4.4

Proof. Write $G = G_F = F + \lambda I$ and suppress the base point; since λI is constant, $\partial G = \partial F$ at every order. Throughout, β and β_2 denote the hypothesis constants of Theorem 4.4,

$$\beta := \sup_{w \in U} \sup_{\|z\|_2 = 1} \|\partial_z F(w)\|_{\mathrm{op}}, \qquad \beta_2 := \sup_{w \in U} \sup_{\|h\|_2 = \|z\|_2 = 1} \|\partial_h \partial_z F(w)\|_{\mathrm{op}}, \tag{29}$$

and the width-dependent rates enter only at the two concluding displays, where Assumption 4.2 licenses $\beta = c_1 n^{-\alpha(\sigma)}$ and $\beta_2 = c_2 n^{-\alpha(\sigma)}$.

For the Christoffel bound, contracting the Levi-Civita symbols $\Gamma^k_{ij} = \frac{1}{2}\sum_\ell (G^{-1})^{k\ell}(\partial_i G_{\ell j} + \partial_j G_{\ell i} - \partial_\ell G_{ij})$ with $u^i v^j$ reassembles the three contracted sums into $((\partial_u F)v)_\ell$, $((\partial_v F)u)_\ell$ and $m_w(u,v)_\ell := u^\top(\partial_\ell F)v$, giving

$$\Gamma_w(u,v) = \frac{1}{2}\, G_F(w)^{-1}\big[(\partial_u F)v + (\partial_v F)u - m_w(u,v)\big]. \tag{30}$$

Since $F \succeq 0$ gives $G_F \succeq \lambda I$, the eigenvalues of G_F^{-1} are at most $1/\lambda$:

$$\big\|G_F^{-1}\big\|_{\mathrm{op}} \le \frac{1}{\lambda}. \tag{31}$$

Linearity of $z \mapsto \partial_z F$ and equation 29 give $\|(\partial_u F)v\|_2 \le \beta\|u\|_2\|v\|_2$, and the same with u, v exchanged. For the third term of equation 30, probing $m_w(u,v)$ against a unit direction y reassembles the P coordinate derivatives into one directional derivative,

$$y^\top m_w(u,v) = \sum_{\ell=1}^{P} y_\ell\, u^\top(\partial_\ell F)v = u^\top(\partial_y F)v, \tag{32}$$

so $|y^\top m_w(u,v)| \le \beta\|u\|_2\|v\|_2$ for $\|y\|_2 = 1$; taking the supremum over y, the dual characterisation of the Euclidean norm, gives $\|m_w(u,v)\|_2 \le \beta\|u\|_2\|v\|_2$. Inserting the three bounds and equation 31 into equation 30,

$$\|\Gamma_w(u,v)\|_2 \le \frac{1}{2\lambda}\big(\beta+\beta+\beta\big)\|u\|_2\|v\|_2 = \frac{3\beta}{2\lambda}\,\|u\|_2\|v\|_2 =: \kappa\,\|u\|_2\|v\|_2, \tag{33}$$

and specialising equation 29 to $\beta = c_1 n^{-\alpha(\sigma)}$ yields the first claim,

$$\|\Gamma_w(u,v)\|_2 \le \frac{K_\Gamma^{\mathrm{op}}}{n^{\alpha(\sigma)}}\,\|u\|_2\|v\|_2, \qquad K_\Gamma^{\mathrm{op}} = \frac{3c_1}{2\lambda}, \tag{34}$$

in which K_Γ^{op} carries no factor of P because all three bounds entering equation 33 are aggregate.

For the sectional-curvature bound, differentiating equation 30 along u splits $(\partial_u \Gamma_w)(v,z)$ into the term where the derivative hits the bracket and the term where it hits G_F^{-1}. From $G_F G_F^{-1} = I$ and the product rule, $\partial_u G_F^{-1} = -G_F^{-1}(\partial_u F)G_F^{-1}$, so equation 31 gives

$$\big\|\partial_u G_F^{-1}\big\|_{\mathrm{op}} \le \frac{\beta}{\lambda^2}\,\|u\|_2. \tag{35}$$

In the bracket-differentiated term, $\|\partial_u\partial_v F\|_{\mathrm{op}} \le \beta_2\|u\|_2\|v\|_2$ bounds two of the three components. The third needs the second-order probe

$$y^\top \partial_u m_w(v,z) = \sum_{\ell=1}^{P} y_\ell\, v^\top(\partial_u\partial_\ell F)z = v^\top(\partial_u\partial_y F)z, \tag{36}$$

where Schwarz's theorem licenses reading $\sum_\ell y_\ell\,\partial_u\partial_\ell F$ as $\partial_u\partial_y F$; taking the supremum over unit y gives $\|\partial_u m_w(v,z)\|_2 \le \beta_2\|u\|_2\|v\|_2\|z\|_2$. With equation 31 the bracket-differentiated term is at most $\frac{3\beta_2}{2\lambda}\|u\|_2\|v\|_2\|z\|_2$. With equation 35 and the bracket norm $\le 3\beta\|v\|_2\|z\|_2$ from the Christoffel part, the inverse-differentiated term is at most $\frac{3\beta^2}{2\lambda^2}\|u\|_2\|v\|_2\|z\|_2$. Adding,

$$\|(\partial_u\Gamma_w)(v,z)\|_2 \le L_\Gamma\,\|u\|_2\|v\|_2\|z\|_2, \qquad L_\Gamma := \frac{3\beta_2}{2\lambda} + \frac{3\beta^2}{2\lambda^2}. \tag{37}$$

For constant vector fields the curvature tensor is $R(u,v)z = (\partial_u\Gamma)(v,z) - (\partial_v\Gamma)(u,z) + \Gamma(u,\Gamma(v,z)) - \Gamma(v,\Gamma(u,z))$; the two derivative terms contribute $2L_\Gamma$ by equation 37 and the two quadratic terms $2\kappa^2$ by applying equation 33 twice, so

$$\|R(u,v)z\|_2 \le C_R\,\|u\|_2\|v\|_2\|z\|_2, \qquad C_R := 2L_\Gamma + 2\kappa^2 = \frac{3\beta_2}{\lambda} + \frac{15\beta^2}{2\lambda^2}. \tag{38}$$

Finally, for G_F-orthonormal u, v the sectional curvature is $\mathrm{Sec}(u,v) = u^\top G_F R(u,v)v$, so

$$\big|\mathrm{Sec}(u,v)\big| \le \|G_F\|_{\mathrm{op}}\, C_R\, \|u\|_2^2 \|v\|_2^2 \le \frac{(M_F+\lambda)\, C_R}{\lambda^2}, \tag{39}$$

the last step by $\|G_F\|_{\mathrm{op}} \le M_F + \lambda$ (Assumption 4.3) and $\|u\|_2^2, \|v\|_2^2 \le 1/\lambda$, which follows from $1 = \|u\|_{G_F}^2 = u^\top(F+\lambda I)u \ge \lambda\|u\|_2^2$. Substituting $\beta = c_1 n^{-\alpha(\sigma)}$ and $\beta_2 = c_2 n^{-\alpha(\sigma)}$ into equation 39 gives the second claim,

$$\big|\mathrm{Sec}(u,v)\big| \le \frac{K_{\mathrm{Sec}}^{\mathrm{op}}}{n^{\alpha(\sigma)}} + O\big(n^{-2\alpha(\sigma)}\big), \qquad K_{\mathrm{Sec}}^{\mathrm{op}} = \frac{3c_2 C_E^2}{\lambda}, \quad C_E = \frac{\sqrt{M_F+\lambda}}{\lambda}, \tag{40}$$

and the hidden constant of the $O(n^{-2\alpha(\sigma)})$ tail is $15c_1^2(M_F+\lambda)/(2\lambda^4)$. □

B.4 Proof of Lemma 4.5

Proof. Let $\xi \in C^2([0,1];\mathbb{R}^P)$ with $\xi(0) = \xi(1) = 0$, and define the Green kernel of the two-point boundary value problem,

$$G(t,s) = \begin{cases} s(1-t), & 0 \le s \le t, \\ t(1-s), & t \le s \le 1, \end{cases} \tag{41}$$

whose two branches agree at $s = t$, both equalling $t(1-t)$, so that $G \ge 0$ on $[0,1]^2$, and set

$$\eta(t) := -\int_0^1 G(t,s)\ddot{\xi}(s)\, ds = -(1-t)\int_0^t s\,\ddot{\xi}(s)\, ds - t\int_t^1 (1-s)\ddot{\xi}(s)\, ds. \tag{42}$$

Then $\eta(0) = \eta(1) = 0$, and differentiating by the product rule and Leibniz's rule, the two boundary contributions $\mp\, t(1-t)\ddot{\xi}(t)$ cancelling, gives

$$\begin{aligned} \dot{\eta}(t) &= \int_0^t s\,\ddot{\xi}(s)\, ds - \int_t^1 (1-s)\ddot{\xi}(s)\, ds, \\ \ddot{\eta}(t) &= t\,\ddot{\xi}(t) + (1-t)\ddot{\xi}(t) = \ddot{\xi}(t). \end{aligned} \tag{43}$$

Hence every component of $\xi - \eta$ is affine and vanishes at the two points $t = 0, 1$, so $\xi - \eta \equiv 0$; that is,

$$\xi(t) = -\int_0^1 G(t,s)\ddot{\xi}(s)\, ds, \qquad t \in [0,1]. \tag{44}$$

At fixed t the mass of the kernel is

$$\int_0^1 G(t,s)\, ds = (1-t)\frac{t^2}{2} + t\frac{(1-t)^2}{2} = \frac{t(1-t)}{2} \le \frac{1}{8}, \tag{45}$$

the maximum being attained at $t = \frac{1}{2}$. Probing equation 44 against a unit direction y, the triangle inequality for integrals and Cauchy–Schwarz pointwise in s give

$$\begin{aligned} \big|\langle y, \xi(t)\rangle\big| &\le \int_0^1 G(t,s)\, \big|\langle y, \ddot{\xi}(s)\rangle\big|\, ds \\ &\le \Big(\int_0^1 G(t,s)\, ds\Big) \sup_{s\in[0,1]} \|\ddot{\xi}(s)\|_2 \\ &\le \frac{1}{8} \sup_{s\in[0,1]} \|\ddot{\xi}(s)\|_2, \end{aligned} \tag{46}$$

the last line by equation 45; taking the supremum over unit y, the dual characterisation of the Euclidean norm, and then over $t \in [0,1]$ gives

$$\sup_{t\in[0,1]} \|\xi(t)\|_2 \le \frac{1}{8} \sup_{t\in[0,1]} \|\ddot{\xi}(t)\|_2. \tag{47}$$

The constant is sharp: $\xi(t) = t(1-t)e$ for any unit vector e attains $\sup_t \|\xi(t)\|_2 / \sup_t \|\ddot{\xi}(t)\|_2 = (1/4)/2 = 1/8$.

Acceleration used in Lemma 4.5. We next formalize the consequences of the preceding Green-kernel bound within the context of the Fisher geodesic. Let $\xi := \gamma_g - \gamma_{\mathrm{lin}}$. Since $\ddot{\gamma}_{\mathrm{lin}} \equiv 0$ and γ_g satisfies the geodesic equation,

$$\ddot{\xi}(t) = \ddot{\gamma}_g(t) = -\Gamma_{\gamma_g(t)}(\dot{\gamma}_g(t), \dot{\gamma}_g(t)).$$

Write $\kappa_n := K_\Gamma^{\mathrm{op}} n^{-\alpha(\sigma)} = 3c_1/(2\lambda n^{\alpha(\sigma)})$. The Christoffel estimate of Theorem 4.4 gives

$$\|\ddot{\xi}(t)\|_2 \le \kappa_n \|\dot{\gamma}_g(t)\|_2^2.$$

Under the speed bound supplied by the Green-operator fixed-point argument, $\sup_t \|\dot{\gamma}_g(t)\|_2 \le 2\|\Delta\|_2$, this yields

$$\sup_{t\in[0,1]} \|\ddot{\xi}(t)\|_2 \le 4\kappa_n \|\Delta\|_2^2 = \frac{6c_1}{\lambda n^{\alpha(\sigma)}} \|\Delta\|_2^2. \tag{48}$$

Applying Lemma 4.5 to ξ, which satisfies $\xi(0) = \xi(1) = 0$, gives

$$\sup_{t\in[0,1]} \|\xi(t)\|_2 \le \frac{1}{8} \sup_t \|\ddot{\xi}(t)\|_2 \le \frac{1}{2}\kappa_n \|\Delta\|_2^2 = \frac{3c_1}{4\lambda n^{\alpha(\sigma)}} \|\Delta\|_2^2. \tag{49}$$

Thus the certified deviation between the Fisher geodesic and the Euclidean interpolation is $O(n^{-\alpha(\sigma)}\|\Delta\|_2^2)$, hence $O(n^{2s-\alpha(\sigma)})$ whenever $\|\Delta\|_2 = \Theta(n^s)$. The estimate decays only for $\alpha(\sigma) > 2s$, and s is measured per regime in Appendix D; it is close to 1 in the lazy regime, where the estimate grows, and well below $1/2$ under the two feature-learning parameterizations. □

C Assumption 4.2 in the NTK Regime: the Frozen-Softmax Fisher

Statement. *Consider an MLP with $L \ge 2$ hidden layers of uniform width n, activation $\sigma \in C^3$ with $\|\sigma^{(j)}\|_\infty \le M_\sigma$ for $j = 1, 2, 3$, under NTK parameterization: weights initialised as $w_{ij} \sim \mathcal{N}(0,1)$ with each forward pass rescaled by $1/\sqrt{n_\ell}$. Write $w(t)$, $t \in [0,T]$, for the training trajectory, and define the frozen Gauss–Newton surrogate*

$$\widetilde{F}(w) := \mathbb{E}_x\big[J(x;w)^\top S_0(x)\, J(x;w)\big], \qquad S_0(x) := S(p_{w(0)}(x)), \tag{50}$$

in which the softmax factor is held at its initial value and only the Jacobian moves. The derivation is conditional on the following stability bounds, imported from the NTK literature rather than re-proved: for every $\delta > 0$ there are constants C_J, C_H, C_T and n_0 such that for $n \ge n_0$, on an event $\mathcal{E}_n$ with $\mathbb{P}(\mathcal{E}_n) \ge 1 - \delta$, uniformly over $x \in \mathrm{supp}(\mathcal{D}_X)$, over $t \in [0,T]$ and over a neighbourhood U of the trajectory,

$$\|J(x;w)\|_{\mathrm{op}} \le C_J, \qquad \max_k \|H_k(x;w)\|_{\mathrm{op}} \le \frac{C_H}{\sqrt{n}}, \qquad \sup_{\|h\|_2=\|z\|_2=1} \|\partial_h \partial_z J(x;w)\|_{\mathrm{op}} \le \frac{\sqrt{K}\, C_T}{\sqrt{n}}. \tag{51}$$

Then on $\mathcal{E}_n$ the surrogate satisfies Assumption 4.2 with $\alpha = \frac{1}{2}$,

$$\sup_t \big\|\partial_z \widetilde{F}(w(t))\big\|_{\mathrm{op}} \le \frac{c_1}{\sqrt{n}}, \qquad \sup_t \big\|\partial_h \partial_z \widetilde{F}(w(t))\big\|_{\mathrm{op}} \le \frac{c_2}{\sqrt{n}}, \tag{52}$$

with $c_1 \le \sqrt{K}\, C_H C_J$ and $c_2 \le \sqrt{K}\, C_T C_J + K C_H^2$, while the true Fisher matrix obeys

$$\sup_{\|z\|_2=1} \big\|\partial_z (F - \widetilde{F})(w(t))\big\|_{\mathrm{op}} \le \frac{c_3}{\sqrt{n}} + 3C_J^3\, \rho_S(w(t)), \qquad c_3 \le 2\sqrt{K}\, C_H C_J, \tag{53}$$

where $\rho_S(w) := \mathbb{E}_x \|S(p_w(x))\|_{\mathrm{op}} \in [0, \frac{1}{2}]$. The term $3C_J^3 \rho_S$ does not contract with width; it contracts only with predictive uncertainty.

Proof. Row k of $\partial_z J$ is $(H_k z)^\top$, since $(\partial_z J)_{ki} = \sum_j z_j\, \partial_j \partial_i f_k = (H_k z)_i$; passing through the Frobenius norm and applying equation 51,

$$\|\partial_z J\|_{\mathrm{op}} \le \Big(\sum_{k=1}^K \|H_k z\|_2^2\Big)^{1/2} \le \sqrt{K}\, \max_k \|H_k\|_{\mathrm{op}}\, \|z\|_2 \le \frac{\sqrt{K}\, C_H}{\sqrt{n}}\, \|z\|_2, \tag{54}$$

the factor $\sqrt{K}$ being in the number of classes, not of parameters, and therefore harmless. Since S_0 is independent of w, the product rule, interchanged with $\mathbb{E}_x$ by dominated convergence on the compact support, gives $\partial_z \widetilde{F} = \mathbb{E}_x[(\partial_z J)^\top S_0 J + J^\top S_0(\partial_z J)]$; each summand is at most $\|\partial_z J\|_{\mathrm{op}}\|S_0\|_{\mathrm{op}}\|J\|_{\mathrm{op}}$, and $\|S(p)\|_{\mathrm{op}} \le \frac{1}{2}$ for every p, because for unit a $a^\top S(p)a = \mathrm{Var}_{k\sim p}(a_k) \le \frac{1}{4}(\max_j a_j - \min_j a_j)^2 \le \frac{1}{2}$ by Popoviciu's inequality (Popoviciu, 1935). Inserting equation 54 and equation 51,

$$\big\|\partial_z \widetilde{F}\big\|_{\mathrm{op}} \le 2 \cdot \frac{\sqrt{K}\, C_H}{\sqrt{n}} \cdot \frac{1}{2} \cdot C_J = \frac{\sqrt{K}\, C_H C_J}{\sqrt{n}} \qquad (\|z\|_2 = 1). \tag{55}$$

Differentiating once more produces two second-derivative terms, bounded together via equation 51 by $\sqrt{K}\, C_T C_J/\sqrt{n}$, and two cross terms of the form $(\partial J)^\top S_0(\partial J)$, bounded by squaring equation 54 as KC_H^2/n in total, one order smaller; adding the four gives the second half of equation 52 with $c_2 \le \sqrt{K}\, C_T C_J + KC_H^2$. Both bounds are uniform in t because equation 51 is. On the same event Assumption 4.3 holds with $M_F = C_J^2/2$, since $\|F(w)\|_{\mathrm{op}} \le \mathbb{E}_x[\|J\|_{\mathrm{op}}^2 \|S(p_w(x))\|_{\mathrm{op}}] \le C_J^2/2$.

For the true Fisher, write $F - \widetilde{F} = \mathbb{E}_x[J^\top(S(p_w) - S_0)J]$. Now $S(p_w)$ does depend on w, and the product rule produces three groups,

$$\partial_z(F - \widetilde{F}) = \mathbb{E}_x\big[(\partial_z J)^\top (S - S_0)J + J^\top(S - S_0)(\partial_z J) + J^\top(\partial_z S(p_w))J\big]. \tag{56}$$

For the first two, $\|S - S_0\|_{\mathrm{op}} \le \|S\|_{\mathrm{op}} + \|S_0\|_{\mathrm{op}} \le \frac{1}{2} + \frac{1}{2} = 1$, so each group is at most $\sqrt{K}\, C_H C_J/\sqrt{n}$ by equation 54 and equation 51; together they contribute $c_3/\sqrt{n}$ with $c_3 \le 2\sqrt{K}\, C_H C_J$. The third group is the one that matters. By the chain rule and the softmax differential, $\partial_z p = S(p)\,\partial_z f = S(p)Jz$, so

$$\|\partial_z p\|_2 \le \|S(p)\|_{\mathrm{op}}\, \|J\|_{\mathrm{op}}\, \|z\|_2 \le \|S(p)\|_{\mathrm{op}}\, C_J \qquad (\|z\|_2 = 1). \tag{57}$$

Differentiating $S(p) = \mathrm{diag}(p) - pp^\top$ in direction z gives

$$\partial_z S = \mathrm{diag}(\partial_z p) - (\partial_z p)p^\top - p(\partial_z p)^\top, \tag{58}$$

whose three blocks are bounded separately: $\|\mathrm{diag}(a)\|_{\mathrm{op}} = \|a\|_\infty \le \|a\|_2$; $\|ab^\top\|_{\mathrm{op}} = \|a\|_2\|b\|_2$, because $ab^\top v = (b^\top v)a$ and $|b^\top v|$ is maximised at $v = b/\|b\|_2$; and $\|p\|_2 \le \|p\|_1 = 1$. Summing and inserting equation 57,

$$\|\partial_z S\|_{\mathrm{op}} \le \|\partial_z p\|_2\,(1 + 2\|p\|_2) \le 3\,\|\partial_z p\|_2 \le 3\,\|S(p)\|_{\mathrm{op}}\, C_J, \tag{59}$$

therefore $\|J^\top(\partial_z S)J\|_{\mathrm{op}} \le C_J^2 \cdot 3\|S(p_w(x))\|_{\mathrm{op}} C_J = 3C_J^3\|S(p_w(x))\|_{\mathrm{op}}$, and averaging over x — bounding the norm of an expectation by the expectation of the norm — gives $3C_J^3\rho_S(w)$. Adding the three groups yields equation 53; adding equation 52 to it by the triangle inequality, the true Fisher satisfies $\sup_{\|z\|_2=1}\|\partial_z F(w(t))\|_{\mathrm{op}} \le c_1'/\sqrt{n} + 3C_J^3\rho_S(w(t))$ with $c_1' \le 3\sqrt{K}\, C_H C_J$, i.e. Assumption 4.2 holds for F at the lazy rate up to the transport term. This completes the proof. □

D Experiments Setup

This section describes the experimental setup for training the networks whose minima are interpolated, and for measuring the Fisher-side quantities reported in Section 5. All models have $K = 10$ outputs, uniform hidden width n, no dropout, and $P = \Theta(n^2)$. Five activations are trained, $\{\mathrm{ReLU}, \mathrm{GELU}, \tanh, \mathrm{Swish}, \mathrm{Softplus}\}$; every reported quantity uses only the four smooth ones, since F and $\partial_z F$ require a C^3 activation that the ReLU kink denies. The measured grid is therefore three architectures × three parameterizations × four smooth activations, each swept over the widths of its architecture; *cell* means that triple throughout, and never a single width within one.

D.1 Model Training

MLP on MNIST. We train a plain fully connected network $784 \to n \to n \to 10$ on flattened 28×28 images: two hidden layers of n units each, no skip connections and no convolution, with n swept over $\{64, 128, \ldots, 4096\}$. Every linear layer carries a bias, initialized at zero and left outside

the per-layer forward rescaling, and the readout is linear. The hidden layers use each of the four smooth activations in turn, and no normalization or dropout is applied, so that F and $\nabla^2 L$ are free of gauge artifacts. We optimize with SGD at momentum 0.9 and a base learning rate of 10^{-1}, under a linear warmup of $\min(8, \lfloor \text{epochs}/5 \rfloor)$ epochs followed by cosine decay, with no weight decay and gradient-norm clipping at 1.0. The batch size and number of epochs are set to 256 and 30, respectively.

Teacher–student. The student reuses the architecture, width grid and activations of the MLP, but is trained on synthetic data rather than MNIST: inputs are $d = 64$ standard Gaussian vectors, and targets come from a frozen width-32 ReLU teacher drawn once and held fixed across every seed, width and parameterization. Each student is therefore over-realised relative to its teacher, which isolates the effect of width from any change in the target function. The optimizer, learning-rate schedule and batch size are those of the MLP; the number of epochs is 100.

CNN on FashionMNIST. A small all-convolutional classifier, not a VGG- or ResNet-style network: three bias-free 3×3 convolutions at stride 1 and padding 1, with channels $(16w_m, 32w_m, 64w_m)$ for a multiplier $w_m \in \{1, 2, 4, 8\}$, and no residual connections. Each convolution is followed by a single-group $\mathrm{GroupNorm}(1, c)$ — a LayerNorm over channels, chosen because it is equivariant under channel permutation and so survives weight matching — and then the activation. A 2×2 max-pool follows the first two blocks only; the third feeds global average pooling to 1×1 and a single linear layer to the 10 classes, so there is no fully connected head beyond that readout. No architecture in the suite uses batch normalization, so no activation-statistics repair (Jordan et al., 2023) is applied at any width. We plot this family against its widest layer $n = 64w_m \in \{64, \dots, 512\}$, which puts it on the same width axis as the other two. The four smooth activations are used as above, and the optimizer, learning-rate schedule and batch size are those of the MLP; the number of epochs is 100.

Parameterisations. Every architecture and activation is trained in three parameterizations: *NTK-lazy* ($O(1)$ initialization variance with a $1/\sqrt{n}$ forward rescaling per layer), *Standard* (fan-in initialization), and *μP* (per-tensor initialization and optimizer scaling transferred from an anchor width 64, Yang & Hu, 2021). The base learning rate above is not retuned per width, so width scaling is not confounded with hyperparameter adaptation.

D.2 Numerical methodology.

Fisher estimators. Neither the Jacobian $J(x; w)$ nor $\widehat{F}$ is ever formed: all Fisher-side quantities are built from matrix-free Fisher–vector products, evaluated at $O(P)$ memory on a fixed batch of 2048 examples. The Fisher length $\mathcal{L}_F$, the Rayleigh quotient $\mathcal{R}_F$ and the Christoffel loads follow from this one primitive, with $\mathcal{L}_F$ using the unregularised F that appears in Theorem 4.1. Every Fisher-side quantity we report is read at the trained endpoint w_A, where the gradient term of that theorem vanishes: the Rayleigh quotient of Figure 3 is $\widehat{\Delta}^\top F(w_A)\widehat{\Delta}$ and the Fisher length of Figures 4 and 5 is $\Delta^\top F(w_A)\Delta$. Midpoint values are measured on the same grid but are not what the bound is anchored on, and we do not mix the two. Solves with $G_F = \widehat{F} + \lambda I$ use conjugate gradients, and the damping is set relative to the spectrum, $\lambda = 10^{-2}\lambda_{\max}(F)$, so that G_F stays comparably conditioned across regimes whose Fisher scales differ by orders of magnitude. We form no Hessian–vector products: the deviation of $\nabla^2 L$ from F enters only through the fit residual $\widehat{\rho}^*(w)$ of Theorem 3.4, which needs one forward pass.

Metric derivatives and operator norms. The directional derivative $\partial_z F$ is taken by central differences at $\varepsilon = 3 \times 10^{-3}$ with Richardson extrapolation, and every evaluation carries a finite-difference stability diagnostic; a small number of CNN measurements were withdrawn on it, and since they are not spread uniformly over the grid, the CNN exponents of App. E rest on slightly thinner coverage than the other two families. Operator norms are computed by power iteration rather than by random probing, which would estimate a normalized trace instead of a supremum over directions and, because the adversarial direction for ∂F lies in a subspace far smaller than P, would report a flattening exponent that is too large.

Barrier and alignment. At every width of every cell we train 5 seeds and measure all $\binom{5}{2} = 10$ pairs, which is 2160 measured pairs in all; losses, barriers and ρ^* are evaluated on a fixed subset of 10,000 examples, identical across widths, regimes and seeds. $L(t)$ is evaluated on a uniform grid of $T = 41$ points, a density we verified rather than assumed: refining to a grid ten times finer that contains the reported one as a subset moves the barrier by less than 0.4% on every pair tested, and every pair satisfies the endpoint floor $B(\gamma_{\text{lin}}) \geq \frac{1}{2}|L(w_A) - L(w_B)|$, which every one of the 2160 pairs satisfies. Before interpolating, a function-preserving permutation is applied to one endpoint by layerwise weight matching (Entezari et al., 2022; Ainsworth et al., 2023), as post-processing only and never during training.

Geodesic residual. The shape diagnostic $D_{\text{rel}} = \sup_t \|\xi_G(t)\|_2 / \|\Delta\|_2$ is the first-order deviation of Section 4, reconstructed from the Christoffel load along the segment through the Green kernel on a quadrature grid coarser than the barrier grid, since each load costs a linear solve where a barrier point costs a forward pass. Two limits should be stated. It is the residual of the *linear* path, not a certified distance to an exact geodesic, because the smallness condition an iteration to a true geodesic requires fails at the displacement scale of the lazy regime. And it is normalized by $\|\Delta\|_2$, which grows with width in every cell of the grid but at rates differing by a factor of 23 between the slowest and the fastest — it tracks $\Theta(\sqrt{P})$ under NTK-lazy and grows far more slowly elsewhere — so it compares across widths within a parameterization and not across parameterizations. Fitting $\|\Delta\|_2 \propto n^s$ on the per-width medians of each cell gives a median s of 0.94 under NTK-lazy, 0.26 under Standard and 0.39 under μP, with no cell of the grid returning a negative s; these are the exponents Section 4 reads the certified estimate against. Every curve in Figure 2 is therefore normalized by its own smallest-width value and read for its trend alone, never for the vertical ordering between regimes.

Reporting. Christoffel, curvature and geodesic-residual magnitudes carry negative powers of λ, which is itself set relative to $\lambda_{\max}(F)$ and so varies with width, and Fisher magnitudes carry the parameterization's scale. Absolute values are therefore not comparable across widths or regimes, and we report only width exponents from log–log fits, normalized ratios, and rank correlations. Each width carries the four smooth activations and all ten seed pairs, forty measurements in all; a curve is their median, the band their interquartile range, and every curve is normalized by its own smallest-width value. A width exponent is the OLS slope of the log median on $\log n$, quoted as a decay rate $n^{-\alpha}$, and we give its R^2 wherever the fit is not a clean power law: in Figure 3 the Standard MLP saturates from $n = 1024$ and reaches only $R^2 = 0.91$, against 0.99 for NTK-lazy and μP, so its exponent describes the first half of the grid rather than the whole of it. The one resampling in the paper is the 95% bootstrap band on the cross-cell regression line of Figure 5.

E Additional Results

Every quantity is summarised by its width exponent, written as a decay rate $Q(n) \sim n^{-\alpha}$, so $\alpha > 0$ means it vanishes with width (App. D).

E.1 The Central Contrast on the Other Two Architectures

In this subsection, Figure 6 shows the interpolation barrier and $\|\partial F\|_{\text{op}}$ against width for the teacher–student family and the CNN, the two architectures Figure 1 omits. $\|\partial F\|_{\text{op}}$ decays as a power law in all six cells, whereas the barrier collapses under Standard and μP and barely moves under NTK-lazy ($\alpha_B = -0.20$ for the teacher–student and $+0.15$ for the CNN). Both observations repeat Figure 1 on the MLP, so the contrast is not a feature of one architecture. They separate because flattening is a Shape-axis property that Theorem 4.4 grants in every regime, whereas collapse additionally requires the Length-axis control of Theorem 4.1. A flattening metric is therefore not evidence for linear mode connectivity in any of the three architectures. The central contrast of Section 5 is thus a property of the parameterization rather than of the MLP. The one MLP-only feature is the non-monotone Standard curve, absent from both rows here, so we read that reversal as belonging to that cell and not to the regime.

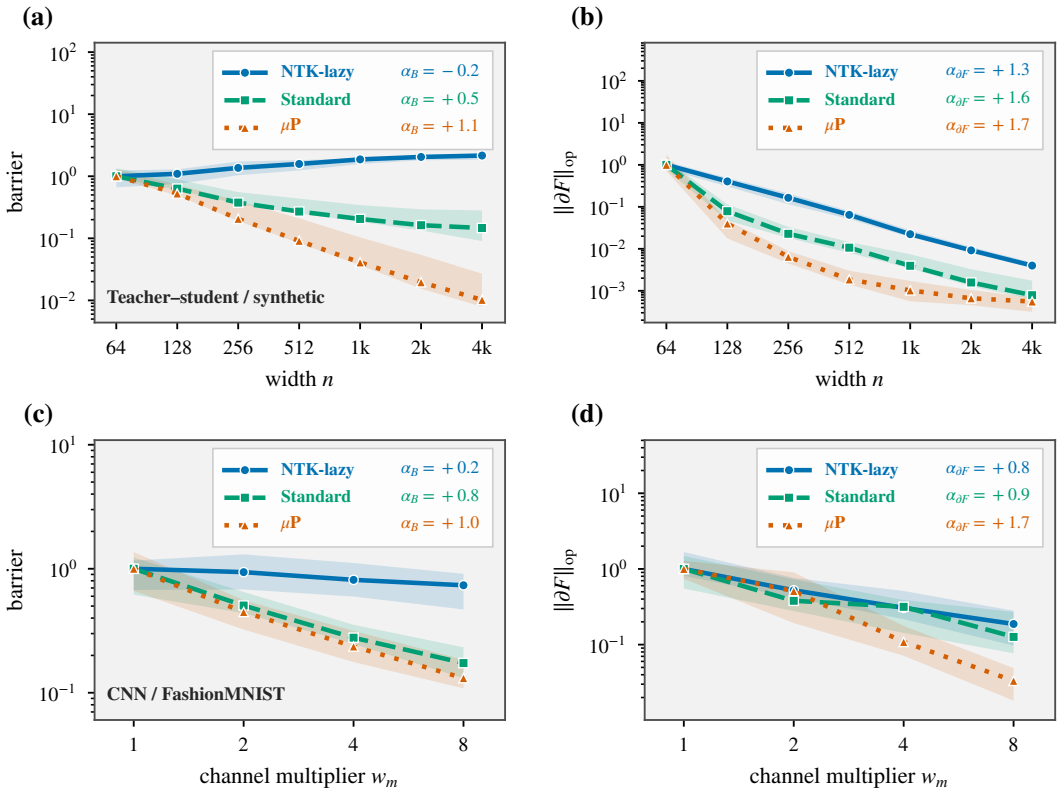

Figure 6: Interpolation barrier and metric-derivative norm $\|\partial F\|_{\mathrm{op}}$ against width for the teacher–student family (a, b) and the CNN (c, d), the two architectures Figure 1 omits.

E.2 Spectral Alignment on the Other Two Architectures

In this subsection, Figure 7 shows the Rayleigh quotient $\mathcal{R}_F$ against width for the teacher–student family and the CNN, the two architectures Figure 3 omits. $\mathcal{R}_F$ falls as a power law in all six cells, between $n^{-1.44}$ and $n^{-2.19}$. This repeats the MLP decay of Figure 3, so the concentration of Δ in the low-eigenvalue subspace of F holds across architectures. The decay is a Shape-axis consequence of the flattening of Theorem 4.4, which acts in every regime and therefore cannot separate them. Accordingly the two orderings disagree: the regime in which $\mathcal{R}_F$ falls most slowly is Standard on the teacher–student family ($n^{-1.44}$) and μP on the CNN ($n^{-1.96}$), in neither case the NTK-lazy cell in which the barrier collapses most slowly ($\alpha_B = -0.20$ and $+0.15$, against $+0.47$ and $+0.84$ under Standard). The negative control of Section 5 therefore holds on all three architectures, and the full Fisher length remains the quantity that predicts the barrier. The saturation that weakens the Standard MLP fit to $R^2 = 0.91$ has no counterpart in the Standard cells here, which reach $R^2 = 0.94$ and $R^2 = 0.996$; where a fit is weaker on the teacher–student family it is the μP one ($R^2 = 0.92$), so the saturation too belongs to that cell and not to the regime.

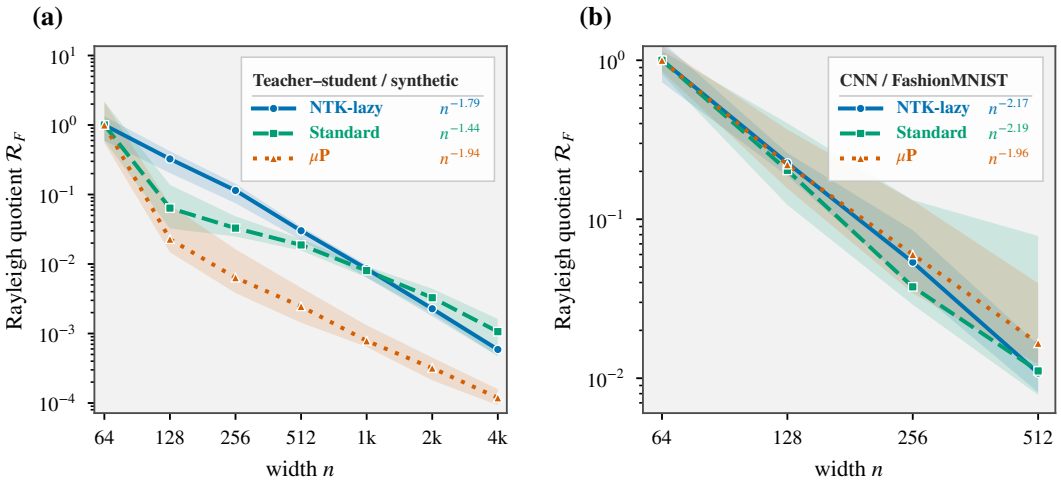

Figure 7: Rayleigh quotient $\mathcal{R}_F$ against width for the teacher–student family (a) and the CNN (b), the two architectures Figure 3 omits.

E.3 The Two Conditions of the Summary Claim

In this subsection, Figures 8–10 show the two conditions of the summary claim — a reduction of the Fisher length and a small endpoint predictive uncertainty ρ^* — over the full experimental grid.

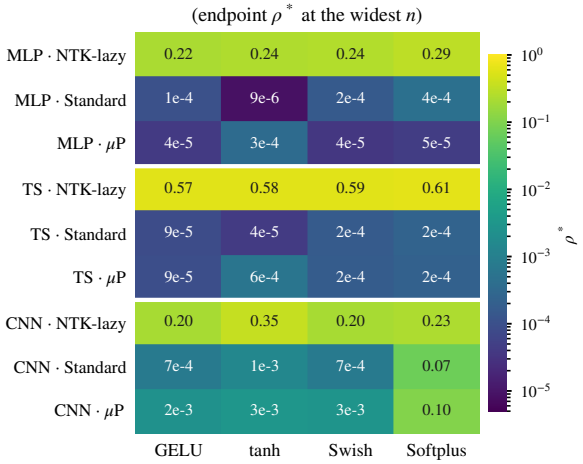

Figure 8: Endpoint predictive uncertainty $\rho^*(w_A)$ at the widest width of every cell of the full experimental grid, value printed in every tile. Colour is a log magnitude, on the scale Figure 10 reuses.

Figure 8 shows the endpoint uncertainty at the widest width of every cell of the grid, and the values sort by parameterization alone into a lazy band and a feature-learning band, whatever the architecture or the activation: every NTK-lazy cell sits above every Standard and μP cell of its architecture, while Standard and μP overlap with each other. Uncertainty is therefore a property of the regime rather than of anything else the grid varies, which makes the uncertainty condition of the summary claim a consequence of how the model is parameterized rather than of how large it is.

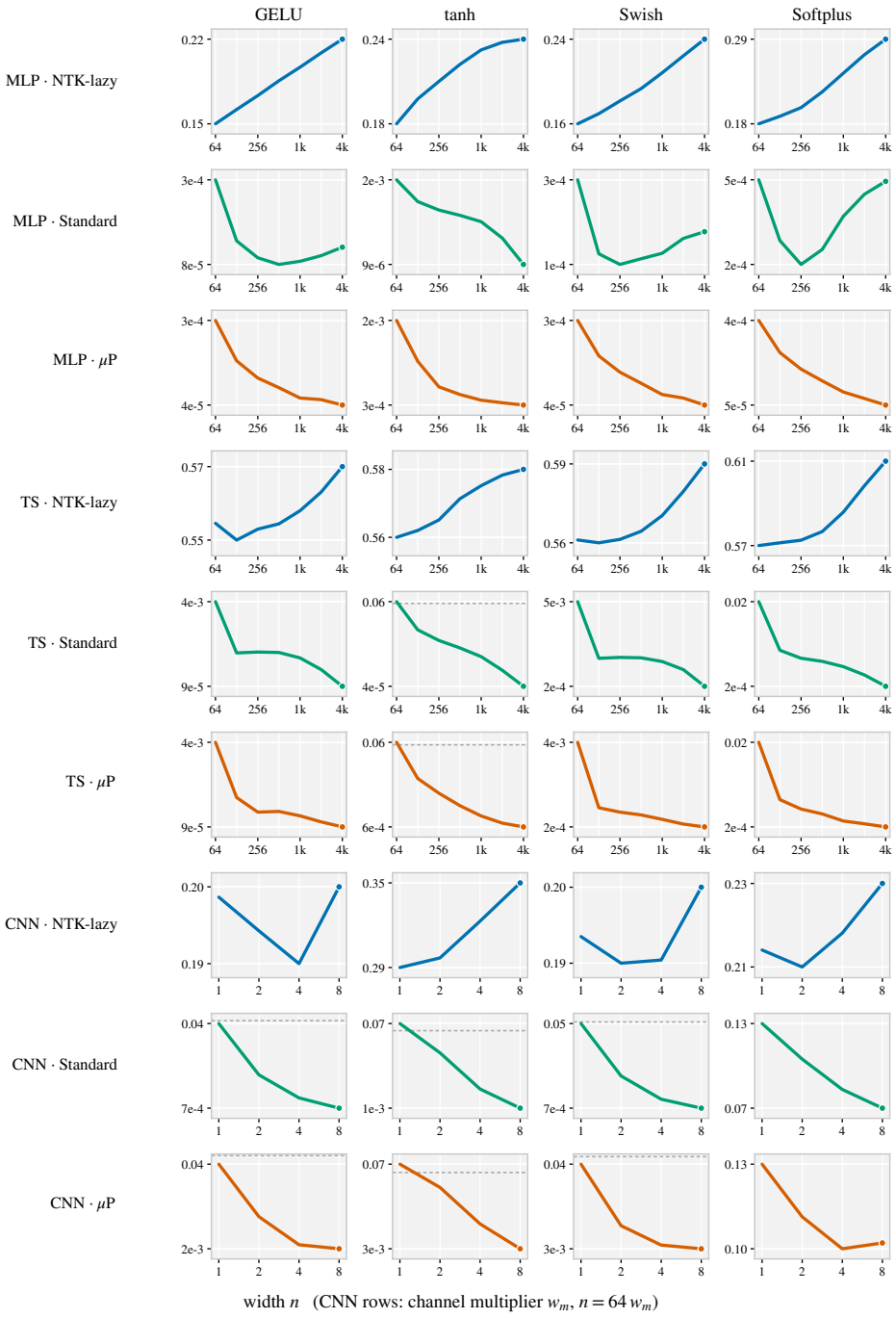

Figure 9: Endpoint uncertainty $\rho^*(w_A)$ against width, for every cell of the full experimental grid.

Figure 9 traces, one panel per cell, the width trajectory of the endpoint uncertainty behind the values of Figure 8: every NTK-lazy cell rises, every μP cell falls, and the Standard MLP alone does neither, dipping to a minimum at intermediate width and climbing back. In every panel the trajectories keep their regime's ordering from the narrowest width to the widest, the lazy ones staying above the feature-learning ones with no crossing, so each cell remains inside its band throughout. Width therefore moves a cell within its band and never between bands, which makes the uncertainty condition a regime-level prerequisite rather than something width can supply or remove.

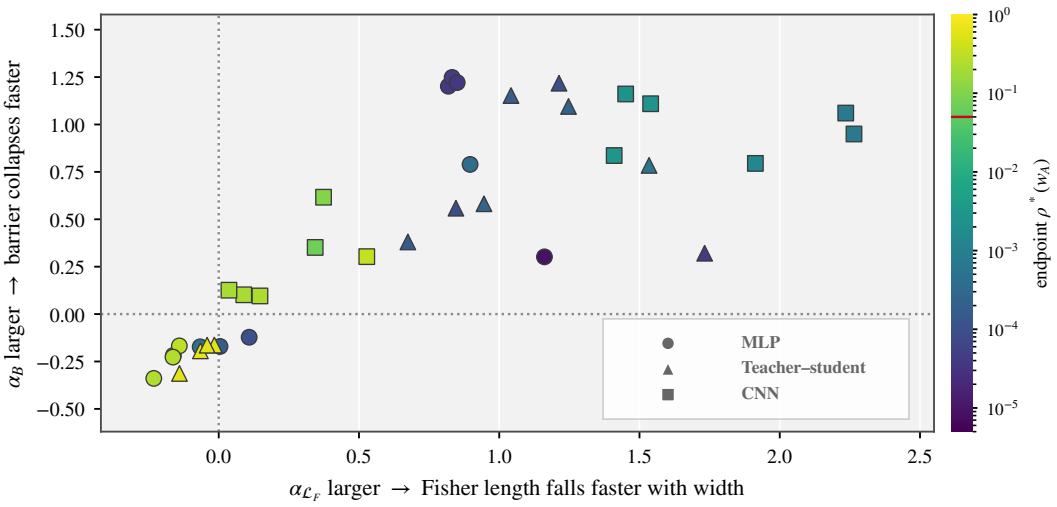

Figure 10: The two conditions of the summary claim against the barrier exponent: the endpoint Fisher-length exponent on the horizontal axis and the endpoint uncertainty ρ^* as the colour.

Figure 10 shows that two conditions of the summary claim act asymmetrically: cells that reduce their length while carrying a large ρ^* still collapse, only more weakly than the cells that meet both conditions, whereas cells that keep their uncertainty controlled without reducing their length never collapse. Length reduction therefore decides whether a cell collapses, and uncertainty only how fast. High uncertainty never coincides with fast length reduction on this grid, so it cannot be what drives the trend. The coupling behind the trend is what the length condition of the summary claim rests on, and it persists within each regime and each architecture taken alone.

E.4 The Method on a Harder Dataset

In this subsection, Figures 11–12 show the paper's width measurements and its Figure 5 prediction repeated on CIFAR-10. The architecture, the three parameterizations, the four smooth activations and the channel multipliers $w_m \in \{1, 2, 4, 8\}$ are all held at the setting of the released CNN run, so the dataset is the only thing that changes.

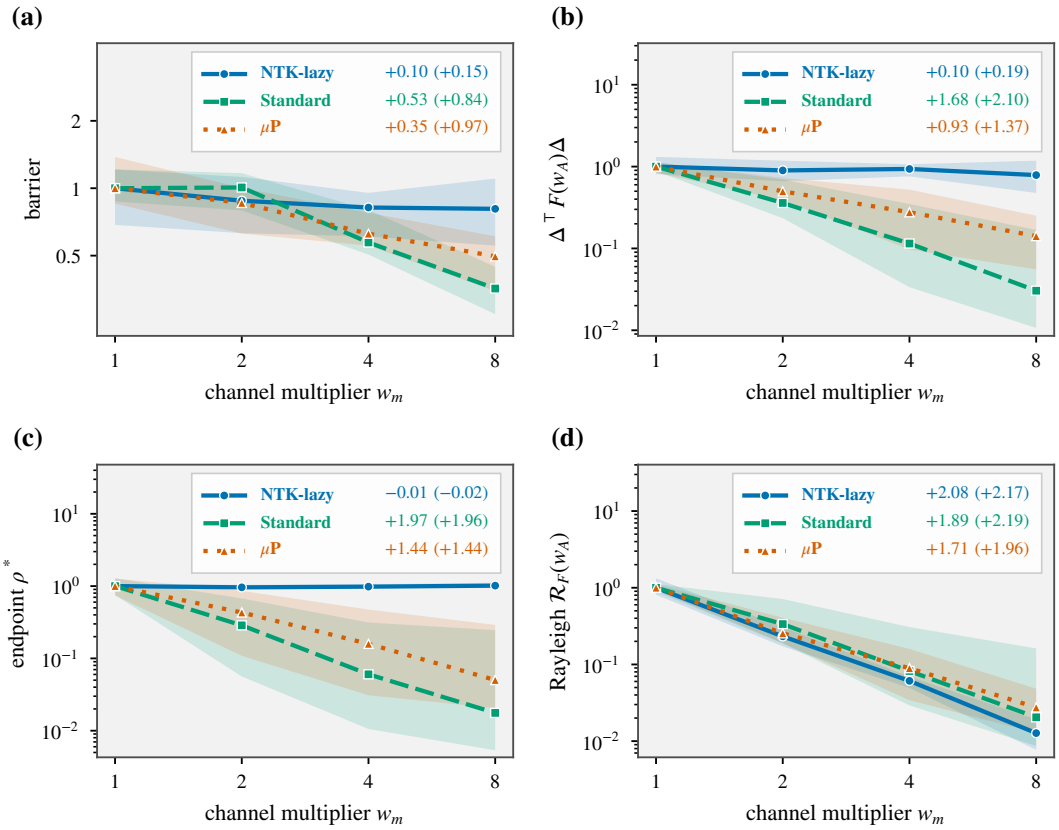

Figure 11: The width measurements of Section 5 repeated on CIFAR-10 for the CNN: barrier (a), Fisher length (b), endpoint uncertainty ρ^* (c) and Rayleigh quotient $\mathcal{R}_F$ (d), each key giving the CIFAR-10 exponent with the released FashionMNIST CNN's in brackets.

Figure 11 carries this paper's width-scaling measurements beyond the small-scale settings of the main text, to CIFAR-10 with every other setting left at the released CNN run; taken unchanged and without retuning, the measurements return the same picture, each quantity behaving on CIFAR-10 as it does on the released dataset. No panel departs from that picture, so the account is not a feature of convenient toy settings but extends, unchanged, into a genuinely harder problem. The relations the measurements expose are therefore properties of the geometry rather than of the dataset.

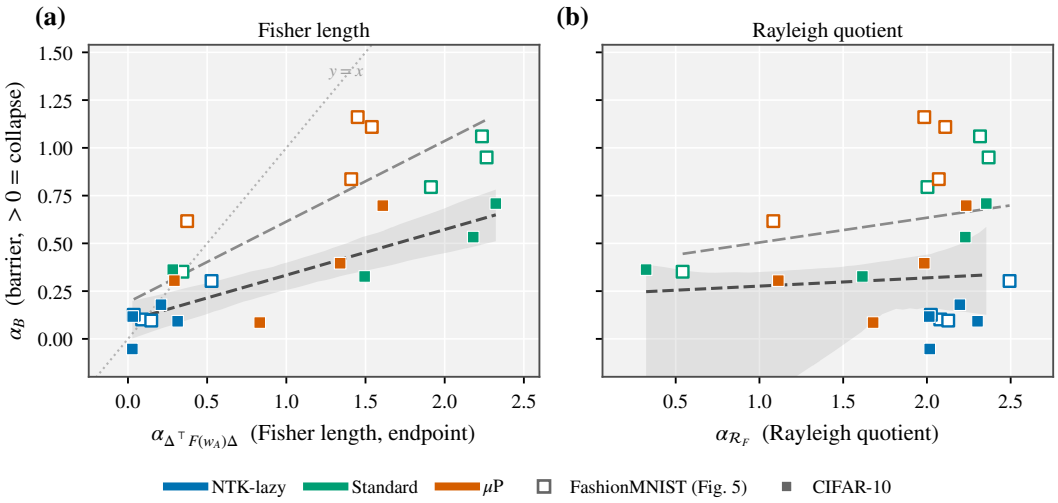

Figure 12: Barrier decay exponents against Fisher-length (panel a) and Rayleigh-quotient (panel b) exponents for the CIFAR-10 cells (solid) against the matched released FashionMNIST CNN cells (hollow), with the two fits and a bootstrap band on the CIFAR-10 fit.

Figure 12 shows that on the CIFAR-10 cells the coupling between barrier and Fisher length is as tight as on the matched released cells, while the Rayleigh quotient remains as powerless on the harder dataset as on the released one. The Fisher length, not directional spectral alignment, therefore governs barrier decay once the convenient toy settings are left behind.

E.5 The Geodesic Deviation on the Other Two Architectures

In this subsection, Figure 13 extends Figure 2 to the remaining two architectures, showing the relative geodesic–linear deviation D_{rel} against width. The first panel mirrors Figure 2 regime by regime: D_{rel} decays as a power law under NTK-lazy and μP, and is non-monotonic under Standard. This indicates that the path-level behavior of Section 5 is not specific to the MLP.

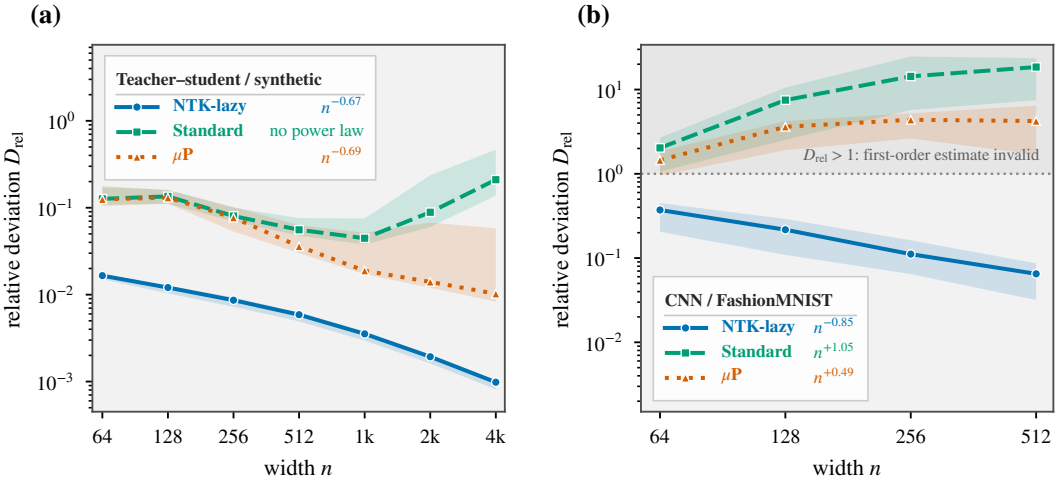

Figure 13: Relative geodesic–linear deviation D_{rel} against width for the teacher–student family (a) and the CNN (b), the two architectures Figure 2 omits. Unlike that figure these panels are drawn at absolute scale, so that panel (b) can mark $D_{\mathrm{rel}} = 1$, above which the first-order estimate no longer holds.

Figure 13 shows whether the path-level picture of Figure 2 extends to the two architectures it omits. In the first panel it does: every regime keeps the behavior Figure 2 reports on the MLP, including the non-monotone Standard curve, so that reversal is not a property of one cell. In the second panel it does not: under the two feature-learning regimes D_{rel} rises with width and lies above the marked threshold at every width, above which the first-order estimate no longer holds and the curves are read qualitatively only. The rise remains informative through its denominator: the displacement between aligned endpoints barely grows with width under those regimes while the absolute deviation keeps growing, and the displacement is read directly off the endpoints without any linear solve. Therefore the increase reports a stalling displacement rather than a bending path, which is the direction Theorem 4.4 leaves open.