CANONICAL REGULARISATION OF WIDE FEATURE-LEARNING NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wide neural networks in the feature-learning regime drive modern deep learning, and yet they remain far less studied than their kernel-regime counterparts. We consider a critical yet under-explored difference between these two regimes: the regulariser and prior implied by gradient flow training. This *canonical regularisation* property is well-studied in kernel regime networks – of all the infinite global minima, gradient flow selects exactly the vanishing ridge solution – and underpins the celebrated NN-GP correspondence, precisely allowing the modelling of noise during training. However, we prove ridge regularisation biases gradient flow in feature-learning regime networks, even in the infinitesimal limit of vanishing regularisation. Over training, ridge distorts the inductive bias of the network, with a particular damage done to pretrained networks where the implicit prior is informative. We resolve this by deriving the canonical regulariser from the training dynamics alone: under a single conditioning assumption, the initialisations that train to the same network form a curved surface, the basin manifold, and the optimiser's own kinetic energy, measured over the motions that preserve the training outcome, uniquely identifies ridge in the kernel regime and generalises it to *geodesic ridge* in the feature-learning regime. As a practical contribution, we propose *arc ridge* as a minimax-robust, scalable surrogate to geodesic ridge, revealing a deep relationship between early stopping and canonical regularisation across learning regimes. Finally, we demonstrate the consequences of our theory empirically on both image processing and NLP transfer-learning problems.

1 INTRODUCTION

Wide neural networks now drive nearly every advance in modern deep learning, yet our theoretical understanding of these networks operating in the regime in which they actually *learn features* – rather than evolving as effectively fixed kernel machines – remains limited. The kernel regime, formalised by the Neural Tangent Kernel (NTK) of Jacot et al. (2018) and elaborated by Lee et al. (2019) and Du et al. (2019), supports a complete account of training dynamics, generalisation, and Bayesian inference; the feature-learning regime supports no comparable toolkit, yet inherits its practices by default. This paper confronts one such case, *which regulariser, and hence which implicit prior, gradient-flow training favours*, and identifies the geometric reason.

The kernel-regime picture is well understood. An overparametrised model has an entire continuum of zero-loss interpolators, yet gradient flow selects a single, well-defined point from this continuum (Lee et al., 2019, eq. (8)): the unique vanishing-regularisation limit of *anchored ridge* (Pearce et al., 2020), $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$, the squared parameter-space distance from initialisation (He et al., 2020; Calvo-Ordoñez et al., 2026, eq. (36), Thm. 4.3). Anchored ridge equals the squared reproducing kernel Hilbert space (RKHS) norm of the change in network output under the NTK over training, so the same penalty governs both parameter and function space, and the corresponding implicit prior over functions is a Gaussian process (GP) with the NTK as covariance, the NTK–GP correspondence (He et al., 2020). From this single equivalence flow the analytical conveniences of the kernel regime: principled treatment of label noise, calibrated uncertainty, predictable transfer behaviour. Anchored ridge is, in this sense, the *canonical* regulariser of the kernel regime. Standard ridge $\|\boldsymbol{\theta}\|^2$ (Hoerl & Kennard, 1970; Tikhonov, 1963) recovers the same selection only when $\boldsymbol{\theta}_0 \in \operatorname{Im} J^\top$, which a

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.

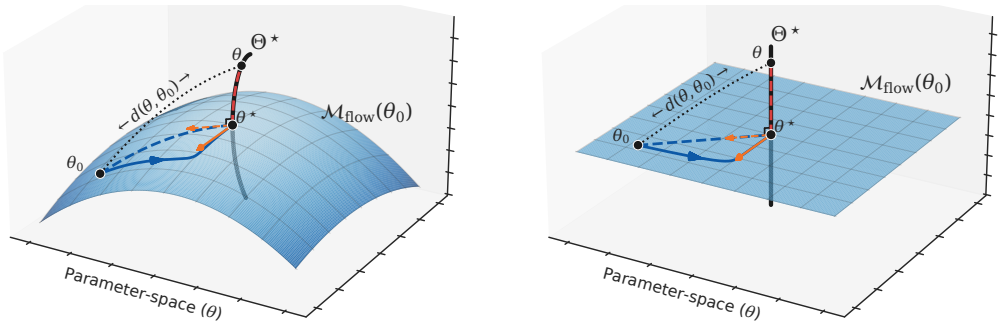


Figure 1: *Left*: feature-learning regime. The flow manifold $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (blue surface), the initialisations that train to $\boldsymbol{\theta}^\star$, curves near $\boldsymbol{\theta}^\star$; $\Theta^\star$ (black curve) is the interpolating set. The gradient flow trajectory (solid blue) lies on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$; the minimising geodesic from $\boldsymbol{\theta}_0$ (dashed blue) realises the distance $d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star)$ that defines geodesic ridge. In orange at $\boldsymbol{\theta}^\star$: the intrinsic gradient of geodesic ridge (dashed), along the terminal tangent of the geodesic, and the arc-ridge force (solid), along the realised flow direction; both are tangent to $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ and, at $\boldsymbol{\theta}^\star$, orthogonal to the output fibre, so neither biases gradient flow as standard and anchored ridge do. The surface depends on the labels through $\boldsymbol{\theta}^\star$. *Right*: kernel regime. The flow manifold is the affine plane $\boldsymbol{\theta}_0 + \operatorname{Im} J^\top$, geodesics are chords, and the canonical regulariser is anchored ridge.

random initialisation fails almost surely; otherwise it already biases the gradient-flow limit *within* the kernel regime (Remark D.11, Figure 2, right).

Outside the kernel regime, the same picture breaks down. Yang & Hu (2021) establish a sharp dynamical dichotomy: in the regime in which networks learn features and transfer between tasks, the Jacobian evolves throughout training, so trajectories leave any single affine plane and the function-space identity behind anchored ridge breaks. In this regime, formalised by Tensor Programs (Yang & Hu, 2021; Yang et al., 2021) and mean-field analyses (Chizat & Bach, 2018; Mei et al., 2018), no analogous canonical regulariser has been identified, nor a geometry in which to seek one (related work is collected in Appendix B). Practitioners apply weight decay (Krogh & Hertz, 1992) or anchored ridge, as in L2-SP for fine-tuning (Li et al., 2018), by default, even though the kernel-regime arguments justifying these choices no longer apply. The stakes are not abstract: weight decay is implicated in grokking (Power et al., 2022; Xu et al., 2026), and fine-tuning is prone to distorting pretrained feature representations (Kumar et al., 2022).

We argue that the gap is geometric. Behind every overparametrised network lie two parameter-space objects (Section 3.1): the *flow manifold*, the n-dimensional surface of initialisations from which training produces the same network as from $\boldsymbol{\theta}_0$; and the *output fibres*, the directions along which the training predictions do not change but predictions at unseen inputs can. In the kernel regime the flow manifold is the affine plane $\boldsymbol{\theta}_0 + \operatorname{Im}(J^\top)$, J the Jacobian of the training outputs with respect to the parameters (Section 2), and anchored ridge, the squared chord from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$, lives within it (Figure 1, right): its gradient never deflects parameters along the fibres, and the gradient-flow limit is preserved as the regularisation strength shrinks. In the feature-learning regime the flow manifold curves in general and the chord leaves it (Figure 1, left): at the gradient-flow limit the ridge gradient acquires a component along the output fibres, invisible to the training loss and consequential at test points, and anchored ridge selects the gradient-flow limit only when the total displacement is aligned with the terminal horizontal space. The bias is silent at the loss level and persists as regularisation shrinks to zero (Theorem 3.1).

The same geometry supplies the remedy, given one modelling principle, least energy. Training itself fixes a rule for pricing motion, the kinetic energy of the optimiser, once one asks that the rule use nothing but the optimiser's metric and agree with the rate at which training decreases the loss; and it fixes the motions a regulariser may add without changing the trained network, those along the flow manifold. The least such energy from initialisation is *geodesic ridge* (Definition 4.1): squared geodesic distance from initialisation along the flow manifold, plus an analogous on-fibre term. Its intrinsic gradient, that of its restriction to the flow manifold, is at the gradient-flow limit orthogonal to the output fibre (Figure 1, left, dashed orange), and its energy on the training outputs, read through the maximum a posteriori (MAP) correspondence, generalises the NTK Gaussian-process energy to curved geometry. In the kernel regime the manifold flattens, geodesics straighten into chords,

and geodesic ridge reduces to anchored ridge: the progression standard ridge $\rightarrow$ anchored ridge $\rightarrow$ geodesic ridge is a sequence of geometric corrections, one Euclidean for initialisation, then one Riemannian for curvature.

Geodesic ridge presupposes the trained network and is inaccessible at scale, so our final contribution is a tractable upper bound that retains the alignment property: *arc ridge*, the squared path length of the training trajectory itself, a scalar accumulated over training that bounds geodesic ridge from above, is minimax-optimal among such bounds, and coincides with early stopping under exact gradient flow.

Contributions.

- **Training defines the surface it lives on.** Under a ball-local conditioning assumption alone (Section 2), the parameters from which gradient flow produces the same network as from $\boldsymbol{\theta}_0$ form a smooth n-dimensional surface, the flow manifold: it carries the whole trajectory, is invariant under training, is the unique such surface meeting the interpolating set only at the trained network, and depends on the labels by theorem: no invariant surface is label-free unless the horizontal directions integrate, fitting together into n-dimensional surfaces, as in the kernel regime (Theorem 2.2; Appendix D.2).
- **Anchored ridge is unbiased only in the aligned case.** Anchored (and, with $\boldsymbol{\theta}_0$ replaced by $\mathbf{0}$, standard) ridge selects the gradient-flow limit only if the total displacement over training lies in the terminal horizontal space, which holds when that space does not rotate and can fail once it does, curvature deflecting the ridge gradient along output fibres invisible to the training loss; the bias persists as regularisation vanishes, and it is the drift of a two-timescale dynamics that any smooth regulariser escapes exactly when the trained network is a critical point of its restriction to the interpolating set (Theorem 3.1, Lemma D.5, Theorem D.12, Figure 2, left). Standard ridge fails in both regimes (Remark D.11, Figure 2, right).
- **The canonical regulariser.** The optimiser fixes the rule that prices motion, the outcome fixes the admissible motions, and, under a least-energy principle, the least such energy is *geodesic ridge* on the flow manifold (Theorem 4.1, Definition 4.1): unique for every $n \geq 1$ (Theorem D.19), anchored ridge in the kernel regime, and unbiased for the gradient-flow limit by construction, being anchored on it (Theorem 4.2). It induces a label-independent *training-cost function* that prices every trained network, exactly the NTK Gaussian-process energy in the kernel regime and, like the output energy of anchored ridge, NTK-Gaussian to second order at initialisation in every regime (Remark 4.1).
- **Arc ridge.** The geodesic ridge of the training curve: an admissible force, not a gradient on Θ, that upper-bounds geodesic ridge, is the least upper bound over all geometries consistent with the observed path, and coincides with early stopping under gradient flow (Theorems 4.3 and E.7, Propositions E.2 and E.3), and behaves as predicted on transfer-learning problems (Section 5).

2 Problem Setting

We study overparametrised supervised models $f(\boldsymbol{x}; \boldsymbol{\theta}) : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}$ with scalar output $\mathcal{Y} \subseteq \mathbb{R}$, trained on a dataset $\mathcal{D} = (X, Y) \in (\mathcal{X} \times \mathcal{Y})^n$ with $n \ll m = \dim \Theta$ finite and distinct inputs $X_i \neq X_j$ for $i \neq j$. We assume f is expressible in the *Tensor Programs framework* (Yang, 2019; 2020a;b; Yang & Hu, 2021) with smooth non-linearities. We write $g(\boldsymbol{\theta}) : \Theta \rightarrow \mathcal{Y}^n$ for the stacked output vector with $g(\boldsymbol{\theta})_i = f(X_i; \boldsymbol{\theta})$, $J(\boldsymbol{\theta}) \in \mathbb{R}^{n \times m}$ for its Jacobian, and $K(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) J(\boldsymbol{\theta})^\top \in \mathbb{R}^{n \times n}$ for the instantaneous NTK Gram matrix.

Training is gradient flow on the sum-of-squared-errors (SSE) loss $\mathcal{L}(g(\boldsymbol{\theta}), Y) = \|g(\boldsymbol{\theta}) - Y\|^2$, that is, $\dot{\boldsymbol{\theta}} = -2J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y)$, with function-space dynamics $\frac{d}{dt} g(\boldsymbol{\theta}) = -2K(\boldsymbol{\theta})(g(\boldsymbol{\theta}) - Y)$; we write $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ for the parameters after running gradient flow from $\boldsymbol{\theta}_0$ for a time t. The loss is minimised by the *interpolating* set $\Theta^\star = \{\boldsymbol{\theta} : g(\boldsymbol{\theta}) = Y\}$. Extensions to strictly convex, smooth losses are given in Appendix A (A.3).

Our results rest on one assumption, imposed on a ball about the initialisation: that the Jacobian be well conditioned there:

Assumption 2.1 (Uniform Jacobian conditioning)**.** There exist a radius $R > 0$ and constants $0 < c \leq C < \infty$ such that $c \leq \sigma_{\min}(J(\boldsymbol{\theta}))$ and $\sigma_{\max}(J(\boldsymbol{\theta})) \leq C$ for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R) := \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < R\}$, and $R > C\|g(\boldsymbol{\theta}_0) - Y\|/c^2$.

The quantifier *for all* $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R)$ requires the same bounds c, C uniformly throughout the entire ball of the required radius, including points off the realised trajectory. Proposition A.2 gives a sufficient certificate using an independently bounded variation of the Jacobian throughout the ball.

Intuitively, the assumption asks for a ball on which every training output responds to the parameters, large enough to contain the whole of training: gradient flow cannot travel further than $C\|g(\boldsymbol{\theta}_0) - Y\|/c^2$, so nothing need be assumed about the trajectory or about $\Theta^\star$ away from the ball; containment, convergence, and finite path length follow:

Proposition 2.1 (Containment and convergence of gradient flow)**.** *Under Assumption 2.1, the gradient flow trajectory* $\boldsymbol{\theta}(t) = \phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ *is defined for all* $t \geq 0$*, satisfies* $\|g(\boldsymbol{\theta}(t)) - Y\| \leq \|g(\boldsymbol{\theta}_0) - Y\| e^{-2c^2 t}$ *and* $\int_0^\infty \left\|\dot{\boldsymbol{\theta}}(t)\right\| dt \leq C\|g(\boldsymbol{\theta}_0) - Y\|/c^2 < R$*, so it remains in* $B(\boldsymbol{\theta}_0, R)$ *and converges to an interpolator* $\boldsymbol{\theta}^\star := \lim_{t\to\infty} \phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t) \in \Theta^\star \cap B(\boldsymbol{\theta}_0, R)$.

Assumption 2.1 is the gradient-flow form of the local conditioning hypothesis under which convergence of gradient descent on overparametrised non-linear least squares is proved (Oymak & Soltanolkotabi, 2019; Liu et al., 2022; Plenk et al., 2026), and is weaker than laziness: it bounds the singular values of $J(\boldsymbol{\theta})$ and nothing else, so it lets $\operatorname{Im} J(\boldsymbol{\theta})^\top$ rotate freely within the ball and places no bound on the Hessian of g, the quantity that defines lazy training (Chizat et al., 2019). The lower bound at initialisation is supplied by infinite-width theory; its persistence on a ball of the required radius is the substantive hypothesis, open in the feature-learning regime, and its consequences are verifiable post hoc along any realised trajectory (Appendix A).

The ball also fixes what we mean by a regime:

Definition 2.1 (Kernel and feature-learning regimes)**.** Under Assumption 2.1, the pair $(f, \boldsymbol{\theta}_0)$ is in the *kernel regime* if the model-output Jacobian is constant on the ball, $J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R)$, equivalently $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ there; it is in the *feature-learning regime* otherwise.

Intuitively, in the kernel regime the network coincides with its linearisation about $\boldsymbol{\theta}_0$ throughout training and the NTK is frozen; in the feature-learning regime the Jacobian, and with it the directions in which gradient flow can move, rotates as training proceeds.

Training has an outcome. From $\boldsymbol{\theta}_0$, gradient flow converges to one interpolator, $\boldsymbol{\theta}^\star$, and the initialisations from which this training problem produces exactly $\boldsymbol{\theta}^\star$ form a surface through $\boldsymbol{\theta}_0$: the map sending an initialisation to its trained network is smooth and of full rank on an explicit region around $\boldsymbol{\theta}_0$, so its level sets are surfaces of dimension n, one per trained network. That surface, not the ball, is where the theory lives.

Theorem 2.2 (The basin manifold, informal)**.** *Under Assumption 2.1, training from every point of an explicit neighbourhood* $\mathfrak{B}$ *of* $\boldsymbol{\theta}_0$ *converges, and the training-limit map* $\pi : \mathfrak{B} \to \Theta^\star \cap B(\boldsymbol{\theta}_0, R)$*,* $\pi(\boldsymbol{\theta}) := \lim_{t\to\infty} \phi_{\mathrm{gf}}(\boldsymbol{\theta}, t)$*, is smooth and of full rank. The set* $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) := \pi^{-1}(\boldsymbol{\theta}^\star)$ *of initialisations training to the same network as* $\boldsymbol{\theta}_0$ *is a smooth* n*-dimensional surface through* $\boldsymbol{\theta}_0$*, closed in* $\mathfrak{B}$*, containing the whole trajectory and* $\boldsymbol{\theta}^\star$*, invariant under training, meeting* $\Theta^\star$ *only at* $\boldsymbol{\theta}^\star$ *and tangent there to* $\operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$*; it is the unique closed connected invariant* n*-dimensional surface in* $\mathfrak{B}$ *meeting* $\Theta^\star$ *only at* $\boldsymbol{\theta}^\star$*. Its tangent space at* $\boldsymbol{\theta}$ *is the space of perturbations of* $\boldsymbol{\theta}$ *that training forgets,* $\{\boldsymbol{v} : D\phi_{\mathrm{gf}}(\boldsymbol{\theta}, t)\, \boldsymbol{v} \to \mathbf{0}\}$*. It depends on the labels: an* n*-dimensional surface invariant under training for an open set of label vectors is tangent to* $\operatorname{Im} J(\boldsymbol{\theta})^\top$ *at each of its points, so a label-free invariant surface exists only where the horizontal directions integrate, as they do in the kernel regime, where* $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = (\boldsymbol{\theta}_0 + \operatorname{Im} J^\top) \cap \mathfrak{B}$ *(Theorems D.2, D.3, D.8; Corollary D.3).*

Intuitively, the flow manifold is what the optimiser and the problem determine together, the optimiser through the flow and the problem through the outcome; its tangent directions away from $\boldsymbol{\theta}^\star$ are the perturbations training forgets rather than the directions gradient flow can move, which is why it may tilt out of the horizontal space, curve (Proposition D.7), and depend on the labels, as the trained network does. The training-limit map is the limit map of Li et al. (2022a), smooth by Falconer (1983), and the flow manifold is the stable fibre of $\boldsymbol{\theta}^\star$ in Fenichel's theory (Fenichel, 1974); new is their

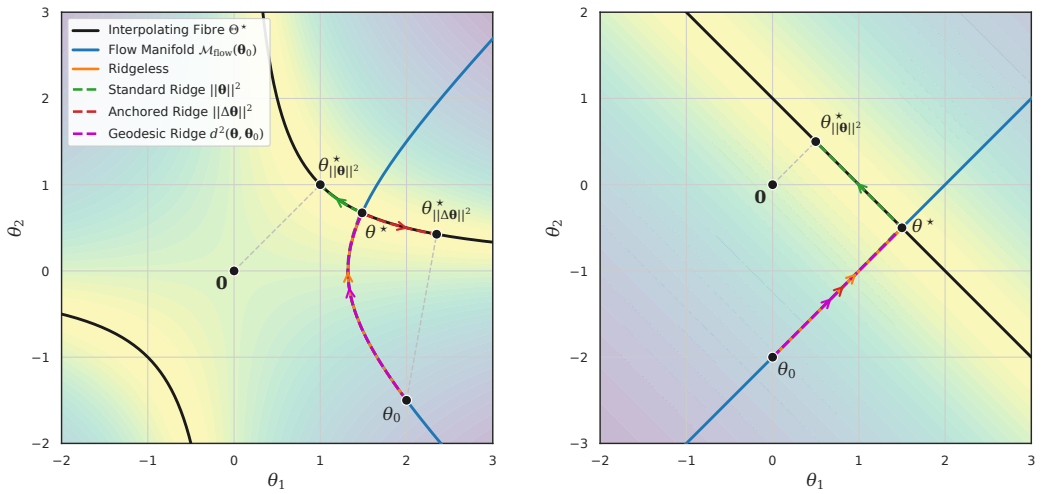


Figure 2: Gradient flow trajectories for the minimal overparametrised non-linear model $f(x;\boldsymbol{\theta}) = \theta_1\theta_2 x$ (left) and linear model $f(x;\boldsymbol{\theta}) = (\theta_1+\theta_2)\,x$ (right) on the singleton dataset $\mathcal{D} = \{(1,1)\}$ under infinitesimal-strength regularisers; shading indicates data loss. Only the canonical regulariser prevents a biased equilibrium in the non-linear case: standard and anchored ridge drift along the interpolating fibre after the data loss converges; in the linear case anchored ridge coincides with the canonical regulariser and standard ridge still biases the solution. In both models the flow manifold is the gradient line of g through $\boldsymbol{\theta}_0$, one-dimensional, so arc ridge coincides with geodesic ridge (Corollary D.3, Proposition E.2); the canonical trajectory shown is the arc-ridge one, which halts short of $\boldsymbol{\theta}^\star$ by a distance invisible here (Proposition E.3).

use as the carrier of a regulariser. The ridge-bias theorem, the early-stopping equivalence, and the predictions of Section 5 use only the assumption; the canonical regulariser uses the surface (Table 1).

3 Ridge Biases Gradient Flow in the Feature-Learning Regime

3.1 Parameter-space geometry

Throughout, Assumption 2.1 holds and we work on the region $\mathfrak{B}$ of Theorem 2.2. Directions in $\operatorname{Im} J(\boldsymbol{\theta})^\top$, which move the training outputs, we call *horizontal*; directions in $\ker J(\boldsymbol{\theta})$, which do not, we call *vertical*. The two are orthogonal complements of $\mathbb{R}^m$ (Lemma C.5), and gradient flow moves horizontally. Using these, and the dynamics of gradient flow, we form two geometric objects:

Definition 3.1 (Flow manifold)**.** The *flow manifold* (or *basin manifold*) $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) := \pi^{-1}(\boldsymbol{\theta}^\star)$ is the set of initialisations from which training produces the network $\boldsymbol{\theta}^\star$, a connected n-dimensional surface through $\boldsymbol{\theta}_0$ (Theorem 2.2), equipped with the metric induced from $\mathbb{R}^m$; $d_{\text{flow}}(\cdot,\cdot)$ denotes its geodesic distance.

Definition 3.2 (Output fibre)**.** The *output fibre* over an output vector $\boldsymbol{c}$ is its preimage $\Theta(\boldsymbol{c}) := g^{-1}(\boldsymbol{c}) = \{\boldsymbol{\theta} : g(\boldsymbol{\theta}) = \boldsymbol{c}\}$.

Intuitively, the flow manifold is the set of initialisations that training cannot tell apart from $\boldsymbol{\theta}_0$, horizontal at $\boldsymbol{\theta}^\star$ and, in general, only there; the output fibre through $\boldsymbol{\theta}$ is the set of parameters making the same training predictions as $\boldsymbol{\theta}$, so motion along it is invisible to the training loss yet can change the predictions at unseen inputs, and the fibre over Y is the interpolating set $\Theta^\star \cap B(\boldsymbol{\theta}_0, R)$. Near $\boldsymbol{\theta}^\star$ the flow manifold is a chart of training-output space: on a neighbourhood V of $\boldsymbol{\theta}^\star$, g maps it diffeomorphically onto a ball $C_0(\boldsymbol{\theta}_0) \ni Y$ of output vectors, and each fibre $\Theta(\boldsymbol{c})$, $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0)$, meets it in exactly one point, the *representative* $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ of $\boldsymbol{c}$, with $\boldsymbol{\theta}^\star = g|_{\boldsymbol{\theta}_0}^{-1}(Y)$ (Theorem D.4); the chart, like the surface, depends on the labels through $\boldsymbol{\theta}^\star$.

3.2 The failure of ridge in the feature-learning regime

The feature-learning regime is the *curved* case of this geometry: J changes with $\boldsymbol{\theta}$ and $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ can bend between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}^\star$ (Figure 1, left), whereas in the kernel regime of Definition 2.1 it is the

affine plane $\boldsymbol{\theta}_0 + \operatorname{Im} J^\top$ (right). The gradient of anchored ridge, $2(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, is twice the chord from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$. In the kernel regime the chord lies in the plane, so anchored ridge never pushes parameters off the flow manifold – it is *admissible* (Proposition D.15); once the flow manifold curves, the chord leaves it, and the ridge gradient acquires a component along the output fibre, invisible to the loss and thus uncontrollable by the training set, yet able to perturb the inductive bias imbued by initialisation. The pathology is a bias in the vanishing-regularisation limit:

Theorem 3.1 (Anchored ridge and the gradient-flow limit, informal)**.** *Under Assumption 2.1, every limit point $\bar{\boldsymbol{\theta}}$ of the minimisers of $\mathcal{L}\left(g\left(\boldsymbol{\theta}\right), Y\right) + \lambda \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|^2$ as $\lambda \downarrow 0$ is an interpolator minimising the chord $\left\|\cdot - \boldsymbol{\theta}_0\right\|^2$ over $\Theta^\star$, and satisfies $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \in \operatorname{Im} J\left(\bar{\boldsymbol{\theta}}\right)^\top$. Hence if $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \notin \operatorname{Im} J\left(\boldsymbol{\theta}^\star\right)^\top$, no sequence of anchored-ridge minimisers converges to the gradient-flow limit $\boldsymbol{\theta}^\star$, however small the regularisation. The same bias holds for standard ridge $\left\|\boldsymbol{\theta}\right\|^2$ with $\boldsymbol{\theta}_0$ replaced by $\mathbf{0}$ (Theorem D.11).*

The *alignment condition* $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \in \operatorname{Im} J\left(\boldsymbol{\theta}^\star\right)^\top$ asks that the total displacement over training lie in the terminal horizontal space. It holds whenever the horizontal space does not rotate along the trajectory, as in the kernel regime, where anchored ridge is indeed unbiased (Corollary D.6), and can fail once it does, the misalignment being bounded by the rotation (Lemma D.5); for the bilinear model of Figure 2 it fails at every off-diagonal initialisation (Remark D.10). It is necessary but not sufficient: the chord may be stationary rather than minimal at $\boldsymbol{\theta}^\star$ along the curved interpolating set, and anchored ridge is then biased although the condition holds (Remark D.7). Standard ridge is biased already in the kernel regime (Remark D.11; Figure 2, right). The selection principle is classical in nonlinear Tikhonov regularisation (Engl et al., 1989; 1996); what it implies for learning is that the identification of the implicit prior with a parameter norm, on which the kernel-regime toolkit rests, is lost whenever alignment fails, for weight decay, L2-SP, and every other fixed-anchor quadratic penalty (Proposition D.13). The mechanism is a drift on two timescales (Katzenberger, 1991; Li et al., 2022a; Boursier et al., 2025): small-λ regularised trajectories first track the unregularised one to within $O(\lambda \log \frac{1}{\lambda})$ of $\boldsymbol{\theta}^\star$, then move along the interpolating set under the gradient flow of the regulariser restricted to it (Theorem D.12), so a regulariser is drift-free exactly when $\boldsymbol{\theta}^\star$ is a critical point of its restriction; for anchored ridge that is the alignment condition. For larger strengths λ, necessary with little data or noisy targets, the drift is immediate. This is not always harmful: a related drift drives weight-decay-induced *grokking* (Power et al., 2022; Xu et al., 2026) and underpins the adaptive kernel representer theorem of Lauditi et al. (2025); but where the feature geometry suits the task a priori, as in pretrained networks, it replaces useful inductive bias with drifted parameters (Section 5).

4 Canonical Regularisation in the Feature-Learning Regime

In the kernel regime anchored ridge is the *canonical regulariser* and the NTK Gaussian process the *canonical function-space prior*; we now derive the feature-learning counterpart from the training dynamics and one principle, with anchored ridge recovered in the kernel regime as a consistency check. Three inputs are needed, and training supplies all three. The first is a rule for pricing motion in parameter space. It may use nothing but the metric in which the optimiser runs, the Euclidean metric of the effective parameters, and it must agree, along gradient flow for every label vector, with the rate at which training destroys the loss, $-\frac{d}{dt}\mathcal{L} = \left\|\dot{\boldsymbol{\theta}}\right\|^2$; agreement prices the horizontal velocities gradient flow realises, and the first requirement all others, among them the tilted velocities of the flow manifold. The second is the class of motions a regulariser may add: those that keep regularised training on the flow manifold at every strength, since forces that leave it change the trained network, the mechanism of Section 3. The third is the anchor, $\boldsymbol{\theta}_0$. The principle, a modelling choice, is least energy: the canonical regulariser charges a parameter the least energy, by the rule, that any outcome-preserving run must spend to reach it:

Theorem 4.1 (The canonical regulariser, informal)**.** *Under Assumption 2.1: (i) a rule pricing motion in parameter space that uses nothing but the optimiser's metric and agrees, along gradient flow for every label vector, with the rate at which training decreases the loss is the Euclidean kinetic energy $\int \left\|\dot{\boldsymbol{\theta}}\right\|^2 dt$ in effective coordinates, with no free scale, for every $n \geq 1$, agreement pricing the horizontal velocities and the metric alone the rest (Lemma D.6); (ii) a regularising force keeps every regularised run started on $\mathcal{M}_{\text{flow}}\left(\boldsymbol{\theta}_0\right)$ in it, at every strength, if and only if it is tangent to*

$\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, and the runs of such forces from $\boldsymbol{\theta}_0$ are exactly the paths in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (Lemma E.2); (iii) the least energy over such runs of unit duration from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, which is also the largest function of $\boldsymbol{\theta}$ vanishing at $\boldsymbol{\theta}_0$ whose square root never grows faster than the length of any outcome-preserving motion (Theorem D.16).

Intuitively, (i) says the price of motion is what training pays for it, (ii) that the only free motions are those the trained network cannot see, and (iii) that the regulariser is the largest penalty consistent with early stopping on every such run: the value function of the optimiser's energy over the class the outcome fixes. The value function lives on the flow manifold; off it, the remaining *vertical* leg is priced by the same rule along the output fibre from the representative, and in the kernel regime the two legs are orthogonal projections of $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ and the Pythagorean identity gives $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$: anchored ridge (Corollary D.8). The canonical regulariser of the feature-learning regime is therefore:

Definition 4.1 (Geodesic ridge)**.** For $\boldsymbol{\theta} \in V$ with $g(\boldsymbol{\theta}) \in C_0(\boldsymbol{\theta}_0)$, *geodesic ridge* is

$$d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) := d^2_{\mathrm{flow}}\left(\boldsymbol{\theta}_0, g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta}))\right) + d^2_{\mathrm{fibre}}\left(g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta})), \boldsymbol{\theta}\right),$$

the squared geodesic distance along the flow manifold from $\boldsymbol{\theta}_0$ to the representative $g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta}))$ of the outputs of $\boldsymbol{\theta}$, the least energy of Theorem 4.1(iii), plus the squared geodesic distance along the output fibre $\Theta(g(\boldsymbol{\theta})) \cap V$ from that representative to $\boldsymbol{\theta}$; we set $d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) := +\infty$ elsewhere. The vertical leg is priced equivalently; this choice does not affect the regularised estimator (Proposition D.17).

Theorem D.19 makes the uniqueness of geodesic ridge precise; the four properties below follow from anchoring on the flow manifold:

Theorem 4.2 (Geodesic ridge is the canonical regulariser)**.** *Under Assumption 2.1:*

- *(a)* ***Unique minimiser on each fibre.*** *For every $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0)$, $g|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c})$ is the unique minimiser of $d^2(\cdot, \boldsymbol{\theta}_0)$ over $\Theta(\boldsymbol{c}) \cap V$.*
- *(b)* ***Trajectory-level minimality.*** *Whenever the gradient flow trajectory $\boldsymbol{\theta}(t)$ lies in V, $\boldsymbol{\theta}(t) = g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta}(t)))$ and $d^2(\boldsymbol{\theta}(t), \boldsymbol{\theta}_0) = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}(t)) \leq s(t)^2$, $s(t)$ its arc length: every trajectory point is the minimum-cost representative of its output vector.*
- *(c)* ***Unbiased limit property.*** $\lim_{\lambda \downarrow 0} \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \boldsymbol{\theta}^\star$.
- *(d)* ***Intrinsic gradient flow.*** *Gradient flow on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is the gradient flow of the restricted loss for the induced metric, whose only critical point is $\boldsymbol{\theta}^\star$, a strict minimum; the fibre drift of any twice-differentiable regulariser at small λ is the gradient flow of its restriction to $\Theta^\star$ from $\boldsymbol{\theta}^\star$, which for geodesic ridge is stationary whenever its flow term is C^2 near $\boldsymbol{\theta}^\star$, as when minimising geodesics from $\boldsymbol{\theta}_0$ to points near $\boldsymbol{\theta}^\star$ are unique and non-conjugate (Theorem D.12, Corollary D.7, Proposition E.5).*

Figure 2 exemplifies (a)–(c). Property (c) holds by construction, geodesic ridge being anchored on the outcome: the exact regulariser presupposes $\boldsymbol{\theta}^\star$ and cannot be evaluated before training, which is why arc ridge, its surrogate, is the practical object.

Remark 4.1 (Energy of the training problem and training-cost function). Read as a MAP objective, $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ pushes forward exactly to output space: $\inf_{\boldsymbol{\theta}}[\mathcal{L} + \lambda d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)] = \inf_{\boldsymbol{c}}[\|\boldsymbol{c} - Y\|^2 + \lambda E_Y(\boldsymbol{c})]$ with $E_Y(\boldsymbol{c}) := d^2_{\mathrm{flow}}\left(\boldsymbol{\theta}_0, g|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c})\right)$ the least value of geodesic ridge on the fibre over $\boldsymbol{c}$, the *energy of the training problem*, and the regularised estimator is the representative $g|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c}_\lambda)$ of the shrunk outputs (Proposition F.1); $p(\boldsymbol{c} \mid X) \propto \exp(-\beta E_Y(\boldsymbol{c}))$ is its prior reading, the scale β absorbed into λ. E_Y depends on the labels through $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. The label-free objects come from the problems sharing g and $\boldsymbol{\theta}_0$: the *training map* τ sends $\boldsymbol{c}$ to the network trained from $\boldsymbol{\theta}_0$ on labels $\boldsymbol{c}$, its image, the *reachable manifold*, is the n-surface of networks training from $\boldsymbol{\theta}_0$ can produce, and anchoring on τ gives the one lift unbiased for every label vector (Theorems F.2, F.4). The *training-cost function* $E(\boldsymbol{c}) := E_{\boldsymbol{c}}(\boldsymbol{c})$, the cost of training from $\boldsymbol{\theta}_0$ to $\tau(\boldsymbol{c})$, agrees with E_Y at Y and is NTK-Gaussian to second order at initialisation in every regime, with the NTK velocity at third order, an expansion shared by the least value of anchored ridge on each fibre (Theorem F.3). In the kernel regime $E_Y = E$ is the squared NTK RKHS norm for every Y and the prior is the marginal at X of the NTK Gaussian process with mean $f(\cdot; \boldsymbol{\theta}_0)$ and covariance $(2\beta)^{-1}$ times the NTK (Jacot et al., 2018; Rasmussen & Williams, 2006) (Proposition D.15(f)). Whether either extends beyond the training inputs to a process is open.

4.1 ARC RIDGE: A MINIMAX-ROBUST UPPER BOUND TO GEODESIC RIDGE

Geodesic ridge is also intractable at scale: its value needs a boundary-value problem on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, a surface a run knows only through its own path, and its intrinsic gradient the terminal tangent of the minimising geodesic. Both are path functionals of that geodesic, the value its squared length and the gradient $2d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ times its unit terminal tangent, by the first variation of length (Lemma E.1), and both may be evaluated on *any* path on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$, as the realised training path is under exact gradient flow. There the length bounds geodesic ridge from above and the endpoint term gives a force tangent to $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, not the gradient of a potential on Θ, arc length being a functional of the whole path, but with the three properties the theory requires:

Definition 4.2 (Arc ridge and the arc-ridge force)**.** For a piecewise-C^1 training path $\gamma : [0, T] \to \Theta$ with arc length $s(T) := \int_0^T \|\dot\gamma(t)\|\, dt$, *arc ridge* is the squared path length $L^2(\gamma) := s(T)^2$, and the *arc-ridge force* at the endpoint $\boldsymbol{\theta} = \gamma(T)$ is $\boldsymbol{F}_{\text{arc}}(T) := 2\, s(T)\, \dot\gamma(T^-)/\, \|\dot\gamma(T^-)\|$, the terminal unit tangent scaled by $2s(T)$. Both depend only on the image of γ, not on its parametrisation. Under gradient flow $\dot\gamma \propto -\nabla_{\boldsymbol{\theta}}\mathcal{L}$, so $\boldsymbol{F}_{\text{arc}} = -2s\, \nabla_{\boldsymbol{\theta}}\mathcal{L}/\, \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|$, which we take as the definition of the force at any $(\boldsymbol{\theta}, s)$ with $\boldsymbol{\theta} \notin \Theta^\star$.

Theorem 4.3 (Arc ridge is an admissible upper bound to geodesic ridge)**.** *Under Assumption 2.1, let $\gamma : [0, T] \to \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ be any piecewise-C^2 path with $\gamma(0) = \boldsymbol{\theta}_0$, $\gamma(T) = \boldsymbol{\theta} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ and $\dot\gamma(T^-) \neq 0$. Then:*

- *(a)* ***Value bound.*** *$d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = d^2_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \le L^2(\gamma)$, with equality if and only if γ is a minimising geodesic of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$.*
- *(b)* ***Admissibility.*** *$\boldsymbol{F}_{\text{arc}}(T) \in T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, and under gradient flow also in $\operatorname{Im} J(\boldsymbol{\theta})^\top$: the force is admissible, as the intrinsic gradient of geodesic ridge and unlike the gradient of anchored ridge.*
- *(c)* ***Force bound.*** *$\|\boldsymbol{F}_{\text{arc}}(T)\| = 2L(\gamma) \ge 2\, d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \left\|\operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\right\|$ wherever the latter exists, with equality of vectors if and only if γ is a minimising geodesic.*

Under exact gradient flow the realised path is such a γ, and more holds: the training curve $\Gamma_{\text{gf}} := \{\phi_{\text{gf}}(\boldsymbol{\theta}_0, t) : t \ge 0\}$ is a one-dimensional submanifold of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ on which arc length is the intrinsic distance from $\boldsymbol{\theta}_0$ and $\boldsymbol{F}_{\text{arc}}$ the intrinsic gradient of its square, so anchored, geodesic, and arc ridge are the squared intrinsic distances from $\boldsymbol{\theta}_0$ on $\Theta \supset \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \supset \Gamma_{\text{gf}}$, and the sandwich $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \le d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le s^2$ along the trajectory is the monotonicity of intrinsic distance under inclusion (Proposition E.2, Corollary E.1).

The training run reveals the geometry only through the path it traces. For every geometry consistent with that path the canonical regulariser lies in the sandwich $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \le d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le L^2(\gamma_{\text{train}})$, γ_{train} the realised training path, and its upper end is the least upper bound over those geometries (Theorem E.7); the minimax-robust choice, never under-regularising relative to the canonical prior whatever the geometry, is that upper end, and among all functionals of the path with this property the squared path length is the smallest: arc ridge. It is exact when the realised path is a minimising geodesic of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, in the kernel regime precisely when the initial residual is an NTK eigenvector (Proposition E.9); and it is the only end of the sandwich whose force is admissible.

Under exact gradient flow, arc ridge coincides with early stopping: the regularised flow $\dot{\boldsymbol{\theta}} = -\big(1 - 2\lambda s/\, \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|\big)\nabla_{\boldsymbol{\theta}}\mathcal{L}$ runs along the unregularised trajectory and halts at the first point at which $2\lambda s$ balances $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|$ (Proposition E.3). The correspondence between early stopping and ridge is classical in the linear and kernel settings (Engl et al., 1996; Yao et al., 2007); in the feature-learning regime early stopping is MAP estimation under the canonical prior of the least-favourable geometry consistent with the realised training path, the minimax-robust relaxation of Theorem E.7 (Proposition E.8). This offers one explanation for why early stopping stays safe under strong regularisation in the feature-learning regime while anchored ridge does not: its action is confined to the realised flow direction.

5 EMPIRICAL CONSEQUENCES

Figure 2 realises the hypotheses of Theorems 3.1 and 4.2 exactly on the minimal example for each regime. Figure 3 tests the predictions of the theory where its hypotheses are not known to hold,

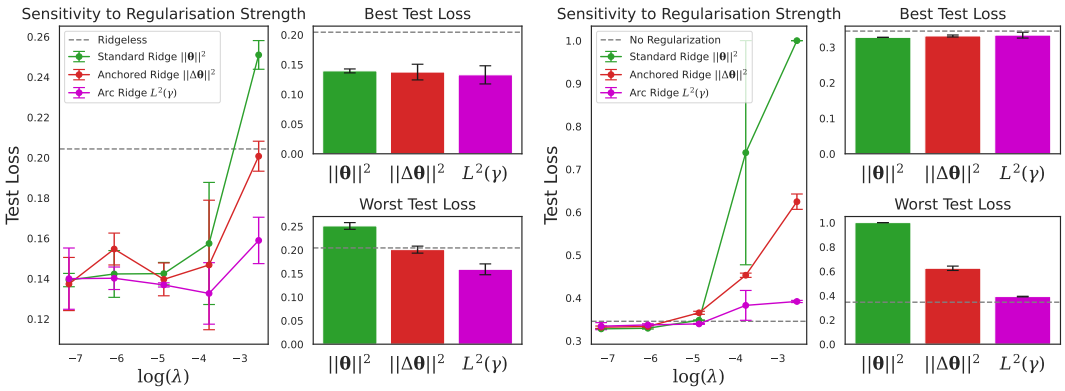


Figure 3: Test-set mean squared error (MSE) on UTKFace (left) and Yelp Review (right) for standard, anchored, and arc ridge across regularisation strengths λ with early stopping, against an unregularised baseline without early stopping (dashed); bars: best and worst test loss over λ. Standard and anchored ridge degrade sharply at high λ; arc ridge degrades mildly, consistent with undertraining.

as kernel-regime predictions are tested on finite networks: age regression on UTKFace (Zhang et al., 2017) from a pretrained ResNet18 (He et al., 2016) and review-score regression on Yelp Review (Zhang et al., 2015) from DistilBERT (Sanh et al., 2019), under momentum SGD with mini-batches and noisy targets (Appendix H; Remark E.1 discusses the deviations from the gradient-flow theory). Each regulariser is run across a range of strengths with early stopping, as in common practice, against an unregularised ("Ridgeless", no early stopping) baseline. Arc ridge being early stopping in force form, we do not claim that it outperforms early stopping; the experiment tests the predicted mechanism. The theory, resting here on Assumption 2.1 alone (Theorem 3.1, Proposition E.3), and the drift criterion of Section 3 predict that at high strengths the fibre component of standard and anchored ridge degrades the pretrained prior, and the test performance with it; that arc ridge, whose force is admissible, degrades only through undertraining, a far more controlled phenomenon; and that at low strengths early stopping – equivalent to optimally tuned arc ridge under exact gradient flow – dominates the fibre components and all regularisers converge in performance. Figure 3 is consistent with all three, each of which could have failed: had standard and anchored ridge held up at the largest strengths, had arc ridge degraded as sharply as they do, or had the three separated at the smallest, the fibre mechanism would not distinguish them.

6 Conclusion and Limitations

We showed that the regulariser and the prior favoured by gradient-flow training are governed by the geometry of training itself – the surface of initialisations sharing its outcome – and not by a norm on parameters. Anchored ridge is aligned with this geometry in the kernel regime; once it curves, the gradient of anchored (and standard) ridge drifts along the interpolating set whenever the alignment condition fails, which matters in transfer learning. The optimiser's own energy, measured over the motions that preserve the trained network, identifies *geodesic ridge* as the unique canonical regulariser, and the training-cost function it induces on the outputs generalises the energy of the NTK Gaussian process. For unbiased regularisation at scale we proposed *arc ridge*, a cheap, minimax-robust surrogate to geodesic ridge that coincides with early stopping under exact gradient flow and degrades mildly where standard and anchored ridge degrade sharply.

One hypothesis defines our setting and bounds what we have shown. The lower singular-value bound of Assumption 2.1 is established along entire trajectories only under conditions that also enforce near-laziness; whether it can be decoupled from laziness is open. Beyond it, one principle is chosen, least energy; the label dependence of the flow manifold is a theorem, not a choice: what depends on the labels is the regulariser of a training problem, as early stopping does, and what does not is the training-cost function that prices every trained network (Appendix F). Open directions include scalable approximations to geodesic ridge, verification of the assumption at scale, the flow manifold in the μP infinite-width limit (Appendix G), non-interpolating regimes (Soudry et al., 2018; Lyu & Li, 2020), and preconditioned optimisers (Kingma & Ba, 2015; Jordan et al., 2024).

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing the experimental code, to polish the text, and to check the proofs. We have not used generative AI tools to generate research ideas, theorems, mathematical arguments, experimental designs or experimental results: the ideas, the theorems and their proofs, the mathematics and the experiments are solely attributable to the authors. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and methodological; its experiments use the public UTKFace and Yelp Review datasets under their published terms. UTKFace contains face images with age labels; we use it only as a regression benchmark and release no derived personal data. We see no direct ethical risk beyond those inherent in the datasets.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 2 and collected in Appendix A; every formal result is stated in full and proved in Appendix D, with the detailed analysis of arc ridge in Appendix E, the pushforward energies in Appendix F, and what width adds in Appendix G; Table 1 maps each claim of the main text to its formal statement, its proof, and the assumptions it uses. Experimental details (architectures, optimiser settings, seeds, data preprocessing, compute) are in Appendix H, and the code, including the arc-ridge implementation and the scripts producing every figure, is at `https://anonymous.4open.science/r/canonical-regularisation-7349/`.

REFERENCES

Andrei Agrachev, Davide Barilari, and Ugo Boscain. *A Comprehensive Introduction to Sub-Riemannian Geometry*, volume 181 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2019. ISBN 9781108476355.

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

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2 edition, 2008.

Bernd Aulbach. *Continuous and Discrete Dynamics near Manifolds of Equilibria*, volume 1058 of *Lecture Notes in Mathematics*. Springer, Berlin, 1984.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Mikhail Belkin. Neural tangent kernel at initialization: linear width suffices. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 110–118, 2023.

Ari S Benjamin, David Rolnick, and Konrad P Körding. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2205.09653.

Etienne Boursier, Scott Pesme, and Radu-Alexandru Dragomir. A theoretical framework for grokking: Interpolation followed by Riemannian norm minimisation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2505.20172.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

Tianle Cai, Ruiqi Gao, Jikai Hou, Siyu Chen, Dong Wang, Di He, Zhihua Zhang, and Liwei Wang. Gram-Gauss-Newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

Sergio Calvo-Ordoñez, Jonathan Plenk, Richard Bergna, Alvaro Cartea, José Miguel Hernández-Lobato, Konstantina Palla, and Kamil Ciosek. A Gaussian process view on observation noise and initialization in wide neural networks. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Luís Carvalho, João L. Costa, José Mourão, and Gonçalo Oliveira. The positivity of the neural tangent kernel. *SIAM Journal on Mathematics of Data Science*, 7(2):495–515, 2025. doi: 10.1137/24M1659534.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton–Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

Alexandru Damian, Jason D Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

deedax. UTK-Face-Revised. Hugging Face dataset, `https://huggingface.co/datasets/deedax/UTK-Face-Revised`, 2023. Revision 91a76d0b7626172b7fc5ca3c8ca2bbf2233fbefa; derived from UTKFace.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, 2 edition, 2010.

Manfredo Perdigão do Carmo. *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser, 1992.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Heinz W Engl, Karl Kunisch, and Andreas Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, 1989.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, 1996.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, revised edition, 2015.

K. J. Falconer. Differentiation of the limit mapping in a dynamical system. *Journal of the London Mathematical Society*, 27(2):356–372, 1983.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959. doi: 10.1090/S0002-9947-1959-0110078-1.

Neil Fenichel. Persistence and smoothness of invariant manifolds for flows. *Indiana University Mathematics Journal*, 21(3):193–226, 1971.

Neil Fenichel. Asymptotic stability with rate conditions. *Indiana University Mathematics Journal*, 23(12):1109–1137, 1974.

Neil Fenichel. Asymptotic stability with rate conditions, II. *Indiana University Mathematics Journal*, 26(1):81–93, 1977.

Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979.

Henry Gouk, Timothy Hospedales, and Massimiliano Pontil. Distance-based regularisation of deep networks for fine-tuning. In *International Conference on Learning Representations*, 2021.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, 2018.

Martin Hanke, Andreas Neubauer, and Otmar Scherzer. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numerische Mathematik*, 72(1):21–37, 1995.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1010–1022, 2020.

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, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Morris W Hirsch, Charles C Pugh, and Michael Shub. *Invariant manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer-Verlag, 1977.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

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, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Christopher K. R. T. Jones. Geometric singular perturbation theory. In Russell Johnson (ed.), *Dynamical Systems (Montecatini Terme, 1994)*, volume 1609 of *Lecture Notes in Mathematics*, pp. 44–118. Springer, Berlin, 1995.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Ryo Karakida and Kazuki Osawa. Understanding approximate Fisher information for fast convergence of natural gradient descent in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10891–10901, 2020.

Gary S Katzenberger. Solutions of a stochastic differential equation forced onto a manifold by a large drift. *The Annals of Probability*, 19(4):1587–1628, 1991.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil 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*, 114(13):3521–3526, 2017.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, pp. 950–957, 1992.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 32617–32648, 2025.

Jaehoon Lee, Lechao Xiao, Samuel 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*, volume 32, 2019.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2012. ISBN 978-1-4419-9981-8. doi: 10.1007/978-1-4419-9982-5.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pp. 2825–2834. PMLR, 2018.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? A mathematical framework. In *International Conference on Learning Representations*, 2022a.

Zhiyuan Li, Tianhao Wang, Jason D Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34626–34640, 2022b.

Pierre-Louis Lions. *Generalized Solutions of Hamilton–Jacobi Equations*, volume 69 of *Research Notes in Mathematics*. Pitman, Boston, 1982.

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*, volume 33, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022. doi: 10.1016/j.acha.2021.12.009.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Carlo Mantegazza and Andrea Carlo Mennucci. Hamilton–Jacobi equations and distance functions on Riemannian manifolds. *Applied Mathematics and Optimization*, 47(1):1–25, 2002.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, 2026.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Tiberiu Musat. The geometry of grokking: Norm minimization on the zero-loss manifold. arXiv preprint arXiv:2511.01938, 2025.

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

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately Bayesian ensembling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 234–244, 2020.

Jonathan Plenk, Sergio Calvo-Ordoñez, Álvaro Cartea, Yarin Gal, Mark van der Wilk, and Kamil Ciosek. The neural tangent kernel for classification. *arXiv preprint arXiv:2605.17606*, 2026.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21174–21187, 2020.

Grant M Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tim G J Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *International Conference on Machine Learning*, pp. 29275–29290. PMLR, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Héctor J. Sussmann. Orbits of families of vector fields and integrability of distributions. *Transactions of the American Mathematical Society*, 180:171–188, 1973. doi: 10.1090/S0002-9947-1973-0321133-2.

Andrei Nikolaevich Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

Gal Vardi and Ohad Shamir. Implicit regularization in ReLU networks with the square loss. In *Conference on Learning Theory*, pp. 4224–4258. PMLR, 2021.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Mingyue Xu, Gal Vardi, and Itay Safran. To grok grokking: Provable grokking in ridge regression. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://arxiv.org/abs/2601.19791`.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Greg Yang. Tensor Programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020a.

Greg Yang. Tensor Programs III: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020b.

Greg Yang and Edward J Hu. Tensor Programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097, 2021.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315, 2007.

Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Overview of the Appendices

Appendix A states the assumptions, the integrable class, and the verification of Assumption 2.1; Appendix B collects related work; Appendix C proves preliminary lemmas; Appendices D and E give, in main-text order, the formal results behind Sections 2–4 and behind arc ridge (Section 4.1); Appendix F treats the pushforward energies; Appendix G separates what width adds into theorems and conjectures; Appendix H gives experimental details. Table 1 maps each claim of the body to its formal statement and proof. The results of Section 4 are local: they hold on the chart neighbourhood V of Theorem D.4, on which geodesic ridge is finite.

A Assumptions

We list the atomic hypotheses of the main results, with their users and justification. The manifold structure of the flow manifold, its chart near the interpolator, and the existence of a unique gradient-flow limit are consequences, proved in Appendix D.

Table 1: Where each claim of the main text is made formal. Every row needs A.1–A.3 and Assumption 2.1; "kernel" marks the kernel regime of Definition 2.1, "integrable" the class of Remark A.1, in which the row's object becomes label-free.

Claim in the main text	Formal statement	Proof	Regime
Gradient flow stays in the ball and converges (Prop. 2.1)	Cors. D.1, D.2	§D.1	all
Training has an outcome; the flow manifold is the unique invariant n-surface of forgotten perturbations (Thm. 2.2, Def. 3.1)	Prop. D.1, Thms. D.2, D.3, Lems. D.1, D.2	§D.2	affine in kernel (Cor. D.3)
It charts output space near $\boldsymbol{\theta}^\star$, is tilted and curved, and depends on the labels unless $\mathcal{H}$ integrates	Thms. D.4, D.5, D.8, D.9, Cor. D.4, Lem. D.3, Prop. D.7	§D.2, §D.3	untilted, label-free in integrable
Output fibres are submanifolds (Def. 3.2)	Prop. D.10, Lem. C.5	§D.4	all
Ridge misses the gradient-flow limit unless aligned; alignment holds in the kernel regime, fails under rotation, is not sufficient (Thm. 3.1)	Thm. D.11, Cor. D.6, Lem. D.5, Rmks. D.7, D.9	§D.5	all
The drift is the regulariser's gradient flow on $\Theta^\star$	Thm. D.12, Cor. D.7	§D.5	all
Fixed-anchor quadratic and output-norm penalties fail	Props. D.13, D.14	§D.6	all
Rule, admissible motions and flow term; the convention does not affect the estimator; uniqueness (Thms. 4.1, D.19)	Lems. D.6, E.2, Thm. D.16, Props. D.17, D.18, Rmk. D.13	§D.8, §E.4	all $n \geq 1$; anchored ridge in kernel
Characterisation of geodesic ridge (Thm. 4.2)	as stated	§D.9	all
Kernel regime: RKHS energy, anchored ridge, NTK Gaussian process	Prop. D.15, Cor. D.8	§D.7	kernel
Energy of the training problem; label-free training-cost function, NTK-Gaussian to second order (Rmk. 4.1)	Prop. F.1, Rmk. F.1, Thms. F.2, F.3, F.4	§F	equal near Y in integrable
Arc ridge upper-bounds geodesic ridge with an admissible force, and is the geodesic ridge of the training curve (Thm. 4.3)	Props. E.1, E.2, Lem. E.1	§E.1, §E.2	all
Admissible forces do not bias; intrinsic vs. ambient dynamics	Thm. E.4, Prop. E.5	§E.6	all
Sandwich; minimax robustness	Prop. E.6, Cor. E.1, Thm. E.7, Prop. E.8	§E.7, §E.8	the path alone
Early stopping; kernel tightness; implementation	Props. E.3, E.9, Rmks. E.1, E.2	§E.5, §E.9, §E.10	all; kernel

(A.1) Smoothness. *Statement.* The network $f(\cdot;\boldsymbol{\theta})$ is C^∞ in $\boldsymbol{\theta}$ on the parameter domain of interest. The parameter domain $\Theta \subseteq \mathbb{R}^m$ is open and contains the closed ball $\overline{B}(\boldsymbol{\theta}_0, R)$ of Assumption 2.1. *Used by.* All main-text results. *Justification.* Compositions of smooth activations are C^∞; the practical restriction is that activations satisfy the pseudo-Lipschitz / polynomially bounded growth hypothesis imposed by the Tensor Programs framework (Yang, 2020a;b), covering analytic variants of $\tanh$, smooth SiLU, and smooth approximations of ReLU.

(A.2) Data. *Statement.* The training dataset $\mathcal{D} = (X, Y)$ has finite n and distinct inputs $X_i \neq X_j$ for $i \neq j$. *Used by.* All main-text results. *Justification.* Distinct inputs are necessary for the output fibre $\Theta(\boldsymbol{c})$ to be well-defined at every achievable $\boldsymbol{c}$; finiteness of n ensures the Jacobian $J(\boldsymbol{\theta}) \in \mathbb{R}^{n\times m}$ has finite domain.

(A.3) Loss and gradient flow dynamics. *Statement.* Training is the continuous-time gradient flow $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(g(\boldsymbol{\theta}), Y) = J(\boldsymbol{\theta})^\top \boldsymbol{u}$ with $\boldsymbol{u} := -\partial_{\boldsymbol{c}}\mathcal{L}(\boldsymbol{c}, Y)|_{\boldsymbol{c}=g(\boldsymbol{\theta})}$. The main text takes $\mathcal{L}(\boldsymbol{c}, Y) = \|\boldsymbol{c} - Y\|^2$ (squared error), giving $\boldsymbol{u} = 2(Y - g(\boldsymbol{\theta}))$, $\nabla^2_{\boldsymbol{c}}\mathcal{L}(\cdot, Y) = 2I$, and Hessian $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(g(\boldsymbol{\theta}), Y) = 2J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta}) + 2\sum_i (g(\boldsymbol{\theta}) - Y)_i \nabla^2 g_i$, which reduces to $2J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta})$ at any $\boldsymbol{\theta} \in \Theta^\star$. *Used by.* All main-text results. *Justification.* A definitional choice; gradient descent approximates it at small step size. The main text states squared error only. Every result extends to a loss $h(\boldsymbol{c}) := \mathcal{L}(\boldsymbol{c}, Y)$ that is μ-strongly convex and L_h-smooth on the sublevel set $\{h \leq h(g(\boldsymbol{\theta}_0))\}$ with minimiser Y, through three substitutions stated once here and used throughout the appendix: the rate $2c^2$ of Proposition 2.1 becomes $c^2\mu^2/L_h$ for the residual $\|g(\boldsymbol{\theta}) - Y\|$ (and $2c^2\mu^2/L_h$ for the excess loss); the a-priori path length $\ell = C\|g(\boldsymbol{\theta}_0) - Y\|/c^2$ becomes $\ell_{\mathrm{gen}} = (CL_h^2/c^2\mu^2)\sqrt{2\Delta_0/\mu}$ with Δ_0 the initial excess loss, so that the radius condition of Assumption 2.1 reads $R > \ell_{\mathrm{gen}}$; and the identity Hessian $\nabla^2_{\boldsymbol{c}}\mathcal{L}(\cdot, Y) = 2I$ becomes $\mu I \preceq \nabla^2_{\boldsymbol{c}}\mathcal{L}(\cdot, Y) \preceq L_h I$ wherever it is used (Remark D.1).

(A.4) Uniform Jacobian conditioning on a ball about the initialisation. *Statement.* There exist a radius $R > 0$ and constants $0 < c \leq C < \infty$ such that $c \leq \sigma_{\min}(J(\boldsymbol{\theta}))$ and $\sigma_{\max}(J(\boldsymbol{\theta})) \leq C$ for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R)$, and $R > C\|g(\boldsymbol{\theta}_0) - Y\|/c^2$ (formally Assumption 2.1). Nothing is assumed about the trajectory or about $\Theta^\star$ outside the ball: the trajectory remains in $\overline{B}(\boldsymbol{\theta}_0, \ell)$, $\ell := C\|g(\boldsymbol{\theta}_0) - Y\|/c^2 < R$, and converges to $\boldsymbol{\theta}^\star \in \Theta^\star \cap B(\boldsymbol{\theta}_0, R)$ as a consequence (Proposition 2.1). *Used by.* All main-text theorems and definitions in Sections 3 and 4; Proposition 2.1 and Corollaries D.1 and D.2; Lemmas C.3 and C.4; and proofs throughout Appendices D–G. *Justification.* Assumption 2.1 is the gradient-flow form of the local Jacobian-conditioning hypothesis under which convergence of gradient descent on overparametrised non-linear least squares is proved by a bootstrap (Oymak & Soltanolkotabi, 2019; Liu et al., 2022); Plenk et al. (2026) give the continuous-time version. The constants are self-consistent rather than circular: c is non-increasing and C non-decreasing in R, so the radius condition is a genuine restriction, asking for conditioning on a ball containing every path of length $C\|g(\boldsymbol{\theta}_0) - Y\|/c^2$. The assumption bounds the singular values of $J(\boldsymbol{\theta})$ and nothing else: it places no bound on the Hessian of g, allows $\mathrm{Im}\, J(\boldsymbol{\theta})^\top$ to rotate arbitrarily within the ball, and so does not imply lazy training, whose defining quantity is the Hessian-to-Jacobian ratio (Chizat et al., 2019). The upper bound holds on any bounded set by A.1; the lower bound is the substantive hypothesis. At initialisation it is supplied for μP MLPs with non-polynomial activations: their learning-rate-weighted NTK equals that of the NTK parametrisation at every width (Yang & Hu, 2021, Table 1; Bordelon & Pehlevan, 2022) and converges almost surely to a positive-definite limit (Yang, 2020a, Cor. 7.3; Carvalho et al., 2025, Thms. 1–2) (Remark A.4). What existing theory does not supply is the radius condition: a width-uniform lower bound on $\sigma_{\min}(J(\boldsymbol{\theta}))$ along entire trajectories is established only under hypotheses that also enforce near-laziness (Du et al., 2019; Oymak & Soltanolkotabi, 2019; Liu et al., 2022), and whether it can be decoupled from laziness is open. Our results are conditional on it. It is verifiable post hoc (Corollary D.2, Appendix A.1) and holds in closed form for the model of Figure 2 (Proposition A.1). Two consequences are derived, not assumed: $\Theta^\star \cap B(\boldsymbol{\theta}_0, R)$ is an embedded C^∞ codimension-n submanifold with tangent space $\ker J(\boldsymbol{\theta})$ (Lemma C.4 (Lee, 2012, Ch. 5)), and on it the loss Hessian $2J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta})$ has n eigenvalues at least $2c^2$ on $\mathrm{Im}\, J(\boldsymbol{\theta})^\top$ and $m - n$ zero eigenvalues on $\ker J(\boldsymbol{\theta})$.

Remark A.1 (The integrable class). Call the horizontal distribution $\mathcal{H}_{\boldsymbol{\theta}} = \mathrm{Im}\, J(\boldsymbol{\theta})^\top = \mathrm{span}\{\nabla g_1(\boldsymbol{\theta}), \ldots, \nabla g_n(\boldsymbol{\theta})\}$ *integrable* on $B(\boldsymbol{\theta}_0, R)$ if through every point of the ball passes an

n-dimensional surface tangent to $\mathcal{H}$ at each of its points. Under A.4 the distribution has constant rank n, so by the Frobenius theorem (Lee, 2012, Ch. 19) integrability is equivalent to *involutivity*, closure of $\mathcal{H}$ under Lie brackets: writing $P_{\boldsymbol{\theta}} = J(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})^{-1} J(\boldsymbol{\theta})$ for the orthogonal projector onto $\mathcal{H}_{\boldsymbol{\theta}}$, the condition is $(I - P_{\boldsymbol{\theta}})[\nabla g_i, \nabla g_j] = (I - P_{\boldsymbol{\theta}})\big(\nabla^2 g_j\, \nabla g_i - \nabla^2 g_i\, \nabla g_j\big) = 0$ for all $i < j$ on the ball, $\binom{n}{2}(m-n)$ scalar equations on the second derivatives of the training outputs. Dually, it holds if and only if every point of the ball has a neighbourhood carrying $m - n$ independent smooth functions ψ_k with $J(\boldsymbol{\theta})\,\nabla\psi_k = 0$, quantities conserved by gradient flow for every label vector and every loss, the maximal number of conservation laws a horizontal control system can have (cf. Marcotte et al., 2023, Thm. 3.3). Integrability is exactly label-independence of the basin manifold: an n-dimensional surface invariant under gradient flow for an open set of label vectors is an integral surface of $\mathcal{H}$ (Theorem D.8), and conversely under integrability the basin manifolds of all label vectors meeting the radius condition, and of all losses, lie in the leaf through $\boldsymbol{\theta}_0$, locally a joint level set of the ψ_k (Theorem D.9). Membership is a joint property of architecture and data, and general networks lie outside the class (Remark A.3(iii)–(iv)). Nothing in this paper assumes membership; the class is where the label-independent objects of Appendix F coincide, near the trained network, with the regulariser of the training problem. The vertical part $B(\nabla g_i, \nabla g_j) := (I - P_{\boldsymbol{\theta}})\nabla^2 g_j\, \nabla g_i = (I - P_{\boldsymbol{\theta}})(D_{\nabla g_i} P_{\boldsymbol{\theta}})\nabla g_j$ of the ambient derivative of horizontal fields is the rotation of $\operatorname{Im} J(\boldsymbol{\theta})^\top$ that drives Lemma D.5; its antisymmetric part $\frac{1}{2}(I - P_{\boldsymbol{\theta}})[\nabla g_i, \nabla g_j]$ is the obstruction to integrability, and its symmetric part, the second fundamental form of the distribution, enters the curvature of the basin manifold at $\boldsymbol{\theta}^\star$ (Proposition D.7).

Remark A.2 (Three regimes). The three classes of models

$$\big\{J(\cdot) \text{ constant on the ball}\big\} \;\subsetneq\; \big\{\mathcal{H} \text{ integrable on the ball}\big\} \;\subsetneq\; \big\{\text{Assumption 2.1 alone}\big\}$$

are nested and distinct: constant J makes $\mathcal{H}$ a fixed subspace, and every single-output problem is in the middle class (Remark A.3(i)) with, for the bilinear model of Figure 2, non-constant J. In the middle class the basin manifold is label-free with tangent space $\operatorname{Im} J(\boldsymbol{\theta})^\top$ everywhere, the metric gap of Theorem D.5 vanishes, and the energies of all training problems are bounded below by one inverse-NTK energy of the outputs, with equality wherever near-minimising paths stay on the flow manifold (Corollary D.5); the kernel regime is its affine case, where geodesic ridge collapses to anchored ridge (Proposition D.15). Feature learning adds curvature of the basin manifold in the middle class and, outside it, also its tilt out of the horizontal space and its dependence on the labels.

Remark A.3 (Examples of integrable and non-integrable models). *(i) A single training output.* For $n = 1$, $\mathcal{H}_{\boldsymbol{\theta}} = \operatorname{span}\{\nabla g(\boldsymbol{\theta})\}$ has rank one wherever $\nabla g \neq 0$ and is involutive, since $[aX, bX] = \big(a\,X(b) - b\,X(a)\big)X$ for any vector field X and functions a, b. The leaves are the gradient lines of g, $K(\boldsymbol{\theta}) = \|\nabla g(\boldsymbol{\theta})\|^2$, and the training-cost function is the explicit one-dimensional energy

$$E(c) \;=\; \Big(\int_{g(\boldsymbol{\theta}_0)}^{c} \big\|\nabla g(g|_{\boldsymbol{\theta}_0}^{-1}(s))\big\|^{-1}\, ds\Big)^2.$$

For the bilinear model $f(x;\boldsymbol{\theta}) = \theta_1\theta_2 x$ of Figure 2 on $\{(1,1)\}$, $\nabla g = (\theta_2, \theta_1)^\top$, the quantity $\delta := \theta_1^2 - \theta_2^2$ is conserved, and for $\delta \neq 0$ the leaf through $\boldsymbol{\theta}_0$ is one branch of the hyperbola $\{\theta_1^2 - \theta_2^2 = \delta\}$, on which $g = \theta_1\theta_2$ is strictly monotone: g restricted to the branch is a global diffeomorphism onto $\mathbb{R}$.

(ii) A curved involutive model with two outputs. Let $f(x;\boldsymbol{\theta}) = \theta_1(\theta_2 x + \theta_3 x^2)$ on two distinct non-zero inputs x_1, x_2, so that $m = 3$, $n = 2$ and $\nabla g_i = (\theta_2 x_i + \theta_3 x_i^2,\ \theta_1 x_i,\ \theta_1 x_i^2)^\top$. Then $\det K(\boldsymbol{\theta}) = \theta_1^2 x_1^2 x_2^2 (x_1 - x_2)^2\, \|\boldsymbol{\theta}\|^2$, so the spectral bounds of Assumption 2.1 hold on any bounded ball whose closure avoids $\{\theta_1 = 0\}$, and

$$[\nabla g_1, \nabla g_2] \;=\; x_1 x_2 (x_1 - x_2)\,(0,\ \theta_3,\ -\theta_2)^\top,$$

which is non-zero but orthogonal to $\xi := (\theta_1, -\theta_2, -\theta_3)^\top$, the generator of the rescaling symmetry $(\theta_1, \theta_2, \theta_3) \mapsto (e^s\theta_1, e^{-s}\theta_2, e^{-s}\theta_3)$, which spans $\ker J(\boldsymbol{\theta})$: the bracket is horizontal and $\mathcal{H}$ is involutive. The conserved quantity is the layer-balance invariant $\theta_1^2 - \theta_2^2 - \theta_3^2$, and the leaves are its level sets, hyperboloids (each sheet a leaf), intrinsically curved and with non-constant J. Geodesic ridge differs from anchored ridge, and Theorem D.11(c) applies whenever the chord from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}^\star$ is not tangent to the hyperboloid at $\boldsymbol{\theta}^\star$, although $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \mathcal{H}_{\boldsymbol{\theta}}$ everywhere: the mechanism is curvature, not tilt (Theorem D.5).

(iii) Involutivity depends on the data. Let $f(x;\boldsymbol{\theta}) = a\tanh(wx+b)$ with $\boldsymbol{\theta} = (a,w,b)$ on two distinct inputs, and write $z_i := wx_i + b$. With $\xi := \nabla g_1 \times \nabla g_2$ spanning $\ker J(\boldsymbol{\theta})$, the Frobenius obstruction of the rank-two distribution $\mathcal{H} = \xi^{\perp}$ in $\mathbb{R}^3$ is

$$\xi \cdot \operatorname{curl} \xi \;=\; a^3 \operatorname{sech}^4 z_1 \operatorname{sech}^4 z_2 \,(x_1 - x_2)(1 + x_1 x_2)\big(\cosh 2z_1 - \cosh 2z_2\big),$$

which vanishes identically if and only if $\xi^{\top}[\nabla g_1, \nabla g_2]$ does. Hence $\mathcal{H}$ is involutive on any open set if and only if $x_1 x_2 = -1$: for the symmetric dataset $x \in \{-1,+1\}$ the neuron is involutive with non-constant J and conserved quantity $4a^2 - \cosh 2(b-w) - \cosh 2(b+w)$, whereas for $x_1 x_2 \neq -1$ the same architecture is non-involutive at every point off the measure-zero set $\{w = 0\} \cup \{w(x_1+x_2) + 2b = 0\}$. Involutivity is thus a joint property of architecture and data, not of architecture alone.

(iv) Two-layer linear networks. For $f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{v}^{\top} W \boldsymbol{x}$ with hidden width and input dimension above one, the span of the Jacobian vector fields over all inputs is not closed under Lie brackets at full-rank parameters (Marcotte et al., 2023, Prop. I.1), and on a fixed dataset the $\boldsymbol{v}$-block of $[\nabla g_i, \nabla g_j]$ vanishes and its W-block is $W(\boldsymbol{x}_i \boldsymbol{x}_j^{\top} - \boldsymbol{x}_j \boldsymbol{x}_i^{\top})$, so the obstruction $(I - P_{\boldsymbol{\theta}})[\nabla g_i, \nabla g_j]$ is non-zero at generic full-rank parameters already for hidden width two, input dimension two and two linearly independent inputs (it vanishes when $\boldsymbol{x}_i \parallel \boldsymbol{x}_j$, in particular in input dimension one): the balance invariants $\boldsymbol{v}\boldsymbol{v}^{\top} - WW^{\top}$ are conserved, but they are too few. *(v) Diagonal linear networks and commuting parametrisations.* For $f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_k u_k v_k x_k$ with $\boldsymbol{\theta} = (\boldsymbol{u},\boldsymbol{v})$, the gradient fields $\nabla g_i = (\boldsymbol{v} \odot \boldsymbol{x}_i,\ \boldsymbol{u} \odot \boldsymbol{x}_i)$ satisfy $\nabla^2 g_j \, \nabla g_i = (\boldsymbol{u} \odot \boldsymbol{x}_i \odot \boldsymbol{x}_j,\ \boldsymbol{v} \odot \boldsymbol{x}_i \odot \boldsymbol{x}_j)$, symmetric in (i,j), so every bracket $[\nabla g_i, \nabla g_j]$ vanishes and $\mathcal{H}$ is involutive wherever it has constant rank, with a Jacobian that moves throughout training. The same holds for every model whose training outputs are linear in a commuting reparametrisation in the sense of Li et al. (2022b), $g_i = \langle \boldsymbol{x}_i, G(\boldsymbol{\theta})\rangle$ with $[\nabla G_k, \nabla G_l] = 0$, since then $[\nabla g_i, \nabla g_j] = \sum_{k,l} x_{ik} x_{jl} [\nabla G_k, \nabla G_l] = 0$; this is the class in which gradient flow is a continuous-time mirror descent, with an explicit potential in the diagonal case (Woodworth et al., 2020; Li et al., 2022b). For outputs non-linear in G the brackets need not vanish, and Assumption 2.1 further restricts the initialisation scale.

Remark A.4 (μP scaling and Assumption 2.1). For MLPs in the abc-parametrisation of Yang & Hu (2021, Table 1), μP differs from the NTK parametrisation by dividing the readout by the square root of the width and multiplying the learning rate of every other tensor by the width; with shared initial Gaussians the learning-rate-weighted Gram matrix $K(\boldsymbol{\theta}_0)$ of Proposition G.1 therefore coincides at every width with the NTK-parametrisation neural tangent kernel (cf. Bordelon & Pehlevan, 2022, Eqs. (14)–(15) and App. N). By Yang (2020a, Cor. 7.3) it converges almost surely, as the width tends to infinity, to a deterministic limit K_{∞}, which is strictly positive definite for non-polynomial activations when biases are present, or without biases when the training inputs are pairwise non-proportional (Carvalho et al., 2025, Thms. 1–2); for networks without biases the distinct inputs of Section 2 must therefore be pairwise non-proportional. A finite-width lower bound on $\lambda_{\min}(K(\boldsymbol{\theta}_0))$ for smooth MLPs of width $\tilde{\Omega}(n)$ (Banerjee et al., 2023, Thm. 4.1) transfers to μP by the same identity (proof of Proposition G.2). Continuity of $\sigma_{\min}$ and $\sigma_{\max}$ in $\boldsymbol{\theta}$ then propagates the spectral bounds to some ball about $\boldsymbol{\theta}_0$. This is motivation only: it supplies no radius, and the radius condition of Assumption 2.1 is not a consequence of it.

For the model of Figure 2, parts (i)–(iii) below verify the *conclusion* of Proposition 2.1 off the diagonals $\theta_1 = \pm\theta_2$, and part (iv) the *hypothesis*, for near-interpolating initialisations.

Proposition A.1 (Closed-form conditioning for the minimal non-linear model)**.** *Let $f(x;\boldsymbol{\theta}) = \theta_1\theta_2 x$ on the single datum $\mathcal{D} = \{(1,1)\}$, so $g(\boldsymbol{\theta}) = \theta_1\theta_2 =: p$, $Y = 1$, $\mathcal{L}(g(\boldsymbol{\theta}), Y) = (p-1)^2$, $J(\boldsymbol{\theta}) = (\theta_2, \theta_1) \in \mathbb{R}^{1\times 2}$ and $K(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2$. Write $\delta := \theta_1^2 - \theta_2^2$ and p_0, δ_0 for the initial values. Along gradient flow from any $\boldsymbol{\theta}_0$ with $\delta_0 \neq 0$:*

- *(i) δ is conserved and $\dot{p} = -2(p-1)\|\boldsymbol{\theta}\|^2$, so $p(t)$ is monotone towards 1 and $|p(t)| \leq \max(|p_0|, 1)$;*
- *(ii) $\|\boldsymbol{\theta}\|^2 = \sqrt{\delta_0^2 + 4p^2}$, and since $n = 1$, $\sigma_{\min}(J(\boldsymbol{\theta})) = \sigma_{\max}(J(\boldsymbol{\theta})) = \|\boldsymbol{\theta}\|$;*
- *(iii) hence, for all $t \in [0,\infty)$,*

$$|\delta_0|^{1/2} \;\leq\; \sigma\big(J(\boldsymbol{\theta}(t))\big) \;\leq\; \big(\delta_0^2 + 4\max(p_0^2, 1)\big)^{1/4},$$

so the conclusion of Proposition 2.1 – uniform conditioning along the entire trajectory – holds with explicit constants, and $\boldsymbol{\theta}^{\star}$ is the point of the hyperbola $\theta_1\theta_2 = 1$ on the branch of $\boldsymbol{\theta}_0$ with $\theta_1^2 - \theta_2^2 = \delta_0$;

(iv) Assumption 2.1 itself is verifiable in closed form: on $B(\boldsymbol{\theta}_0, R)$ with $R < \|\boldsymbol{\theta}_0\|$ one may take $c = \|\boldsymbol{\theta}_0\| - R$ and $C = \|\boldsymbol{\theta}_0\| + R$, and with $R = \|\boldsymbol{\theta}_0\|/2$ the radius condition $R > C|p_0 - 1|/c^2$ is equivalent to

$$|\theta_{0,1}\theta_{0,2} - 1| \;<\; \tfrac{1}{12}\|\boldsymbol{\theta}_0\|^2 .$$

When $\delta_0 = 0$ and $p_0 < 0$ (the anti-diagonal $\theta_1 = -\theta_2$), gradient flow converges to the non-interpolating saddle at the origin; the lower bound in (iii) is then vacuous and Assumption 2.1 fails for every R, as it must.

Proof. The gradient flow is $\dot\theta_1 = -2(p-1)\theta_2$, $\dot\theta_2 = -2(p-1)\theta_1$. *(i)* Direct differentiation gives $\dot\delta = 0$ and $\dot p = -2(p-1)(\theta_1^2 + \theta_2^2) = -2(p-1)\sqrt{\delta_0^2 + 4p^2}$ by (ii); this scalar equation is locally Lipschitz with solution $p \equiv 1$, so $p-1$ keeps its sign and $|p-1|$ decreases. *(ii)* $(\theta_1^2+\theta_2^2)^2 = \delta^2 + 4p^2$, and for $n = 1$ the only singular value of $J(\boldsymbol{\theta})$ is $\|\boldsymbol{\theta}\|$. *(iii)* The bounds follow from (i)–(ii); $\frac{d}{dt}(p-1)^2 \le -4|\delta_0|(p-1)^2$ gives exponential convergence, on the level set $\delta = \delta_0$, which meets $\{p = 1\}$ on the branch of $\boldsymbol{\theta}_0$ in one point. *(iv)* $\|\boldsymbol{\theta}_0\| - R \le \|\boldsymbol{\theta}\| \le \|\boldsymbol{\theta}_0\| + R$ on the ball, and with $R = \|\boldsymbol{\theta}_0\|/2$, $C|p_0 - 1|/c^2 = 6|p_0 - 1|/\|\boldsymbol{\theta}_0\|$. On the anti-diagonal ($\delta \equiv 0$, $p = -\theta_1^2$), $\dot p = 4|p|(1-p) > 0$, so $\boldsymbol{\theta} \to \mathbf{0}$, where $J(\mathbf{0}) = 0$, and no admissible ball can contain the trajectory. □

Proposition A.2 (A sufficient certificate for uniform conditioning on the required ball)**.** *Let $z \in \Theta$, $a := \sigma_{\min}(J(z)) > 0$ and $b := \sigma_{\max}(J(z))$. Choose $\rho > 0$ with $\overline{B}(z,\rho) \subset \Theta$, and suppose an independently verified constant $H \ge 0$ satisfies*

$$\|J(u) - J(v)\|_{\mathrm{op}} \le H\|u - v\| \qquad \text{for all } u, v \in \overline{B}(z,\rho).$$

Define $c_\rho := a - H\rho$ and $C_\rho := b + H\rho$. If

$$c_\rho > 0, \qquad \rho > \frac{C_\rho\|g(z) - Y\|}{c_\rho^2},$$

then Assumption 2.1 holds with initialisation z, radius ρ and constants c_ρ, C_ρ. Exact squared-error gradient flow from z exists for all time, remains in $B(z,\rho)$, and converges to an interpolator, with total path length at most $C_\rho\|g(z) - Y\|/c_\rho^2 < \rho$.

Proof. For every $u \in \overline{B}(z,\rho)$, singular-value perturbation bounds yield

$$\sigma_{\min}(J(u)) \ge a - \|J(u) - J(z)\|_{\mathrm{op}} \ge c_\rho, \qquad \sigma_{\max}(J(u)) \le b + \|J(u) - J(z)\|_{\mathrm{op}} \le C_\rho.$$

These bounds hold uniformly throughout the ball and satisfy the required radius inequality. Proposition 2.1 therefore applies from z. □

A verified bound on $\|DJ\|_{\mathrm{op}}$ throughout the closed ball supplies H by integration along line segments. Smoothness or differences between sampled iterates alone do not verify a usable value of H. Taking $z = \boldsymbol{\theta}_0$ certifies the original assumption. Taking z to be a later exact iterate certifies only the restarted tail; taking z to be a numerical iterate certifies exact flow started there. Identifying that flow with an earlier exact trajectory additionally requires control of the numerical integration error.

A.1 Empirical verification of Assumption 2.1

The ball-local form is what the model of Figure 2 requires: its interpolating set is an unbounded hyperbola along which $\sigma_{\max}(J(\boldsymbol{\theta})) = \|\boldsymbol{\theta}\|$ is unbounded, whereas Assumption 2.1 holds on a ball in closed form (Proposition A.1). Along a computed trajectory, $\lambda_{\min}(K(\boldsymbol{\theta}(t))) = \sigma_{\min}(J(\boldsymbol{\theta}(t)))^2$, $\lambda_{\max}(K(\boldsymbol{\theta}(t)))$ and $\|g(\boldsymbol{\theta}(t)) - Y\|$ are measurable on any finite horizon $[0,T]$, which checks the *conclusion* of Proposition 2.1 on the observed segment; by Corollary D.2 the unobserved tail lies within $C\|g(\boldsymbol{\theta}(T)) - Y\|/c^2$ of $\boldsymbol{\theta}(T)$, a radius decaying as e^{-2c^2T}, so the assumption reduces to conditioning on an exponentially small ball about the last iterate. Controls of $\sigma_{\min}(J(\boldsymbol{\theta}_0))$ at initialisation (Yang, 2020a; Carvalho et al., 2025) and along near-lazy trajectories (Du et al., 2019; Oymak & Soltanolkotabi, 2019; Liu et al., 2022) do not settle the feature-learning case; all our results are conditional on the assumption.

Setup. We train a μP-parametrised (Yang & Hu, 2021; Yang et al., 2021) multi-layer perceptron (MLP) with two hidden layers and Gaussian error linear unit (GELU) activations (Hendrycks & Gimpel, 2016) on the one-dimensional toy regression dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{16}$ with $x_i \in [-4, 4]$ uniformly spaced, $y_i = \sin(\frac{\pi}{2}x_i) + \varepsilon_i$ for $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.2$. Training uses full-batch gradient descent (a discrete-time approximation of gradient flow) with no regularisation (Ridgeless), at hidden widths $d_{\text{model}} \in \{64, 256, 1024, 4096, 16384\}$ across multiple seeds. At each epoch we record $\lambda_{\min}(K(\boldsymbol{\theta}))$, $\lambda_{\max}(K(\boldsymbol{\theta}))$, and the training loss; epochwise quantities are aggregated across seeds with shaded confidence regions denoting standard error.

Results. In Figure 4, (i) $\lambda_{\min}(K(\boldsymbol{\theta}))$ stays bounded away from 0 throughout training, stable across widths and growing to a plateau; (ii) $\lambda_{\max}(K(\boldsymbol{\theta}))$ stays finite, with a width-dependent magnitude that stabilises as $d_{\text{model}} \to \infty$; (iii) the training loss decreases monotonically. (i)–(ii) are the spectral condition of Assumption 2.1 on the observed segment and (iii) is consistent with Proposition 2.1; the tail is covered as above. These observations are consistent with the assumption across widths under μP; they do not prove it, and a verification across architectures and tasks is left to future work.

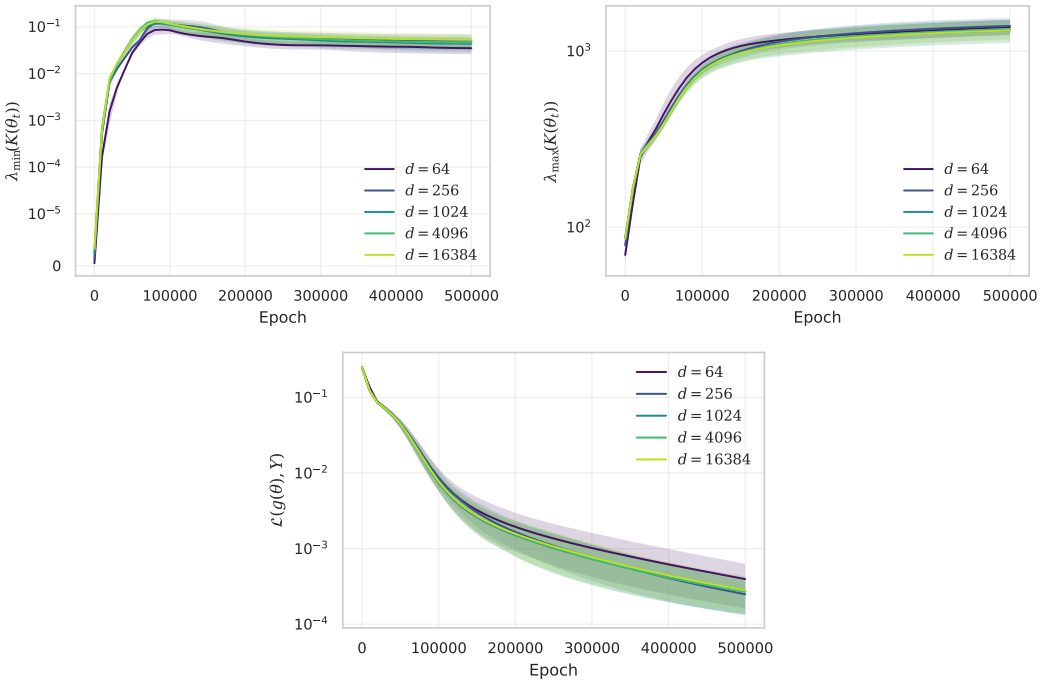


Figure 4: Empirical verification of Assumption 2.1 along gradient-flow trajectories of a two-hidden-layer μP MLP with GELU activations on a 16-point one-dimensional regression toy ($y = \sin(\frac{\pi}{2}x) + \varepsilon$), at widths $d_{\text{model}} \in \{64, 256, 1024, 4096, 16384\}$. *Top*: $\lambda_{\min}(K(\boldsymbol{\theta}))$ (the bound c^2) and $\lambda_{\max}(K(\boldsymbol{\theta}))$ (the bound C^2) over training. *Bottom*: training loss.

B Extended Related Work

Implicit bias and minimum-norm interpolation. The connection between gradient flow and minimum-norm solutions in overparametrised linear models is classical (Engl et al., 1996; Gunasekar et al., 2018): gradient flow from $\boldsymbol{\theta}_0$ on a linear model $g(\boldsymbol{\theta}) = X\boldsymbol{\theta}$ converges to the interpolator closest to $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_0 + X^+(Y - X\boldsymbol{\theta}_0)$, which reduces to the Moore–Penrose pseudoinverse solution X^+Y, the ℓ_2-minimum-norm interpolator, when $\boldsymbol{\theta}_0 = \mathbf{0}$ (more generally, when $\boldsymbol{\theta}_0 \in \operatorname{Im} X^\top$). The same identity holds in kernel regression and in NTK-regime neural networks (Jacot et al., 2018; Lee et al., 2019): the gradient flow limit is the minimum-RKHS-norm interpolator, and the implicit prior $p(\boldsymbol{c}) \propto \exp(-\beta \|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\|^2_{\text{RKHS}})$ is the Gaussian-process prior with NTK covariance kernel, $\mathcal{GP}(f(\cdot; \boldsymbol{\theta}_0), (2\beta)^{-1} k_{\text{NTK}})$ (Rasmussen & Williams, 2006; He et al., 2020; Calvo-Ordoñez

et al., 2026). The *kernel vs. rich* dichotomy of Woodworth et al. (2020) delineates when networks behave like linear models. We identify the *energy* underlying this selection and generalise it to the feature-learning regime, where the minimum-norm interpolator and the gradient-flow limit diverge (Theorem D.11). The selection principle behind Theorem D.11(a)–(b) is the vanishing-Tikhonov principle of nonlinear inverse problems (Engl et al., 1989; 1996), and the alignment condition is the range (source) condition of that literature, evaluated at the gradient-flow limit; its positive side, that a horizontal range fixed on a ball returns gradient descent to the $\boldsymbol{\theta}_0$-minimum-norm solution, is the null-space condition of nonlinear Landweber theory (Hanke et al., 1995), of which Corollary D.6 is the constant-Jacobian case (trajectory-wise fixity does not suffice, Remark D.8). What is new is the evaluation at the gradient-flow limit, the rotation bound (Lemma D.5), the witnesses of Remark D.7, and the extension to fixed-anchor penalties (Proposition D.13). That no data-independent function of the parameters captures the implicit bias of gradient descent in several non-linear models is known (Razin & Cohen, 2020; Vardi & Shamir, 2021); our positive claim escapes those results because geodesic ridge depends on the training problem through the flow manifold, as the training-cost function depends on the inputs through the reachable manifold. The slow drift along the interpolating set under a small explicit penalty (Section 3) is the slow-manifold dynamics of Katzenberger (1991), developed for the manifold of minimisers by Li et al. (2022a) (Theorem D.12). Closest to Theorem D.12, Boursier et al. (2025) prove, as a theory of grokking, that gradient flow with small weight decay on a general loss first interpolates and then follows the Riemannian gradient flow of $\|\boldsymbol{\theta}\|^2$ on the manifold of minimisers (concurrently, Musat, 2025); their result covers weight decay only, assumes a Morse–Bott manifold of minimisers and gives no rates, whereas Theorem D.12 and Corollary D.7 treat an arbitrary C^2 regulariser on the ball with an explicit remainder, which is what yields the drift criterion and its reading as the alignment condition for anchored ridge.

Implicit bias in classification and structured models. For classification with the logistic loss on linearly separable data, gradient descent converges in direction to the max-margin solution, that of a hard-margin support-vector machine (SVM) (Soudry et al., 2018; Ji & Telgarsky, 2019); for homogeneous networks this extends to a characterisation of the margin through the Karush–Kuhn–Tucker (KKT) conditions (Lyu & Li, 2020); and for wide two-layer networks trained with the logistic loss, Chizat & Bach (2020) characterise the infinite-width dynamics as a Wasserstein gradient flow over the distribution of neurons, whose limit is a max-margin classifier for a function-space norm. For matrix factorisation, gradient descent was conjectured to induce an implicit nuclear-norm penalty (Gunasekar et al., 2017). There the interpolating set is empty and the object characterised is a direction or a limiting classifier; our framework presupposes an interpolating set that gradient flow reaches (Proposition 2.1), so these results lie outside our scope, and whether geodesic ridge has a counterpart for losses without an interpolating set is open.

Feature-learning theory. The mean-field limit (Chizat & Bach, 2018; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018) describes the infinite-particle dynamics of two-layer networks in a regime where the Jacobian varies substantially, complementing the NTK picture. The Tensor Programs framework of Yang & Hu (2021) provides a unified characterisation of wide networks via a dynamical dichotomy between kernel and feature-learning regimes, with μP scaling (Yang & Hu, 2021; Yang et al., 2021) as the canonical parametrisation enabling stable feature learning. Recent work has analysed what gradient descent actually *learns* in the feature-learning regime: Damian et al. (2022) prove that gradient descent can learn representations unavailable to fixed-kernel methods, and Lauditi et al. (2025) characterise the adaptive kernel that emerges at the feature-learning infinite-width limit. Closest to our geometric setting, Marcotte et al. (2023) characterise the conservation laws of gradient flow, for every dataset, through the Lie algebra generated by the Jacobian vector fields over all inputs, the maximal number of conserved quantities being attained only when that algebra closes, and Marcotte et al. (2026) show that a Frobenius property of the analogous distribution spanned over all inputs is what allows training to be reduced to an intrinsic Riemannian gradient flow on a lower-dimensional space (their Def. 2.20, Prop. 2.21 and Thm. 2.17); the integrable class of Remark A.1 is the case of maximal conservation, in which the flow manifold is label-free; in general it is the fibre of the training-limit map through $\boldsymbol{\theta}_0$ (Theorem D.3). We characterise the canonical *regulariser* rather than what the dynamics converge to. Diagonal linear networks (Woodworth et al., 2020) and commuting reparametrisations (Li et al., 2022b) lie inside the integrable class (Remark A.3(v)); how geodesic ridge relates to their mirror-descent potentials at finite regularisation strength is open.

Weight decay beyond RKHS regularisation. Grokking, delayed generalisation long after the data loss reaches zero, was documented by Power et al. (2022) and analysed for ridge regression by Xu et al. (2026); the effect of weight decay on learned features is analysed by Lauditi et al. (2025). Theorem D.11 and Remark D.11 add that in the feature-learning regime standard and anchored ridge select interpolators systematically different from the gradient-flow limit: their effect is *solution-set modification* rather than *prior specification*.

Geometry-aware optimisers. Natural gradient descent (Amari, 1998) preconditions the gradient by the pseudo-inverse of the Fisher information matrix (Martens, 2020; Ollivier, 2015); for Gaussian regression the parameter-space Fisher matrix is proportional to $J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta})$, of rank n, and preconditioning by its pseudo-inverse gives $\dot{\boldsymbol{\theta}} = -J(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{c}} \mathcal{L}$, that is, the function-space dynamics preconditioned by $K(\boldsymbol{\theta})^{-1}$. The function-space dynamics then become isotropic, which yields fast convergence (Zhang et al., 2019; Cai et al., 2019; Karakida & Osawa, 2020), and in the integrable class the trajectories stay on the flow manifold; recent optimisers implement related principles at scale (Jordan et al., 2024; Bernstein & Newhouse, 2024). We characterise a regulariser, not an optimiser; preconditioning gradient flow by the inverse NTK straightens the kernel-regime trajectory into a geodesic of the affine flow manifold, so that arc ridge, anchored ridge, and geodesic ridge coincide exactly (Proposition E.9); its development beyond the kernel regime we leave to future work.

Regularisation for fine-tuning. Anchored ridge on the pretrained weights, with standard ridge on the new head, is the standard explicit regulariser for fine-tuning, L2-SP (Li et al., 2018); its relatives replace the Euclidean anchor by a Fisher-weighted one, as in elastic weight consolidation (Kirkpatrick et al., 2017), whose purpose is to retain a previous task rather than to recover the new task's gradient-flow limit, or impose a distance from the pretrained weights as a constraint (Gouk et al., 2021). The penalties among them are quadratic and anchored at a fixed point, so Proposition D.13 applies to each: outside the kernel regime the vanishing-regularisation limit is the interpolator nearest the anchor in the chosen metric, which is the gradient-flow limit only under the corresponding alignment condition. Function-space regularisers penalise the change in the network's outputs rather than in its weights (Benjamin et al., 2019; Rudner et al., 2023); restricted to the training inputs they are the output-shaped penalties of Proposition D.14, which select an entire output fibre and cannot single out the gradient-flow limit; evaluated on held-out inputs, as they are in practice, they penalise motion along the output fibres through a data-dependent proxy for the fibre distance that geodesic ridge measures intrinsically (Definition 4.1), and a comparison of the two is open. Arc ridge differs from all of these in being a functional of the training path rather than of its endpoint.

C Preliminary Lemmas

The standing assumptions of Appendix A hold throughout.

C.1 Gradient flow basics

Lemma C.1 (Gradient formula)**.** *For any loss with output gradient* $\boldsymbol{u} := -\partial_{\boldsymbol{c}} \mathcal{L}(\boldsymbol{c}, Y)|_{\boldsymbol{c}=g(\boldsymbol{\theta})}$, $\nabla_{\boldsymbol{\theta}} \mathcal{L}(g(\boldsymbol{\theta}), Y) = -J(\boldsymbol{\theta})^\top \boldsymbol{u}$. *The SSE special case* $\mathcal{L} = \|g(\boldsymbol{\theta}) - Y\|^2$ *gives* $\boldsymbol{u} = 2(Y - g(\boldsymbol{\theta}))$ *and* $\nabla_{\boldsymbol{\theta}} \mathcal{L} = 2J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y)$.

Proof. Chain rule applied to $\mathcal{L}(g(\cdot), Y)$. □

Lemma C.2 (Lyapunov decrease)**.** *Along any gradient flow solution* $\boldsymbol{\theta}(t)$, $\frac{d}{dt}\mathcal{L}(g(\boldsymbol{\theta}(t)), Y) = -\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(g(\boldsymbol{\theta}(t)), Y)\|^2 \leq 0$, *with equality if and only if* $\nabla_{\boldsymbol{\theta}} \mathcal{L} = 0$.

Proof. $\frac{d}{dt}\mathcal{L} = \langle \nabla \mathcal{L}, \dot{\boldsymbol{\theta}} \rangle = -\|\nabla \mathcal{L}\|^2$ by A.3. □

Lemma C.3 (Critical points in the ball are interpolators)**.** *Under Assumptions 2.1 and A.3, for every* $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R)$, $\nabla_{\boldsymbol{\theta}} \mathcal{L}(g(\boldsymbol{\theta}), Y) = 0$ *if and only if* $g(\boldsymbol{\theta}) = Y$. *Consequently, the critical set of* $\mathcal{L}$ *in* $B(\boldsymbol{\theta}_0, R)$ *is* $\Theta^\star \cap B(\boldsymbol{\theta}_0, R)$.

Proof. By Lemma C.1, $\nabla_{\boldsymbol{\theta}}\mathcal{L} = -J(\boldsymbol{\theta})^\top \boldsymbol{u}$. If $g(\boldsymbol{\theta}) = Y$ then $\boldsymbol{u} = -\partial_{\boldsymbol{c}}\mathcal{L}|_Y = 0$ by A.3 (loss minimum at Y), so $\nabla_{\boldsymbol{\theta}}\mathcal{L} = 0$. Conversely, if $J(\boldsymbol{\theta})^\top \boldsymbol{u} = 0$, left-multiplying by $J(\boldsymbol{\theta})$ gives $K(\boldsymbol{\theta})\boldsymbol{u} = 0$; on $B(\boldsymbol{\theta}_0, R)$ Assumption 2.1 makes $K(\boldsymbol{\theta}) \succeq c^2 I_n$ invertible, so $\boldsymbol{u} = 0$, and strict convexity of $\mathcal{L}(\cdot, Y)$ in $\boldsymbol{c}$ (A.3) with minimum at Y forces $g(\boldsymbol{\theta}) = Y$. Outside the ball non-interpolating critical points can exist: for the model of Proposition A.1 the origin is one. □

C.2 Structural lemmas

Lemma C.4 (Output fibres are embedded submanifolds)**.** *Under Assumption 2.1, for every $\boldsymbol{c} \in g(B(\boldsymbol{\theta}_0, R))$ the output fibre $\Theta(\boldsymbol{c}) \cap B(\boldsymbol{\theta}_0, R) = g|^{-1}_{B(\boldsymbol{\theta}_0,R)}(\boldsymbol{c})$ is an embedded C^∞ submanifold of $B(\boldsymbol{\theta}_0, R)$ of codimension n.*

Proof. $\boldsymbol{c}$ is a regular value of $g|_{B(\boldsymbol{\theta}_0,R)}$ since $J(\boldsymbol{\theta})$ has rank n on $B(\boldsymbol{\theta}_0, R)$ by Assumption 2.1. The regular-value theorem (Lee, 2012, Ch. 5) then gives the conclusion. □

C.3 Geometric identities

Lemma C.5 (Infinitesimal orthogonality)**.** *At every $\boldsymbol{\theta}$ with full-rank $J(\boldsymbol{\theta})$, $\mathrm{Im}(J(\boldsymbol{\theta})^\top) \perp \ker J(\boldsymbol{\theta})$ in the ambient Euclidean inner product on $\mathbb{R}^m$, and $\mathbb{R}^m = \mathrm{Im}(J(\boldsymbol{\theta})^\top) \oplus \ker J(\boldsymbol{\theta})$ as an orthogonal decomposition.*

Proof. If $\boldsymbol{v} \in \mathrm{Im}(J(\boldsymbol{\theta})^\top)$ and $\boldsymbol{w} \in \ker J(\boldsymbol{\theta})$, write $\boldsymbol{v} = J(\boldsymbol{\theta})^\top \boldsymbol{a}$; then $\boldsymbol{v}^\top \boldsymbol{w} = \boldsymbol{a}^\top J(\boldsymbol{\theta})\boldsymbol{w} = 0$. Dimension count: $\dim \mathrm{Im}(J(\boldsymbol{\theta})^\top) = \mathrm{rank}\, J(\boldsymbol{\theta}) = n$ and $\dim \ker J(\boldsymbol{\theta}) = m - n$, summing to m. □

Lemma C.6 (Geodesic length dominates ambient chord)**.** *For any embedded C^1 submanifold $M \subseteq \mathbb{R}^m$ and any two points $p, q \in M$ in the same path component, the intrinsic geodesic distance $d_M(p, q)$ (defined as infimum of arc lengths of C^1 curves in M) satisfies $d_M(p, q) \geq \|p - q\|$.*

Proof. For any C^1 curve $\gamma : [0, 1] \to M$ with $\gamma(0) = p$, $\gamma(1) = q$, $\mathrm{Length}(\gamma) = \int_0^1 \|\dot\gamma(t)\|\, dt \geq \left\| \int_0^1 \dot\gamma(t)\, dt \right\| = \|q - p\|$ by the triangle inequality for integrals. Taking the infimum over γ preserves the inequality. □

Lemma C.7 (Energy–length identity)**.** *Let Γ be any non-empty family of piecewise-C^1 paths $[0, 1] \to \mathbb{R}^m$ with common endpoints, closed under increasing piecewise-C^1 time reparametrisation. Then*

$$\inf_{\gamma\in\Gamma} \int_0^1 \|\dot\gamma(t)\|^2\, dt \;=\; \Big(\inf_{\gamma\in\Gamma} \int_0^1 \|\dot\gamma(t)\|\, dt\Big)^2,$$

with the left infimum attained exactly at constant-speed length-minimisers whenever such minimisers exist.

Proof. Let $\gamma \in \Gamma$ have length $L := \int_0^1 \|\dot\gamma(t)\|\, dt$. By Cauchy–Schwarz,

$$L^2 = \left(\int_0^1 \|\dot\gamma(t)\| \cdot 1\, dt \right)^2 \leq \int_0^1 \|\dot\gamma(t)\|^2\, dt \cdot \int_0^1 1^2\, dt = \mathcal{E}(\gamma),$$

with equality if and only if $\|\dot\gamma(t)\|$ is constant in t; taking infima over Γ gives $\inf_\Gamma \mathcal{E} \geq (\inf_\Gamma L)^2$. Conversely, fix $\gamma \in \Gamma$ of length L and $\varepsilon > 0$, and let $\phi(t) := \int_0^t \big(\|\dot\gamma(s)\| + \varepsilon\big)\, ds \,/\, (L + \varepsilon)$: an increasing piecewise-C^1 bijection of $[0, 1]$ with $\phi' \geq \varepsilon/(L + \varepsilon) > 0$, so $\gamma \circ \phi^{-1} \in \Gamma$ with speed $(L + \varepsilon)\, \|\dot\gamma\| / (\|\dot\gamma\| + \varepsilon) \leq L + \varepsilon$, hence energy at most $(L + \varepsilon)^2$. Letting $\varepsilon \downarrow 0$ and taking infima over γ gives $\inf_\Gamma \mathcal{E} \leq (\inf_\Gamma L)^2$. Attainment at a constant-speed length-minimiser, when one exists, is the equality case above. □

Lemma C.8 (Intrinsic distance of nearby points on a curved surface)**.** *Let $M \subseteq \mathbb{R}^m$ be an embedded C^2 submanifold and $K \subseteq M$ a compact set of reach at least $\rho > 0$ in $\mathbb{R}^m$ (Federer, 1959), so that*

every point of $\mathbb{R}^m$ within distance ρ of K has a unique nearest point in K. Then for all $p, q \in K$ with $\delta := \|p - q\| < \rho$,

$$\delta \;\le\; d_M(p,q) \;\le\; \delta\left(1 + \delta^2/\rho^2\right);$$

in particular $d_M(p,q) = \delta\,(1 + O(\delta^2))$ with a constant depending on ρ only.

Proof. The lower bound is Lemma C.6. For the upper bound write ξ for the nearest-point map onto K and $\operatorname{dist}(\cdot, K)$ for the Euclidean distance to K. Two properties of sets of positive reach are used (Federer, 1959, §4): for $0 \le h < \rho$ the map ξ is Lipschitz with constant $\rho/(\rho - h)$ on $\{x : \operatorname{dist}(x, K) \le h\}$; and for $a, b \in K$ and every x with $\xi(x) = a$,

$$\langle b - a,\, x - a\rangle \;\le\; \frac{\|x - a\|\,\|b - a\|^2}{2\rho},$$

because the open ball of radius ρ tangent to K at a in the direction $x - a$ contains no point of K. Let $x_t := (1-t)p + tq$ for $t \in [0,1]$, $d_t := \operatorname{dist}(x_t, K) \le \min(t, 1-t)\,\delta < \rho$, and $a_t := \xi(x_t)$. Since $x_t - a_t = (1-t)(p - a_t) + t(q - a_t)$, the second property with $b = p$ and $b = q$ gives

$$d_t^2 \;=\; (1-t)\langle p - a_t,\, x_t - a_t\rangle + t\langle q - a_t,\, x_t - a_t\rangle \;\le\; \frac{d_t}{2\rho}\Big((1-t)\,\|p - a_t\|^2 + t\,\|q - a_t\|^2\Big),$$

and $\|p - a_t\| \le \|p - x_t\| + d_t \le 2t\delta$, $\|q - a_t\| \le 2(1-t)\delta$, so $d_t \le 2t(1-t)\delta^2/\rho \le \delta^2/(2\rho) =: h$. The chord therefore lies in $\{x : \operatorname{dist}(x, K) \le h\}$ with $h < \rho/2$, and $t \mapsto \xi(x_t)$ is a Lipschitz curve in $K \subseteq M$ from p to q of length at most $\delta\rho/(\rho - h) = \delta/(1 - \delta^2/2\rho^2) \le \delta(1 + \delta^2/\rho^2)$, the last step using $\delta \le \rho$. The infimum defining $d_M(p,q)$ is unchanged if Lipschitz curves in M are admitted, since such a curve is approximated in length by piecewise-C^1 curves in M (piecewise-linear interpolation in charts of M), whose corners are removed by reparametrising to zero speed (Burago et al., 2001). □

C.4 Optimisation lemma

Lemma C.9 (Vanishing-regularisation selection principle)**.** *Let $\mathcal{L} : \Theta \to [0, \infty)$ be continuous with non-empty $\arg\min \mathcal{L}$, and let $R : \Theta \to [0, \infty]$ be continuous and finite at some point of $\arg\min \mathcal{L}$. For each $\lambda > 0$ let $\hat{\boldsymbol{\theta}}_\lambda \in \arg\min(\mathcal{L} + \lambda R)$ be any minimiser of the regularised problem, and assume that every sequence $\{\hat{\boldsymbol{\theta}}_{\lambda_k}\}$ with $\lambda_k \downarrow 0$ has a convergent subsequence. Then every limit point of $\hat{\boldsymbol{\theta}}_\lambda$ as $\lambda \downarrow 0$ lies in $\arg\min \mathcal{L}$ and, among $\arg\min \mathcal{L}$, minimises R.*

Proof. Let $\bar{\boldsymbol{\theta}} \in \arg\min \mathcal{L}$ with $R(\bar{\boldsymbol{\theta}}) < \infty$ (for the other points of $\arg\min \mathcal{L}$ the final inequality is trivial) and set $\mathcal{L}^\star := \min \mathcal{L}$. By optimality, $\mathcal{L}(\hat{\boldsymbol{\theta}}_\lambda) + \lambda R(\hat{\boldsymbol{\theta}}_\lambda) \le \mathcal{L}^\star + \lambda R(\bar{\boldsymbol{\theta}})$, so $\mathcal{L}(\hat{\boldsymbol{\theta}}_\lambda) \le \mathcal{L}^\star + \lambda R(\bar{\boldsymbol{\theta}})$ and $R(\hat{\boldsymbol{\theta}}_\lambda) \le (\mathcal{L}^\star - \mathcal{L}(\hat{\boldsymbol{\theta}}_\lambda))/\lambda + R(\bar{\boldsymbol{\theta}}) \le R(\bar{\boldsymbol{\theta}})$. Taking $\lambda \downarrow 0$ along a convergent subsequence with limit $\boldsymbol{\theta}^\infty$, continuity yields $\mathcal{L}(\boldsymbol{\theta}^\infty) \le \mathcal{L}^\star$, so $\boldsymbol{\theta}^\infty \in \arg\min \mathcal{L}$. Passing to the limit in $R(\hat{\boldsymbol{\theta}}_\lambda) \le R(\bar{\boldsymbol{\theta}})$ gives $R(\boldsymbol{\theta}^\infty) \le R(\bar{\boldsymbol{\theta}})$ for every $\bar{\boldsymbol{\theta}} \in \arg\min \mathcal{L}$; hence $\boldsymbol{\theta}^\infty$ minimises R over $\arg\min \mathcal{L}$. □

D Formal Results and Proofs

Formal statements and proofs follow in main-text order, for Sections 2, 3.1, 3 and 4; the proof of Theorem 4.3 is in Appendix E.

D.1 Convergence of gradient flow under uniform Jacobian conditioning

Proof of Proposition 2.1. Write $r(t) := g(\boldsymbol{\theta}(t)) - Y$, $r_0 := g(\boldsymbol{\theta}_0) - Y$ and $\ell := C\,\|r_0\|/c^2 < R$; the argument is a bootstrap.

Step 1. By A.1 the field $-2J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y)$ is locally Lipschitz on Θ, so there is a unique maximal solution on $[0, T_{\max})$, which leaves every compact subset of Θ if $T_{\max} < \infty$ (Khalil, 2002, Ch. 3). *Step 2.* Let $T_{\text{exit}} > 0$ be the supremum of the times up to which $\boldsymbol{\theta}$ stays in $B(\boldsymbol{\theta}_0, R)$. *Step 3.* For $t < T_{\text{exit}}$, $\dot r = -2K(\boldsymbol{\theta})\,r$ with $K(\boldsymbol{\theta}) \succeq c^2 I_n$ gives $\frac{d}{dt}\|r\|^2 \le -4c^2\|r\|^2$, so $\|r(t)\| \le \|r_0\|\,e^{-2c^2 t}$ by Grönwall (Khalil, 2002), $\left\|\dot{\boldsymbol{\theta}}\right\| = 2\left\|J(\boldsymbol{\theta})^\top r\right\| \le 2C\,\|r_0\|\,e^{-2c^2 t}$, and $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\| \le \int_0^t \left\|\dot{\boldsymbol{\theta}}\right\| \le$

$\ell(1-e^{-2c^2t}) < \ell$. *Step 4.* If $T_{\text{exit}} < T_{\max}$, continuity gives $\|\boldsymbol{\theta}(T_{\text{exit}}) - \boldsymbol{\theta}_0\| \le \ell < R$, and openness of the ball contradicts maximality of T_{exit}. *Step 5.* If $T_{\max} < \infty$, the solution stays in the compact $\overline{B}(\boldsymbol{\theta}_0, \ell) \subset \Theta$, contradicting Step 1. So the bounds of Step 3 hold for all $t \ge 0$, and integrating gives the path-length bound. *Step 6.* $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(s)\| \le \ell e^{-2c^2 s}$ for $s \le t$, so $\boldsymbol{\theta}^\star := \lim \boldsymbol{\theta}(t)$ exists, $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\| \le \ell < R$, and $g(\boldsymbol{\theta}^\star) = Y$; $B(\boldsymbol{\theta}^\star, R-\ell) \subset B(\boldsymbol{\theta}_0, R)$, so $\boldsymbol{\theta}^\star$ is interior to the ball. □

Three consequences of Proposition 2.1 follow; throughout, $r(t) := g(\boldsymbol{\theta}(t)) - Y$, $r_0 := g(\boldsymbol{\theta}_0) - Y$, $\ell := C\|r_0\|/c^2$, and $s(T) := \int_0^T \left\|\dot{\boldsymbol{\theta}}(t)\right\| dt$ is the Euclidean length of the training path up to time T.

Corollary D.1 (A-priori path length; arc ridge is bounded)**.** *Under Assumption 2.1, the training path γ_{gf} has finite Euclidean length*

$$s(\infty) \;=\; \int_0^\infty \left\|\dot{\boldsymbol{\theta}}(t)\right\| dt \;\le\; \ell \;=\; \frac{C\|r_0\|}{c^2} \;<\; R,$$

so for every $T \ge 0$ the arc-ridge regulariser of Definition 4.2 satisfies $L^2(\gamma_{\text{gf}}|_{[0,T]}) = s(T)^2 \le \ell^2$, and, whenever $\boldsymbol{\theta}(T) \in V$, by Corollary E.1

$$\|\boldsymbol{\theta}(T) - \boldsymbol{\theta}_0\|^2 \;\le\; d^2(\boldsymbol{\theta}(T), \boldsymbol{\theta}_0) \;\le\; s(T)^2 \;\le\; \ell^2.$$

Moreover, for every $\lambda > 0$ every minimiser $\boldsymbol{\theta}_\lambda$ of $\mathcal{L}(g(\cdot), Y) + \lambda\|\cdot - \boldsymbol{\theta}_0\|^2$ over Θ satisfies $\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}_0\| \le \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\| \le \ell$, so all anchored-ridge minimisers and all their $\lambda \downarrow 0$ limit points lie in $\overline{B}(\boldsymbol{\theta}_0, \ell) \subset B(\boldsymbol{\theta}_0, R)$.

Proof. The length bound is the display of Proposition 2.1, and $s(T) \le s(\infty)$. The anchored-ridge bound compares the objective at $\boldsymbol{\theta}_\lambda$ and at $\boldsymbol{\theta}^\star$: $\mathcal{L}(g(\boldsymbol{\theta}_\lambda), Y) + \lambda\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}_0\|^2 \le 0 + \lambda\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\|^2$ and $\mathcal{L}(\cdot, Y) \ge 0$, while $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\| \le s(\infty) \le \ell$; limit points of a set inside the closed ball $\overline{B}(\boldsymbol{\theta}_0, \ell)$ lie in it. □

Corollary D.2 (Finite-horizon tail bound)**.** *Under Assumption 2.1, for every $T \ge 0$,*

$$\int_T^\infty \left\|\dot{\boldsymbol{\theta}}(t)\right\| dt \;\le\; \frac{C\|r(T)\|}{c^2} \;\le\; \ell\, e^{-2c^2T}, \qquad \textit{hence} \qquad \boldsymbol{\theta}^\star \in \overline{B}\Big(\boldsymbol{\theta}(T),\, \tfrac{C\|r(T)\|}{c^2}\Big).$$

Consequently the assumption is checkable from a finite observation window: if spectral bounds $c_T \le \sigma_{\min}(J(\boldsymbol{\theta})) \le \sigma_{\max}(J(\boldsymbol{\theta})) \le C_T$ are verified on a ball $B(\boldsymbol{\theta}(T), R_T)$ with $R_T > C_T\|r(T)\|/c_T^2$, then Proposition 2.1 applied from the initial point $\boldsymbol{\theta}(T)$ confines the unobserved tail $\{\boldsymbol{\theta}(t) : t \ge T\}$ to $\overline{B}(\boldsymbol{\theta}(T), C_T\|r(T)\|/c_T^2)$.

Proof. Step 3 of the proof of Proposition 2.1, run from time T, gives the first bound, and $\|r(T)\| \le \|r_0\|\, e^{-2c^2T}$ the second; the inclusion follows from $\boldsymbol{\theta}^\star - \boldsymbol{\theta}(T) = \int_T^\infty \dot{\boldsymbol{\theta}}$, and the restart statement is Proposition 2.1 from $\boldsymbol{\theta}(T)$, the flow being autonomous. □

Remark D.1 (General strictly convex losses). Let $h(\boldsymbol{c}) := \mathcal{L}(\boldsymbol{c}, Y)$ be μ-strongly convex and L_h-smooth on the sublevel set $\{h \le h(g(\boldsymbol{\theta}_0))\}$, with minimiser Y and $h^\star := h(Y)$, and set $\Delta_0 := h(g(\boldsymbol{\theta}_0)) - h^\star$. Replacing the residual equation by

$$\frac{d}{dt}\big(h - h^\star\big) \;=\; -\left\|J(\boldsymbol{\theta})^\top \nabla h\right\|^2 \;\le\; -c^2\|\nabla h\|^2 \;\le\; -c^2\mu^2\|r\|^2 \;\le\; -\frac{2c^2\mu^2}{L_h}\big(h - h^\star\big),$$

the same first-exit argument gives, while $\boldsymbol{\theta}(t) \in B(\boldsymbol{\theta}_0, R)$, $h(t) - h^\star \le \Delta_0 e^{-2c^2\mu^2 t/L_h}$, $\|r(t)\| \le \sqrt{2\Delta_0/\mu}\, e^{-c^2\mu^2 t/L_h}$, $\left\|\dot{\boldsymbol{\theta}}\right\| \le CL_h\|r\|$, and

$$\int_0^\infty \left\|\dot{\boldsymbol{\theta}}\right\| dt \;\le\; \ell_{\text{gen}} \;:=\; \frac{CL_h^2}{c^2\mu^2}\sqrt{\frac{2\Delta_0}{\mu}}.$$

Assumption 2.1 then reads $R > \ell_{\text{gen}}$; squared error ($\mu = L_h = 2$, $\Delta_0 = \|r_0\|^2$) recovers $\ell = C\|r_0\|/c^2$ and the rate $2c^2$ for $\|r\|$, and Proposition 2.1 and Corollaries D.1 and D.2 hold verbatim with ℓ_{gen} in place of ℓ, $c^2\mu^2/L_h$ in place of $2c^2$, and $\mathcal{L}(\cdot, Y) - h^\star$ in place of $\mathcal{L}(\cdot, Y)$ in the anchored-ridge comparison.

Proof. Lemma C.2 gives the equality, $\sigma_{\min}(J(\boldsymbol{\theta})) \geq c$ the first inequality, and $\mu \|r\| \leq \|\nabla h\| \leq L_h \|r\|$, $\frac{\mu}{2}\|r\|^2 \leq h - h^\star \leq \frac{L_h}{2}\|r\|^2$ (strong convexity and smoothness) the rest; Grönwall gives the decay of $h - h^\star$, hence of $\|r\|$, and $\left\|\dot{\boldsymbol{\theta}}\right\| \leq C L_h \|r\|$ integrates to ℓ_{gen}. Steps 2, 4, 5, and 6 of the proof of Proposition 2.1 use only these three bounds and the radius condition, so they repeat with ℓ_{gen} in place of ℓ; the anchored-ridge comparison of Corollary D.1 uses $\mathcal{L}(g(\boldsymbol{\theta}^\star), Y) - h^\star = 0$ and $\mathcal{L}(\cdot, Y) - h^\star \geq 0$. □

D.2 Structure of the flow manifold

Throughout this subsection A.1–A.3 and Assumption 2.1 hold, $B := B(\boldsymbol{\theta}_0, R)$, and $X := -\nabla_{\boldsymbol{\theta}}\mathcal{L}$, so that $X(\boldsymbol{\theta}) = -2J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y)$ and ϕ_t denotes its flow. Three facts about $\Theta^\star \cap B$ are used: it is an embedded C^∞ submanifold of codimension n with $T_{\boldsymbol{\theta}}(\Theta^\star \cap B) = \ker J(\boldsymbol{\theta})$ (Lemma C.4); every point of it is an equilibrium of X; and $DX(\boldsymbol{\theta}) = -2J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta})$ there, with spectrum $\{0\}$ on $\ker J(\boldsymbol{\theta})$ and n eigenvalues in $[-2C^2, -2c^2]$ on $\mathrm{Im}\, J(\boldsymbol{\theta})^\top$ (A.3, A.4). A manifold of equilibria whose normal linearisation is uniformly contracting is *normally attracting*; the theory of normally hyperbolic invariant manifolds supplies, near any compact piece of it, a smooth foliation by the sets of points converging to each equilibrium. We state the form of that theory we use, then derive the flow manifold from it. Neither object is new: the training-limit map π of Theorem D.2 is the limit map Φ of Li et al. (2022a), with $d\pi = I - P$ on $\Theta^\star$ as there, its smoothness near a normally attracting manifold of minimisers goes back to Falconer (1983), and the flow manifold $\pi^{-1}(\boldsymbol{\theta}^\star)$ is the stable fibre of $\boldsymbol{\theta}^\star$ in Fenichel's theory, the set of points with asymptotic phase $\boldsymbol{\theta}^\star$. What is new is their use as the carrier of a regulariser.

Lemma D.1 (Stable fibration of a normally attracting manifold of equilibria)**.** *Let X be a C^∞ vector field on an open set of $\mathbb{R}^m$ and $\Sigma \subset \{X = 0\}$ an embedded C^∞ submanifold of dimension $m - n$ such that at every $p \in \Sigma$ the spectrum of $DX(p)$ is $\{0\}$ on $T_p\Sigma$ and lies in $\{\mathrm{Re} \leq -\mu\}$, $\mu > 0$, on a complementary n-dimensional invariant subspace E^s_p. Let $K \subset \Sigma$ be compact and $r \in \mathbb{N}$. There are an open neighbourhood U' of K, forward-invariant under the flow ϕ_t of X and contained in any prescribed neighbourhood of K, and a C^r map $\pi_{\mathrm{loc}} : U' \to \Sigma \cap U'$ such that:*

- *(i)* $\pi_{\mathrm{loc}} \circ \phi_t = \pi_{\mathrm{loc}}$ *on U' for all $t \geq 0$, and $\pi_{\mathrm{loc}}|_{\Sigma \cap U'} = \mathrm{id}$;*
- *(ii)* π_{loc} *is a submersion, and each fibre $W^s_{\mathrm{loc}}(p) := \pi^{-1}_{\mathrm{loc}}(p)$, $p \in \Sigma \cap U'$, is an embedded C^r n-disc tangent at p to E^s_p;*
- *(iii)* $W^s_{\mathrm{loc}}(p) = \{\boldsymbol{\theta} \in U' : \phi_t(\boldsymbol{\theta}) \to p\}$ *for every $p \in \Sigma \cap U'$, and for every $\mu' < \mu$ there is K_0 with $\|\phi_t(\boldsymbol{\theta}) - p\| \leq K_0 e^{-\mu' t}\|\boldsymbol{\theta} - p\|$ and $\|D\phi_t(\boldsymbol{\theta})\boldsymbol{v}\| \leq K_0 e^{-\mu' t}\|\boldsymbol{v}\|$ for $\boldsymbol{\theta} \in W^s_{\mathrm{loc}}(p)$, $\boldsymbol{v} \in T_{\boldsymbol{\theta}} W^s_{\mathrm{loc}}(p)$, $t \geq 0$;*
- *(iv)* $D\pi_{\mathrm{loc}}(p)$ *is the projection onto $T_p\Sigma$ along E^s_p.*

Proof. The statement is the stable-fibration theorem, with asymptotic phase, for a normally hyperbolic invariant manifold whose unstable bundle is trivial, specialised to a manifold of equilibria: persistence and smoothness of the invariant manifold and of its stable manifold are due to Fenichel (1971), the asymptotic phase and its rate conditions to Fenichel (1974; 1977), the compact case to Hirsch, Pugh and Shub (1977); the form used in geometric singular perturbation theory, for a critical manifold of equilibria, is Fenichel (1979) and Jones (1995) at $\varepsilon = 0$, and manifolds of equilibria are treated as such by Aulbach (1984). In the form we use it: if X_1 is C^{r+1} and K' a compact submanifold with boundary, overflowing invariant under X_1, whose normal contraction rate μ exceeds $r + 1$ times the largest tangential Lyapunov exponent, then a neighbourhood of K' is fibred by C^{r+1} discs $W^s(p)$ tangent at p to the stable subspace, consisting of the points whose trajectories approach that of p at every rate $\mu' < \mu$, and depending C^r on p.

Rate and boundary. On Σ the flow is the identity, so every tangential exponent is 0 and the rate condition holds for every r (the neighbourhood may depend on r; Theorem D.2 removes this for the limit map). K is invariant but not overflowing invariant. Choose a compact $K' \subset \Sigma$ with boundary and $K \subset \mathrm{int}_\Sigma K'$, a field ζ on Σ vanishing on a neighbourhood K'' of K and pointing outward along $\partial K'$, extended constantly along the normal discs of a tubular neighbourhood with coordinates $(p, \boldsymbol{w})$, $\boldsymbol{w} \in E^s_p$. For $X_\eta := X + \eta\zeta$, K' is overflowing invariant, the normal linearisation is that of X, and the tangential exponents are $O(\eta)$; for small η the theorem applies to X_η.

Transfer to X. Points of K'' are equilibria of X_η. In disc-adapted coordinates the flow reads $\dot{p} = 0$, $\dot{\boldsymbol{w}} = F(p, \boldsymbol{w})$ with $F(p, \mathbf{0}) = \mathbf{0}$, $D_{\boldsymbol{w}} F(p, \mathbf{0}) = DX(p)|_{E_p^s}$; a continuous Lyapunov inner product on E^s and continuity of $D_{\boldsymbol{w}} F$ give $\delta > 0$ with $\frac{d}{dt} \|\boldsymbol{w}\|_p \leq -\mu' \|\boldsymbol{w}\|_p$ and the same for the variational equation whenever $\|\boldsymbol{w}\|_p < \delta$. Then $U' := \{(p, \boldsymbol{w}) : p \in \operatorname{int}_\Sigma K'', \ \|\boldsymbol{w}\|_p < \delta\}$ is forward-invariant, can be made to lie in any prescribed neighbourhood of K and in $\{\zeta = 0\}$, where $X_\eta = X$; set $W^s_{\mathrm{loc}}(p) := W^s_\eta(p) \cap U'$ and $\pi_{\mathrm{loc}}(p, \boldsymbol{w}) := p$.

Properties. (i) is $\dot{p} = 0$; (ii) holds for a coordinate projection whose fibres are the discs; (iii) is the asymptotic-phase statement with the two Lyapunov estimates, K_0 comparing $\|\cdot\|_p$ with the Euclidean norm uniformly on K''; (iv) $D\pi_{\mathrm{loc}}(p)$ is the identity on $T_p\Sigma$ by (i) and vanishes on $E^s_p = T_p W^s_{\mathrm{loc}}(p)$ by (ii). □

Proposition D.1 (Trapping region)**.** *Let*

$$\mathfrak{B} \ := \ \left\{\boldsymbol{\theta} \in B : \ \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + C \left\|g\left(\boldsymbol{\theta}\right) - Y\right\| / c^2 < R\right\}.$$

Then $\mathfrak{B}$ is open, contains $\boldsymbol{\theta}_0$ and $\Theta^\star \cap B$, and is forward-invariant: for every $\boldsymbol{\theta} \in \mathfrak{B}$ the gradient flow $\phi_t(\boldsymbol{\theta})$ exists for all $t \geq 0$, stays in $\mathfrak{B}$, has $\|g\left(\phi_t(\boldsymbol{\theta})\right) - Y\| \leq \|g\left(\boldsymbol{\theta}\right) - Y\| e^{-2c^2 t}$, total length at most $C \|g\left(\boldsymbol{\theta}\right) - Y\| / c^2$, and converges to a point $\pi\left(\boldsymbol{\theta}\right) \in \Theta^\star \cap B$ with $\|\pi\left(\boldsymbol{\theta}\right) - \boldsymbol{\theta}\| \leq C \|g\left(\boldsymbol{\theta}\right) - Y\| / c^2$; $\pi\left(\cdot\right)$ is the identity on $\Theta^\star \cap B$ and $\pi\left(\phi_t(\boldsymbol{\theta})\right) = \pi\left(\boldsymbol{\theta}\right)$ for all $t \geq 0$.

Proof. $\mathfrak{B}$ is open by continuity, contains $\boldsymbol{\theta}_0$ by the radius condition and $\Theta^\star \cap B$ trivially. For $\boldsymbol{\theta} \in \mathfrak{B}$ and $r(t) := g\left(\phi_t(\boldsymbol{\theta})\right) - Y$, while $\phi_s(\boldsymbol{\theta}) \in B$ the spectral bounds give $\|r(s)\| \leq \|r(0)\| e^{-2c^2 s}$ and $\left\|\dot{\boldsymbol{\theta}}\right\| \leq 2C \|r\| \leq -\frac{C}{c^2} \frac{d}{ds} \|r\|$, so

$$\|\phi_t(\boldsymbol{\theta}) - \boldsymbol{\theta}_0\| + \frac{C \|r(t)\|}{c^2} \ \leq \ \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \frac{C}{c^2}\big(\|r(0)\| - \|r(t)\|\big) + \frac{C \|r(t)\|}{c^2} \ < \ R :$$

the defining quantity of $\mathfrak{B}$ does not increase. Steps 2–5 of the proof of Proposition 2.1 then keep the trajectory in $\mathfrak{B}$ for all $t \geq 0$; finite length gives the limit $\pi\left(\boldsymbol{\theta}\right)$ with the stated bound, $g\left(\pi\left(\boldsymbol{\theta}\right)\right) = Y$, and the strict inequality places it in B. Points of $\Theta^\star \cap B$ are equilibria, and $\pi\left(\phi_t(\boldsymbol{\theta})\right) = \pi\left(\boldsymbol{\theta}\right)$ is immediate. □

Theorem D.2 (The training-limit map is a smooth submersion)**.** *The training-limit map $\pi\left(\cdot\right) : \mathfrak{B} \to \Theta^\star \cap B$, $\pi\left(\boldsymbol{\theta}\right) := \lim_{t \to \infty} \phi_t(\boldsymbol{\theta})$, is a C^∞ surjective submersion with $\pi\left(\cdot\right) \circ \phi_t = \pi\left(\cdot\right)$ for all $t \geq 0$, $d\pi(\boldsymbol{\theta}) X(\boldsymbol{\theta}) = \mathbf{0}$ on $\mathfrak{B}$, and*

$$d\pi(p) \ = \ I - P_p, \qquad P_p := J\left(p\right)^\top K\left(p\right)^{-1} J\left(p\right), \qquad p \in \Theta^\star \cap B.$$

The rank-n distribution $\ker d\pi$ *on $\mathfrak{B}$ is integrable; its maximal connected integral manifolds are the components of the level sets of π, and $X(\boldsymbol{\theta}) \in \ker d\pi(\boldsymbol{\theta})$ at every point.*

Proof. The setting of Lemma D.1. Take $\Sigma := \Theta^\star \cap B$, an embedded C^∞ submanifold of dimension $m - n$ consisting of equilibria of X (Lemma C.4), and at $p \in \Sigma$ take $E^s_p := \operatorname{Im} J\left(p\right)^\top$. Since $DX(p) = -2J\left(p\right)^\top J\left(p\right)$, the subspace $\ker J\left(p\right) = T_p\Sigma$ is its kernel, and E^s_p is invariant, $DX(p) J\left(p\right)^\top \boldsymbol{a} = -2J\left(p\right)^\top K\left(p\right) \boldsymbol{a}$, with the eigenvalues $-2\sigma_i(J\left(p\right))^2 \in [-2C^2, -2c^2]$; so the hypotheses hold with $\mu := 2c^2$.

Smoothness and rank. Fix $r \in \mathbb{N}$ and $\boldsymbol{\theta} \in \mathfrak{B}$, and put $p := \pi\left(\boldsymbol{\theta}\right) \in \Theta^\star \cap B$ (Proposition D.1). Choose R' with $\|p - \boldsymbol{\theta}_0\| < R' < R$ and apply Lemma D.1 to the compact piece $K := \Theta^\star \cap \overline{B}(\boldsymbol{\theta}_0, R')$, obtaining U'_r and π_{loc}; we take U'_r inside the tube $\{\boldsymbol{\theta} : \operatorname{dist}(\boldsymbol{\theta}, K) < \delta'\}$ with $\delta' < (R - R')/(1 + C^2/c^2)$, which lies in $\mathfrak{B}$ because $\|g\left(\boldsymbol{\theta}\right) - Y\| \leq C \|\boldsymbol{\theta} - q\|$ for the nearest $q \in K$. By (iii) of the lemma $\pi = \pi_{\mathrm{loc}}$ on U'_r. Choose T with $\phi_T(\boldsymbol{\theta}) \in U'_r$; by continuity of ϕ_T the same T serves on a neighbourhood N of $\boldsymbol{\theta}$, so $\pi = \pi_{\mathrm{loc}} \circ \phi_T$ on N, a C^r submersion since $D\phi_T$ is invertible. As r is arbitrary, π is a C^∞ submersion; it is surjective, being the identity on $\Theta^\star \cap B$.

Invariance and differential. Differentiating $\pi\left(\phi_t(\boldsymbol{\theta})\right) = \pi\left(\boldsymbol{\theta}\right)$ at $t = 0$ gives $d\pi(\boldsymbol{\theta}) X(\boldsymbol{\theta}) = \mathbf{0}$. At $p \in \Theta^\star \cap B$, (iv) of the lemma makes $d\pi(p)$ the projection onto $\ker J\left(p\right)$ along $\operatorname{Im} J\left(p\right)^\top$, orthogonal by Lemma C.5, so $d\pi(p) = I - P_p$. The level sets of a submersion are embedded submanifolds with tangent spaces $\ker d\pi$ (Lee, 2012, Ch. 5), which gives the statement on the distribution. □

Theorem D.3 (The basin manifold)**.** *Let $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) := \pi^{-1}(\boldsymbol{\theta}^\star)$. Then:*

- *(i) $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is a closed, connected, embedded C^∞ n-dimensional submanifold of $\mathfrak{B}$ containing $\boldsymbol{\theta}_0$, the trajectory $\{\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t) : t \geq 0\}$ and $\boldsymbol{\theta}^\star$; it is forward-invariant, and invariant under the backward flow within $\mathfrak{B}$.*
- *(ii) $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \ker d\pi(\boldsymbol{\theta})$ at every $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, $T_{\boldsymbol{\theta}^\star}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \mathrm{Im}\, J(\boldsymbol{\theta}^\star)^\top$, and $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap \mathrm{Im}\, J(\boldsymbol{\theta})^\top$ for every $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$.*
- *(iii) $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap \Theta^\star = \{\boldsymbol{\theta}^\star\}$, and $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \mathfrak{B} : \lim_{t\to\infty} \phi_t(\boldsymbol{\theta}) = \boldsymbol{\theta}^\star\}$ is the set of initialisations in $\mathfrak{B}$ with the same training outcome as $\boldsymbol{\theta}_0$.*
- *(iv) (Uniqueness.) If $N \subseteq \mathfrak{B}$ is a connected embedded C^1 n-dimensional submanifold, closed in $\mathfrak{B}$, forward-invariant, with $N \cap \Theta^\star = \{\boldsymbol{\theta}^\star\}$, then $N = \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$.*
- *(v) $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ depends on Y, and on the loss form h, through the flow and its limit, and on nothing else.*

Proof. (i) $\boldsymbol{\theta}^\star$ is a regular value of the submersion π (Theorem D.2), so $\pi^{-1}(\boldsymbol{\theta}^\star)$ is a closed embedded n-submanifold of $\mathfrak{B}$ (Lee, 2012, Ch. 5); it contains $\boldsymbol{\theta}_0$, the trajectory and $\boldsymbol{\theta}^\star$, and both invariances follow from $\pi \circ \phi_t = \pi$. Connectedness: the disc $W^s_{\mathrm{loc}}(\boldsymbol{\theta}^\star)$ of Lemma D.1 is connected and lies in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, and every $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is joined to it by its forward trajectory.

(ii) Level sets of a submersion have tangent space $\ker d\pi$, which at $\boldsymbol{\theta}^\star$ is $\ker(I - P_{\boldsymbol{\theta}^\star}) = \mathrm{Im}\, J(\boldsymbol{\theta}^\star)^\top$; $X \in \ker d\pi$ by Theorem D.2 and $X \in \mathrm{Im}\, J(\boldsymbol{\theta})^\top$ by Lemma C.1.

(iii) π is the identity on $\Theta^\star \cap B$; the second statement is the definition of π.

(iv) Let $\boldsymbol{\theta} \in N$. Its forward trajectory stays in N (forward invariance) and converges to $\pi(\boldsymbol{\theta}) \in \Theta^\star \cap B \subseteq \mathfrak{B}$ (Proposition D.1); N is closed in $\mathfrak{B}$, so $\pi(\boldsymbol{\theta}) \in N \cap \Theta^\star = \{\boldsymbol{\theta}^\star\}$, that is, $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. Hence $N \subseteq \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. By invariance of domain the inclusion $N \hookrightarrow \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ of n-manifolds is open; it is closed and non-empty, and $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is connected, so $N = \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$.

(v) The construction uses only $X = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(g(\boldsymbol{\theta}), Y)$ (Lemma C.1, Remark D.1). □

In (iv), closedness cannot be dropped (an open subset of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ containing the trajectory and $\boldsymbol{\theta}^\star$ satisfies the rest), and the competitor need only be C^1, with no relation to $\mathrm{Im}\, J(\cdot)^\top$: the flow manifold is determined by invariance, dimension n, and having the trained network as its only interpolator, the last being where the labels enter.

Dependence on the conditioning ball. The construction uses a triple (R, c, C) of Assumption 2.1, and $\mathfrak{B}$ depends on it. For two admissible triples with trapping regions $\mathfrak{B}$ and $\mathfrak{B}'$, π is the same function on $\mathfrak{B} \cup \mathfrak{B}'$, the limit of the flow, and a C^∞ submersion there (Theorem D.2 on each), so the two flow manifolds are open subsets of the one embedded manifold $\pi^{-1}(\boldsymbol{\theta}^\star) \subset \mathfrak{B} \cup \mathfrak{B}'$ and coincide on $\mathfrak{B} \cap \mathfrak{B}'$, an open set containing the trajectory and $\boldsymbol{\theta}^\star$; the chart V may be taken inside it, so the representative map and the fibre term do not depend on the triple. The flow term is an infimum over paths in the whole flow manifold and can depend on it: if $\mathfrak{B} \subseteq \mathfrak{B}'$ it can only decrease. It does not near the trajectory when the radius condition holds with room: a path in the flow manifold from $\boldsymbol{\theta}_0$ of length $L < R$ lies in B, where the defining function of $\mathfrak{B}$ is at most $(1 + C^2/c^2)L + \ell$, $\ell := C\,\|g(\boldsymbol{\theta}_0) - Y\|/c^2$, as in the proof of Theorem F.3(i); the paths of length at most L therefore lie in both trapping regions when $(1 + C^2/c^2)L + \ell < R$ for both triples, and $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ is then the same for both wherever it is below L. Since $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot) \leq \ell$ along the trajectory, geodesic ridge is independent of the triple on the trajectory and near $\boldsymbol{\theta}^\star$ whenever both triples satisfy $(2 + C^2/c^2)\hat{\ell} < R$, $\hat{\ell}$ the larger of their two values of ℓ, the stronger radius condition of Theorem F.3; under the bare radius condition it is defined relative to the chosen ball.

What the tangent space of the flow manifold is away from $\boldsymbol{\theta}^\star$, where it need not be horizontal, is the content of the next lemma.

Lemma D.2 (The tangent space is the space of forgotten perturbations)**.** *For $\boldsymbol{\theta} \in \mathfrak{B}$ and $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{v} \in \ker d\pi(\boldsymbol{\theta})$, the tangent space at $\boldsymbol{\theta}$ of the level set of π through $\boldsymbol{\theta}$ (of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ when $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$), if and only if $D\phi_t(\boldsymbol{\theta})\boldsymbol{v} \to \mathbf{0}$ as $t \to \infty$; and for every $\boldsymbol{v}$, $D\phi_t(\boldsymbol{\theta})\boldsymbol{v} \to d\pi(\boldsymbol{\theta})\boldsymbol{v}$. For $\boldsymbol{\theta}$ in a neighbourhood U' supplied by Lemma D.1 the decay of $D\phi_t(\boldsymbol{\theta})\boldsymbol{v}$, $\boldsymbol{v} \in \ker d\pi(\boldsymbol{\theta})$, is exponential at every rate $\mu' < 2c^2$.*

Proof. Differentiating $\pi \circ \phi_t = \pi$ gives $d\pi(\phi_t(\boldsymbol{\theta}))\, D\phi_t(\boldsymbol{\theta}) = d\pi(\boldsymbol{\theta})$ for all $t \geq 0$. Suppose first that $\boldsymbol{\theta} \in U'$, the neighbourhood of Lemma D.1 constructed in the proof of Theorem D.2 about $p := \pi(\boldsymbol{\theta})$, on which $\pi = \pi_{\mathrm{loc}}$. In the coordinates $(p, \boldsymbol{w})$ of that proof the flow has the form $\dot{p} = 0$, $\dot{\boldsymbol{w}} = F(p, \boldsymbol{w})$ with $F(p, \mathbf{0}) = \mathbf{0}$, so $D_p F(p, \mathbf{0}) = 0$ and, taking $r \geq 2$, $\|D_p F(p, \boldsymbol{w})\| \leq L \|\boldsymbol{w}\|$ on U' for some L. Write $\boldsymbol{v} = (\delta p, \delta \boldsymbol{w})$ and $D\phi_t(\boldsymbol{\theta})\boldsymbol{v} = (\delta p, \delta \boldsymbol{w}_t)$, since the p-coordinate is constant along the flow. The fibre coordinate obeys the variational equation $\frac{d}{dt}\delta \boldsymbol{w}_t = D_{\boldsymbol{w}} F(p, \boldsymbol{w}_t)\delta \boldsymbol{w}_t + D_p F(p, \boldsymbol{w}_t)\delta p$, whose homogeneous part contracts at rate μ' in the Lyapunov norm ((iii) of the lemma) and whose source is bounded by $LK_0 e^{-\mu' t} \|\boldsymbol{w}_0\| \, \|\delta p\|$; variation of constants gives $\|\delta \boldsymbol{w}_t\| \leq K_0 e^{-\mu' t} \|\delta \boldsymbol{w}\| + K_0^2 L\, t\, e^{-\mu' t} \|\boldsymbol{w}_0\| \, \|\delta p\| \to 0$. Hence $D\phi_t(\boldsymbol{\theta})\boldsymbol{v} \to (\delta p, \mathbf{0}) = d\pi_{\mathrm{loc}}(\boldsymbol{\theta})\boldsymbol{v}$, in coordinates and therefore in $\mathbb{R}^m$ (the chart is C^1 and $\phi_t(\boldsymbol{\theta}) \to p$), which vanishes if and only if $\delta p = \mathbf{0}$, that is, if and only if $\boldsymbol{v} \in \ker d\pi_{\mathrm{loc}}(\boldsymbol{\theta}) = T_{\boldsymbol{\theta}} W^s_{\mathrm{loc}}(p)$; when $\delta p = \mathbf{0}$ the decay is at rate μ' by the first term alone. For general $\boldsymbol{\theta} \in \mathfrak{B}$ choose T with $\phi_T(\boldsymbol{\theta}) \in U'$ and apply the case just proved to $\boldsymbol{v}' := D\phi_T(\boldsymbol{\theta})\boldsymbol{v}$ at $\phi_T(\boldsymbol{\theta})$: $D\phi_t(\boldsymbol{\theta})\boldsymbol{v} = D\phi_{t-T}(\phi_T(\boldsymbol{\theta}))\boldsymbol{v}' \to d\pi(\phi_T(\boldsymbol{\theta}))\boldsymbol{v}' = d\pi(\boldsymbol{\theta})\boldsymbol{v}$, and $D\phi_T(\boldsymbol{\theta})$ is invertible, so $\boldsymbol{v}' \in \ker d\pi(\phi_T(\boldsymbol{\theta}))$ if and only if $\boldsymbol{v} \in \ker d\pi(\boldsymbol{\theta})$. □

The lemma gives the tangent space of the flow manifold its operational meaning: $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is the null space of the sensitivity of the trained network to a perturbation of the current iterate, the perturbations training forgets, and along the trajectory $T_{\boldsymbol{\theta}(t)}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = D\phi_t(\boldsymbol{\theta}_0)\, T_{\boldsymbol{\theta}_0}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. A horizontal perturbation $J(\boldsymbol{\theta})^\top \boldsymbol{u}$ at $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$ is remembered in general, $d\pi(\boldsymbol{\theta}) J(\boldsymbol{\theta})^\top \boldsymbol{u} \neq \mathbf{0}$: the directions in which gradient flow for some label vector would move next are not the directions the present training problem forgets, and the two spaces agree at $\boldsymbol{\theta}^\star$ (Theorem D.3(ii)), along the trajectory only in the direction $X(\boldsymbol{\theta})$, and everywhere only in the integrable class (Theorem D.9).

Corollary D.3 (Kernel regime by evaluation)**.** *If $J(\boldsymbol{\theta}) = J$ on B, then $\pi(\boldsymbol{\theta}) = \boldsymbol{\theta} - J^+(g(\boldsymbol{\theta}) - Y)$ on $\mathfrak{B}$, $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = (\boldsymbol{\theta}_0 + \operatorname{Im} J^\top) \cap \mathfrak{B}$ is convex, $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im} J^\top$ everywhere, and its geodesics are chords. If $n = 1$, $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is the arc of the gradient line of g through $\boldsymbol{\theta}_0$ within $\mathfrak{B}$, and $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im} J(\cdot)^\top$ everywhere.*

Proof. Constant J. On the convex ball B, $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, and $X(\boldsymbol{\theta}) = -2J^\top(g(\boldsymbol{\theta}) - Y) \in \operatorname{Im} J^\top$, so every plane $\boldsymbol{\theta} + \operatorname{Im} J^\top$ is invariant. For $\boldsymbol{\theta} \in \mathfrak{B}$ the trajectory stays in $\mathfrak{B}$ and converges (Proposition D.1); along it $r(t) := g(\phi_t(\boldsymbol{\theta})) - Y$ obeys $\dot{r} = -2Kr$, $K := JJ^\top$, so $r(t) = e^{-2Kt} r(0)$ and

$$\begin{aligned} \phi_t(\boldsymbol{\theta}) \;&=\; \boldsymbol{\theta} - 2J^\top \int_0^t e^{-2Ks}\, ds\; r(0) \;=\; \boldsymbol{\theta} - J^\top K^{-1}\big(I - e^{-2Kt}\big) r(0) \\ &\longrightarrow\; \boldsymbol{\theta} - J^\top K^{-1}(g(\boldsymbol{\theta}) - Y), \end{aligned}$$

which is the displayed formula, $J^+ = J^\top K^{-1}$ being the pseudoinverse of a matrix of full row rank. Since $g(\boldsymbol{\theta}) - Y = J(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$, it reads $\pi(\boldsymbol{\theta}) = \boldsymbol{\theta}^\star + (I - P)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$ with $P = J^\top K^{-1} J$, so $\pi(\boldsymbol{\theta}) = \boldsymbol{\theta}^\star$ if and only if $\boldsymbol{\theta} - \boldsymbol{\theta}^\star \in \operatorname{Im} J^\top$; as $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 = -J^+(g(\boldsymbol{\theta}_0) - Y) \in \operatorname{Im} J^\top$, $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = (\boldsymbol{\theta}_0 + \operatorname{Im} J^\top) \cap \mathfrak{B}$. The function defining $\mathfrak{B}$ is convex when g is affine, so $\mathfrak{B}$ is a convex open set, its intersection with an affine subspace is convex, the induced metric is flat, and the geodesics are the chords; $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im} J^\top$ is immediate.

One output. For $n = 1$, $\operatorname{Im} J(\boldsymbol{\theta})^\top = \mathbb{R}\, \nabla g(\boldsymbol{\theta})$ with $\|\nabla g\| = \sigma_{\min}(J(\boldsymbol{\theta})) \geq c$ on B. Let $\Gamma : (a, b) \to \mathfrak{B}$ be the maximal integral curve in $\mathfrak{B}$ of the unit field $\nabla g / \|\nabla g\|$ with $\Gamma(0) = \boldsymbol{\theta}_0$. Along it $\frac{d}{ds} g(\Gamma(s)) \geq c$ and g is bounded on $\overline{B}$, so $b - a < \infty$ and, by maximality, Γ converges to points outside $\mathfrak{B}$ at both ends; hence Γ is closed in $\mathfrak{B}$ and, $g \circ \Gamma$ being strictly increasing, a connected embedded curve. The training field $-2(g - Y)\nabla g$ is tangent to it, so Γ is forward-invariant, and strict monotonicity of g gives $\Gamma \cap \Theta^\star = \{\boldsymbol{\theta}^\star\}$. Theorem D.3(iv) gives $\Gamma = \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. □

The basin manifold is horizontal at $\boldsymbol{\theta}^\star$ and transverse to the output fibres there; near $\boldsymbol{\theta}^\star$ this makes g a chart on it, which is the representative map of the main text.

Theorem D.4 (Local manifold chart)**.** *Under Assumption 2.1, there exists an open neighbourhood $V \subseteq B$ of $\boldsymbol{\theta}^\star$ such that:*

(i) $\Theta^\star \cap V$ is an embedded C^∞ codimension-n submanifold of V with $T_{\boldsymbol{\theta}}(\Theta^\star \cap V) = \ker J(\boldsymbol{\theta})$ at every $\boldsymbol{\theta} \in \Theta^\star \cap V$;

(ii) $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ *is an embedded* C^∞ n*-submanifold of* V *tangent at* $\boldsymbol{\theta}^\star$ *to* $\operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$*, and transversal to* $\ker J(\boldsymbol{\theta})$ *at every one of its points; the basin manifolds of the other interpolators* $p \in \Theta^\star \cap V$ *vary* C^∞*-smoothly with* p*;*

(iii) the restriction $g|_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)\cap V}$ *is a* C^∞ *diffeomorphism onto its open image* $C_0(\boldsymbol{\theta}_0) := g(\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V) \subseteq \mathbb{R}^n$*, an open ball centred at* Y*, with inverse the representative map* $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$*.*

Proof of Theorem D.4. Conclusion (i) is Lemma C.4; the first part of (ii) is Theorem D.3(i),(ii), and the smooth dependence on p is that of the level sets of the C^∞ submersion π (Theorem D.2), the basin manifold of p near p being the disc $W^s_{\text{loc}}(p)$ of Lemma D.1(iii). The restriction of $J(\boldsymbol{\theta})$ to $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is an isomorphism iff $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap \ker J(\boldsymbol{\theta}) = \{\mathbf{0}\}$ (transversality). At $\boldsymbol{\theta}^\star$ this holds by Theorem D.3(ii) and Lemma C.5, and it is an open condition, both subspaces varying continuously; shrinking V gives (ii). By the inverse function theorem, shrinking V again so that g is injective on the component of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ through $\boldsymbol{\theta}^\star$ with image an open ball about Y, and intersecting V with an open set cutting out that component, gives (iii) with $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ connected. The trajectory lies in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ for all large t. □

Henceforth V is as constructed in this proof, so that $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ is connected.

Corollary D.4 (Unique fibre–flow intersection in V)**.** *Under Assumption 2.1, for every* $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0)$ *the intersection* $(\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V) \cap \Theta(\boldsymbol{c})$ *consists of exactly one point, which we denote* $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$*. In particular* $\boldsymbol{\theta}^\star = g|_{\boldsymbol{\theta}_0}^{-1}(Y)$ *is the unique interpolator on* $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$*, and* $\lim_{t\to\infty} \phi_{\text{gf}}(\boldsymbol{\theta}_0, t) = \boldsymbol{\theta}^\star$*. The intersection is transversal at every point and orthogonal only where* $T\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im}(J(\cdot)^\top)$*.*

Proof of Corollary D.4. $g|_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)\cap V}$ is a bijection onto $C_0(\boldsymbol{\theta}_0)$ (Theorem D.4); the tail of the trajectory lies in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ and converges to an interpolator, which must be $g|_{\boldsymbol{\theta}_0}^{-1}(Y)$. Transversality is Theorem D.4(ii), and orthogonality to $\ker J(\cdot)$ holds exactly where $T\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im}(J(\cdot)^\top)$ (Lemma C.5). □

The chart $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ pulls the induced metric of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ back to $C_0(\boldsymbol{\theta}_0)$. In the integrable class the pullback is the inverse NTK (Lemma D.4); in general it exceeds it by a term measuring the tilt of the basin manifold out of $\mathcal{H}$:

Theorem D.5 (Metric gap of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ from the horizontal metric)**.** *Let* $A(\boldsymbol{c}) := Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ *and write* $A(\boldsymbol{c}) = A_\parallel(\boldsymbol{c}) + N(\boldsymbol{c})$ *with* $A_\parallel(\boldsymbol{c}) := J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^\top K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1}$ *and* $N(\boldsymbol{c}) := (I - P_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})})\,A(\boldsymbol{c})$*, where* $P_{\boldsymbol{\theta}}$ *is the orthogonal projector onto* $\operatorname{Im}(J(\boldsymbol{\theta})^\top)$*. Then the pullback metric* $G(\boldsymbol{c}) := A(\boldsymbol{c})^\top A(\boldsymbol{c})$ *on* $C_0(\boldsymbol{\theta}_0)$ *admits the orthogonal decomposition*

$$G(\boldsymbol{c}) \;=\; K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1} \;+\; N(\boldsymbol{c})^\top N(\boldsymbol{c}),$$

so $G \succeq K^{-1}$ *pointwise, with equality* $G(\boldsymbol{c}) = K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1}$ *if and only if* $T_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im}(J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^\top)$ *(the kernel-regime alignment). Moreover, along the realised gradient flow trajectory* $\boldsymbol{\theta}(t) = g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}(t))$*, the velocity* $\dot{\boldsymbol{c}}(t)$ *satisfies* $N(\boldsymbol{c}(t))\,\dot{\boldsymbol{c}}(t) = 0$*, so the trajectory's* G*-energy reduces to* $\dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta}(t))^{-1}\,\dot{\boldsymbol{c}}$*.*

Proof of Theorem D.5. Differentiating $g \circ g|_{\boldsymbol{\theta}_0}^{-1}(\cdot) = \mathrm{id}$ gives $JA = I_n$, where J, K, P are evaluated at $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$. Since $JP = J$, $JN = 0$, so N ranges in $\ker J$; and $PA = J^\top B$ with $KB = JPA = I_n$, so $PA = J^\top K^{-1} = A_\parallel$. By Lemma C.5 the cross terms of $A^\top A$ vanish and $A_\parallel^\top A_\parallel = K^{-1}$, which gives the decomposition; equality holds iff $N = 0$, iff $\operatorname{Im} A \subseteq \operatorname{Im} J^\top$.

Along the trajectory, $A(\boldsymbol{c}(t))\dot{\boldsymbol{c}}(t) = \dot{\boldsymbol{\theta}}(t) \in \operatorname{Im}(J(\boldsymbol{\theta}(t))^\top)$ (A.3), so its vertical part $N\dot{\boldsymbol{c}}$ vanishes and the G-energy is $\dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta}(t))^{-1}\,\dot{\boldsymbol{c}}$. □

$N^\top N = G - K^{-1}$, with $\operatorname{Im} N(\boldsymbol{c}) = (I - P_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})})\, T_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})} \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, measures the tilt of the basin manifold out of the horizontal space; Lemma D.3 bounds it along the trajectory and Proposition D.7 computes its source at $\boldsymbol{\theta}^\star$. Its relation to involutivity is weaker than an equivalence.

Proposition D.6 (What the metric gap measures)**.** *Under Assumption 2.1, with V as in Theorem D.4 and N as in Theorem D.5:*

- *(i) $N \equiv 0$ on $C_0(\boldsymbol{\theta}_0)$ if and only if $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ is an integral manifold of $\mathcal{H}$, that is, $T_{\boldsymbol{\theta}} \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \mathcal{H}_{\boldsymbol{\theta}}$ at every one of its points.*
- *(ii) If $\mathcal{H}$ is integrable on a neighbourhood of $\boldsymbol{\theta}^\star$, then, after shrinking V, $N \equiv 0$ on $C_0(\boldsymbol{\theta}_0)$ and $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ is the plaque of the leaf of $\mathcal{H}$ through $\boldsymbol{\theta}^\star$.*
- *(iii) If $N \equiv 0$ on $C_0(\boldsymbol{\theta}_0)$, then $[X, Z]_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}$ for all local sections X, Z of $\mathcal{H}$ and all $\boldsymbol{\theta} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$: $\mathcal{H}$ is involutive along $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$. It need not be involutive on any neighbourhood of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ (Remark D.2).*

Proof. (i) is the equality case of Theorem D.5. (ii) The local arguments of parts (ii) and (iv) of the proof of Theorem D.9, applied in a flat chart about $\boldsymbol{\theta}^\star$, give a plaque $P_0 \ni \boldsymbol{\theta}^\star$ meeting $\Theta^\star$ only at $\boldsymbol{\theta}^\star$ and whose points flow to $\boldsymbol{\theta}^\star$ within it; so P_0 is an open subset of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.3(iii)), and shrinking V gives the claim with (i). (iii) Sections of $\mathcal{H}$ are tangent to the integral manifold $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$, hence so are their brackets (Lee, 2012, Ch. 8); Remark D.2 is the counterexample to the converse of (ii). □

Remark D.2 (An integral surface without involutivity nearby). Let $g(\boldsymbol{\theta}) = (\theta_1 + \theta_3^2\theta_2,\ \theta_2 + \theta_3^2\theta_1)$ on $\mathbb{R}^3$. With $\xi := \nabla g_1 \times \nabla g_2 = (2\theta_1\theta_3^3 - 2\theta_2\theta_3,\ 2\theta_2\theta_3^3 - 2\theta_1\theta_3,\ 1 - \theta_3^4)^\top$ spanning $\ker J(\boldsymbol{\theta})$ where it is non-zero,

$$[\nabla g_1, \nabla g_2] \;=\; (4\theta_2\theta_3^2,\ -4\theta_1\theta_3^2,\ 0)^\top, \qquad \xi^\top [\nabla g_1, \nabla g_2] \;=\; 8\,\theta_3^3\,(\theta_1^2 - \theta_2^2),$$

which vanishes off the plane $\{\theta_3 = 0\}$ only on the nowhere-dense set $\{\theta_1^2 = \theta_2^2\}$, so $\mathcal{H}$ is non-involutive on every open set not contained in the plane $\{\theta_3 = 0\}$. On that plane $\partial_3 g_i = 0$ and $\xi = e_3$: it is an integral surface of $\mathcal{H}$, and horizontal curves starting in it stay in it ($\dot\theta_3 = 2\theta_3(\theta_2 u_1 + \theta_1 u_2)$). For $\boldsymbol{\theta}_0$ in the plane, the trajectory and $\boldsymbol{\theta}^\star = (Y, 0)$ lie in it, $g = (\theta_1, \theta_2)$ on it, and $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ is a piece of it: $N \equiv 0$, although no two-dimensional integral surface passes through any point off the plane with $\theta_1^2 \neq \theta_2^2$. This is why Remark A.1 requires integrability on the whole ball: involutivity along $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ or the trajectory does not make the flow manifold label-independent.

Lemma D.3 (The tilt of the basin manifold vanishes along the trajectory)**.** *Let $M_2 := \sup_B \big(\sum_i \|\nabla^2 g_i\|_{\text{op}}^2\big)^{1/2}$, let M_P be a Lipschitz constant of $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ on B, and let K_0, μ' be the constants of Lemma D.1(iii) for a neighbourhood U' of $\boldsymbol{\theta}^\star$ containing $\boldsymbol{\theta}(t_1)$. For $t \geq t_1$ the tilt*

$$\omega(t) \;:=\; \sup\big\{ \big\|(I - P_{\boldsymbol{\theta}(t)})\boldsymbol{v}\big\| :\ \boldsymbol{v} \in T_{\boldsymbol{\theta}(t)} \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0),\ \|\boldsymbol{v}\| = 1 \big\}$$

satisfies

$$\omega(t) \;\leq\; \frac{2K_0(M_2 + CM_P)}{\mu'}\, \|r(t)\| \;\leq\; \frac{2K_0(M_2 + CM_P)}{\mu'}\, \|r_0\|\, e^{-2c^2 t},$$

and consequently $G(\boldsymbol{c}(t)) = K(\boldsymbol{\theta}(t))^{-1} + O(\|r(t)\|^2)$ along the trajectory tail.

Proof. Both constants are finite by A.1 and Assumption 2.1 on the compact $\overline{B}$. Every $\boldsymbol{v} \in T_{\boldsymbol{\theta}(t)} \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is $\boldsymbol{v}(t) := D\phi_t(\boldsymbol{\theta}_0)\boldsymbol{v}_0$ for some $\boldsymbol{v}_0 \in T_{\boldsymbol{\theta}_0} \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (ϕ_t is a diffeomorphism onto its image preserving $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$), and $\boldsymbol{v}(t) \to \mathbf{0}$ by Lemma D.2. The variational equation is $\dot{\boldsymbol{v}} = DX(\boldsymbol{\theta}(t))\boldsymbol{v}$ with

$$DX(\boldsymbol{\theta}) \;=\; -2J(\boldsymbol{\theta})^\top J(\boldsymbol{\theta}) - 2Q(\boldsymbol{\theta}), \qquad Q(\boldsymbol{\theta}) := \sum_i r_i\, \nabla^2 g_i(\boldsymbol{\theta}), \qquad \|Q\|_{\text{op}} \leq M_2\, \|r\|$$

by the Cauchy–Schwarz inequality. Write $P(t) := P_{\boldsymbol{\theta}(t)}$ and $\boldsymbol{v}_v := (I - P)\boldsymbol{v}$. Since $(I - P)J(\boldsymbol{\theta})^\top = 0$,

$$\dot{\boldsymbol{v}}_v \;=\; -\dot P \boldsymbol{v} + (I - P)\dot{\boldsymbol{v}} \;=\; -\dot P \boldsymbol{v} - 2(I - P)Q\boldsymbol{v},$$

and $\boldsymbol{v}_v(t) \to \mathbf{0}$, so $\boldsymbol{v}_v(t) = \int_t^\infty \big[\dot P(s)\boldsymbol{v}(s) + 2(I - P(s))Q(s)\boldsymbol{v}(s)\big]\, ds$. For $s \geq t \geq t_1$, $\boldsymbol{\theta}(t) \in W^s_{\text{loc}}(\boldsymbol{\theta}^\star) \;=\; \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap U'$ and $\boldsymbol{v}(t) \in T_{\boldsymbol{\theta}(t)} W^s_{\text{loc}}(\boldsymbol{\theta}^\star)$, so Lemma D.1(iii) gives $\|\boldsymbol{v}(s)\| =$

$\|D\phi_{s-t}(\boldsymbol{\theta}(t))\boldsymbol{v}(t)\| \le K_0 e^{-\mu'(s-t)}\|\boldsymbol{v}(t)\|$; moreover $\left\|\dot{P}(s)\right\| \le M_P\left\|\dot{\boldsymbol{\theta}}(s)\right\| \le 2CM_P\|r(s)\|$ and $\|r(s)\| \le \|r(t)\|$ (Proposition D.1). Hence

$$\begin{aligned}\|\boldsymbol{v}_v(t)\| &\le \int_t^\infty \big(2CM_P + 2M_2\big)\|r(t)\|\, K_0 e^{-\mu'(s-t)}\|\boldsymbol{v}(t)\|\, ds \\ &= \frac{2K_0(M_2+CM_P)}{\mu'}\,\|r(t)\|\,\|\boldsymbol{v}(t)\|,\end{aligned}$$

which is the first bound; the second is $\|r(t)\| \le \|r_0\| e^{-2c^2 t}$. For the metric statement, the columns of $A(\boldsymbol{c}) = Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ lie in $T_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, so $N(\boldsymbol{c}(t)) = (I - P_{\boldsymbol{\theta}(t)})A(\boldsymbol{c}(t))$ has $\|N\| \le \omega(t)\|A(\boldsymbol{c}(t))\|$, where $\|A\|$ is bounded along the tail because A is continuous on $C_0(\boldsymbol{\theta}_0)$ and $\boldsymbol{c}(t) \to Y$; Theorem D.5 gives $G - K^{-1} = N^\top N = O(\omega(t)^2) = O(\|r(t)\|^2)$. □

The tilt is proportional to the Hessian scale M_2, zero in the kernel regime, and to the residual; it vanishes along the trajectory because the trajectory converges to a point where $\ker d\pi = \mathcal{H}$, not because $\mathcal{H}$ integrates. At $\boldsymbol{\theta}^\star$ its source is the curvature of the basin manifold:

Proposition D.7 (Curvature of the basin manifold at the interpolator)**.** *At $p := \boldsymbol{\theta}^\star$ write $E := \operatorname{Im} J(p)^\top$, $J := J(p)$, $K := J(p)\,J(p)^\top$, and for $\boldsymbol{v} = J^\top \boldsymbol{a} \in E$, $\boldsymbol{w} \in E$,*

$$\mathcal{B}(\boldsymbol{v},\boldsymbol{w}) \;:=\; (I - P_p)\sum_k a_k \nabla^2 g_k(p)\,\boldsymbol{w} \;\in\; \ker J,$$

so that $\mathcal{B}(\nabla g_i, \nabla g_j) = (I - P_p)\nabla^2 g_i \nabla g_j$, the transpose of the tensor B of Remark A.1. Near p the basin manifold is the graph $\{p + \boldsymbol{v} + q(\boldsymbol{v}) : \boldsymbol{v} \in E\}$ of a smooth $q : E \to \ker J$ with $q(\mathbf{0}) = \mathbf{0}$, $Dq(\mathbf{0}) = 0$, and its second fundamental form at p, $\mathrm{II}_p := D^2 q(\mathbf{0}) : E \times E \to \ker J$, is the unique symmetric bilinear solution of

$$\mathrm{II}_p(\boldsymbol{v}, J^\top J\boldsymbol{v}) \;=\; \sum_k (J\boldsymbol{v})_k\,(I - P_p)\nabla^2 g_k(p)\,\boldsymbol{v} \qquad \textit{for all } \boldsymbol{v} \in E.$$

In the K-eigenframe $\boldsymbol{v}_i := J^\top \boldsymbol{a}^{(i)} / \left\|J^\top \boldsymbol{a}^{(i)}\right\|$, $K\boldsymbol{a}^{(i)} = \lambda_i \boldsymbol{a}^{(i)}$, with S_{ij} and A_{ij} the symmetric and antisymmetric parts of $\mathcal{B}(\boldsymbol{v}_i, \boldsymbol{v}_j)$,

$$\mathrm{II}_p(\boldsymbol{v}_i, \boldsymbol{v}_j) \;=\; S_{ij} + \frac{\lambda_i - \lambda_j}{\lambda_i + \lambda_j}\,A_{ij}.$$

Hence $\mathrm{II}_p(\boldsymbol{v}_i, \boldsymbol{v}_i) = S_{ii}$ on eigen-directions, $\mathrm{II}_p = S$ when $\mathcal{B}$ is symmetric (the integrable class), and $\mathrm{II}_p = 0$ in the kernel regime.

Proof. By Theorem D.3(ii) and Lemma C.5, $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is near p the graph of q as stated, with $q(\boldsymbol{v}) = \frac{1}{2}\mathrm{II}_p(\boldsymbol{v},\boldsymbol{v}) + O(\|\boldsymbol{v}\|^3)$. Invariance makes X tangent to the graph: at $\delta := \boldsymbol{v} + q(\boldsymbol{v})$,

$$(I - P_p)X(p+\delta) \;=\; Dq(\boldsymbol{v})\big[P_p X(p+\delta)\big].$$

Expand $X(p+\delta) = -2J(p+\delta)^\top\,(g(p+\delta) - Y)$ with $g(p+\delta) - Y = J\delta + \frac{1}{2}[\delta^\top \nabla^2 g_k \delta]_k + O(\delta^3)$ and $J(p+\delta)^\top \boldsymbol{r} = \sum_k r_k(\nabla g_k + \nabla^2 g_k \delta) + O(\|\delta\|^2\,\|\boldsymbol{r}\|)$: every term of the form $J^\top(\cdot)$ is killed by $I - P_p$, so

$$(I - P_p)X(p+\delta) \;=\; -2\sum_k (J\delta)_k (I - P_p)\nabla^2 g_k\,\delta + O(\delta^3), \qquad P_p X(p+\delta) \;=\; -2J^\top J\delta + O(\delta^2).$$

Now $J\delta = J\boldsymbol{v}$ and $J^\top J q(\boldsymbol{v}) = \mathbf{0}$, since $q(\boldsymbol{v}) \in \ker J$, while $Dq(\boldsymbol{v}) = \mathrm{II}_p(\boldsymbol{v}, \cdot) + O(\|\boldsymbol{v}\|^2)$; the invariance identity therefore reads

$$-2\sum_k (J\boldsymbol{v})_k (I - P_p)\nabla^2 g_k\,\boldsymbol{v} + O(\boldsymbol{v}^3) \;=\; -2\,\mathrm{II}_p(\boldsymbol{v}, J^\top J\boldsymbol{v}) + O(\boldsymbol{v}^3),$$

and comparing second-order terms gives the displayed equation. The $\boldsymbol{v}_i$ are orthonormal with $J^\top J\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$; for $\boldsymbol{v} = \sum_i x_i \boldsymbol{v}_i$, equating the symmetric parts of the quadratic forms $\sum_{ij} x_i x_j \lambda_j \mathrm{II}_{ij}$ and $\mathcal{B}(J^\top J\boldsymbol{v}, \boldsymbol{v}) = \sum_{ij} x_i x_j \lambda_i \mathcal{B}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ gives $(\lambda_i + \lambda_j)\mathrm{II}_{ij} = \lambda_i(S_{ij} + A_{ij}) + \lambda_j(S_{ij} - A_{ij})$, the formula, and uniqueness since $\lambda_i + \lambda_j \ge 2c^2$. The consequences are read off. □

The formula was checked numerically on the model of Remark D.3 with $\boldsymbol{\theta}_0 = (0, 0, 1)$ and $Y = (0.95, 0.1)$, for which $\boldsymbol{\theta}^\star = (0.0517, 0.0508, 0.95)$, $\lambda_0 = 0.997$, $\lambda_1 = 1.91$, $\boldsymbol{v}_0 = (0.0560, 0.0532, -0.997)$, $\boldsymbol{v}_1 = (0.723, 0.687, 0.0772)$ and the unit vertical is $\hat{\xi} = (-0.689, 0.725, 0)$: the solution agrees with $-D^2\pi(\boldsymbol{\theta}^\star)(\boldsymbol{v}, \boldsymbol{v}) \cdot \hat{\xi}$, computed by finite differences on the basin manifold, to six digits in the four directions $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $(\boldsymbol{v}_0 \pm \boldsymbol{v}_1)/\sqrt{2}$ (values -0.0405, 0.0405, ∓ 0.342), $D^2\pi(\boldsymbol{v}, \boldsymbol{v})$ is vertical to 10^{-8}, and $S_{01} = -0.260$, $\frac{\lambda_0 - \lambda_1}{\lambda_0 + \lambda_1} A_{01} = -0.0824$, $\mathrm{II}_{01} = -0.342$: the twist contributes a quarter of the off-diagonal curvature, which the symmetric formula $\mathrm{II}_p = S$ of the integrable class would miss.

Theorem D.8 (Label dependence is forced)**.** *Let $M \subseteq B$ be a C^1 n-dimensional submanifold and $U \subseteq \mathbb{R}^n$ open. If M is invariant under the gradient flow of $\|g(\boldsymbol{\theta}) - Y'\|^2$ for every $Y' \in U$, then $T_{\boldsymbol{\theta}} M = \operatorname{Im} J(\boldsymbol{\theta})^\top$ at every $\boldsymbol{\theta} \in M$: M is an integral surface of $\mathcal{H}$, and $\mathcal{H}$ is involutive along M. Consequently, if $\mathcal{H}$ is non-involutive at every point of B, no n-dimensional surface in B is invariant under training for an open set of label vectors, and the basin manifold of every training problem is specific to that problem.*

Proof. Invariance puts $X_{Y'}(\boldsymbol{\theta}) = 2J(\boldsymbol{\theta})^\top (Y' - g(\boldsymbol{\theta})) \in T_{\boldsymbol{\theta}} M$ for all $Y' \in U$, hence also the differences $2J(\boldsymbol{\theta})^\top (Y' - Y'')$; $U - U$ spans $\mathbb{R}^n$, so $\operatorname{Im} J(\boldsymbol{\theta})^\top \subseteq T_{\boldsymbol{\theta}} M$, with equality by dimension. Involutivity along an integral surface is the argument of Proposition D.6(iii) (Lee, 2012, Ch. 8), and the last statement follows. □

General networks lie outside the integrable class (Remark A.3(iii)–(iv)), so the label dependence of the flow manifold is the typical case; what is label-free is not the carrier but the training-cost function of Appendix F.

D.3 The integrable special case

When $\mathcal{H}$ is integrable on B (Remark A.1) the basin manifold is label-free and horizontal; we record the identification, which is the only place in this appendix where integrability is a hypothesis.

Theorem D.9 (The basin manifold under integrability)**.** *Under Assumption 2.1 and integrability of $\mathcal{H}$ on B (Remark A.1), let $\mathcal{L}(\boldsymbol{\theta}_0)$ denote the leaf of $\mathcal{H}$ through $\boldsymbol{\theta}_0$. Then:*

- *(i) $\mathcal{L}(\boldsymbol{\theta}_0)$ is the set of parameters reachable from $\boldsymbol{\theta}_0$ by horizontal paths in B; it depends only on g and $\boldsymbol{\theta}_0$, not on Y, is invariant under gradient flow for every loss and every Y, and contains the trajectory $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ and its limit $\boldsymbol{\theta}^\star$.*
- *(ii) The restriction $g|_{\mathcal{L}(\boldsymbol{\theta}_0)}$ is a local diffeomorphism onto an open subset of $\mathbb{R}^n$. In particular there is an open neighbourhood $V \subseteq B(\boldsymbol{\theta}_0, R)$ of $\boldsymbol{\theta}^\star$ such that the component of $\mathcal{L}(\boldsymbol{\theta}_0) \cap V$ containing $\boldsymbol{\theta}^\star$ is embedded and mapped by g diffeomorphically onto an open ball $C_0(\boldsymbol{\theta}_0) \ni Y$.*
- *(iii) For every $\mathbf{c} \in C_0(\boldsymbol{\theta}_0)$ the output fibre $\Theta(\mathbf{c})$ meets this component orthogonally in exactly one point, $g|_{\boldsymbol{\theta}_0}^{-1}(\mathbf{c})$, the preimage of $\mathbf{c}$ under the restriction of g to the component; in particular $\boldsymbol{\theta}^\star = g|_{\boldsymbol{\theta}_0}^{-1}(Y)$.*
- *(iv) Gradient flow from any point of this component stays in it and converges to $\boldsymbol{\theta}^\star$, and $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ lies in it for all large t.*
- *(v) $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is the connected component, in the leaf topology, of $\mathcal{L}(\boldsymbol{\theta}_0) \cap \mathfrak{B}$ containing $\boldsymbol{\theta}_0$; in particular $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V = P_0$, the plaque of the leaf through $\boldsymbol{\theta}^\star$, $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \mathcal{H}$ at every point, $N \equiv 0$, and the basin manifolds of all label vectors in the label ball $\mathcal{N}$ of Appendix F and all losses are components of the same leaf.*

Proof of Theorem D.9. Frobenius charts. $\mathcal{H}$ is spanned on B by $\nabla g_1, \ldots, \nabla g_n$, has constant rank n and is involutive, so (Lee, 2012, Ch. 19) every point of B has a flat chart $\varphi = (x, y) : W \to Q_1 \times Q_2 \subseteq \mathbb{R}^n \times \mathbb{R}^{m-n}$ (Q_i open cubes) with $\mathcal{H}|_W = \operatorname{span}\{\partial_{x_i}\}$, whose slices $\{y = \text{const}\}$ (plaques) are connected embedded integral manifolds; the leaves partition B, are injectively immersed with tangent space $\mathcal{H}$, contain every connected integral manifold they meet, and are weakly embedded. Plaques are open in the leaf topology, which agrees with the subspace topology on them; every statement below is made on a plaque.

(i) An absolutely continuous curve with $\dot{\gamma} \in \mathcal{H}$ a.e. (in particular every solution of $\dot{\boldsymbol{\theta}} = J(\boldsymbol{\theta})^\top \boldsymbol{u}$, $\boldsymbol{u} \in L^2$) has $\dot{y} = 0$ a.e. in each of finitely many flat charts covering it, so it lies in one leaf. Conversely, compositions of the flows of $\pm\nabla g_i$ (the controls $\pm e_i$) map a neighbourhood of $0 \in \mathbb{R}^n$ onto a leaf-neighbourhood of any point, by the inverse function theorem, so the set reachable from $\boldsymbol{\theta}_0$ is open and closed in the connected leaf, hence equal to it. The leaf depends on g and $\boldsymbol{\theta}_0$ only; the negative gradient of any loss is a section of $\mathcal{H}$ (Lemma C.1), so every training trajectory stays in its leaf while in B. In a flat chart about $\boldsymbol{\theta}^\star$ with $\varphi(\boldsymbol{\theta}^\star) = (0,0)$, the tail of $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ has constant y, equal to $y(\boldsymbol{\theta}^\star) = 0$ by continuity, so it lies in the plaque $P^\star$ through $\boldsymbol{\theta}^\star$, and $\boldsymbol{\theta}^\star \in P^\star \subseteq \mathcal{L}(\boldsymbol{\theta}_0)$.

(ii) On the leaf, $dg = J(\boldsymbol{\theta})|_{\operatorname{Im} J(\boldsymbol{\theta})^\top}$ is an isomorphism with inverse of norm at most $1/c$, since $K(\boldsymbol{\theta}) \succeq c^2 I$; by the inverse function theorem there is $P_0 = \varphi^{-1}(A \times \{0\}) \subseteq P^\star$, A open and connected, mapped diffeomorphically onto a ball $C_0(\boldsymbol{\theta}_0) := B(Y, \rho)$. With $V := \varphi^{-1}(A \times Q_2)$, the leaf meets W in countably many plaques (second countability), so y is constant on connected subsets of $\mathcal{L}(\boldsymbol{\theta}_0) \cap V$ and P_0 is the component containing $\boldsymbol{\theta}^\star$, embedded as a slice.

(iii) $g|_{P_0}$ is a bijection onto $C_0(\boldsymbol{\theta}_0)$, and at $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$, $TP_0 = \operatorname{Im} J^\top \perp \ker J = T\Theta(\boldsymbol{c})$ (Lemmas C.4, C.5).

(iv) While the flow from $\boldsymbol{\theta}_1 \in P_0$ stays in P_0, $\boldsymbol{c} := g(\boldsymbol{\theta})$ obeys $\dot{\boldsymbol{c}} = -2K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)(\boldsymbol{c} - Y)$, so $\|\boldsymbol{c}(t) - Y\| \le \|\boldsymbol{c}(0) - Y\| e^{-2c^2 t}$ and $\boldsymbol{\theta}(t)$ stays in the compact $g|_{\boldsymbol{\theta}_0}^{-1}\left(\bar{B}(Y, \|\boldsymbol{c}(0) - Y\|)\right) \subset P_0$; a first-exit argument ($y \equiv 0$ near the exit, A open) shows it never leaves, and $\boldsymbol{\theta}(t) \to g|_{\boldsymbol{\theta}_0}^{-1}(Y) = \boldsymbol{\theta}^\star$. The tail of $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ lies in $P^\star$ with $x \to 0 \in A$, hence in P_0. As $\nabla_{\boldsymbol{\theta}}\mathcal{L} \in T\mathcal{L}(\boldsymbol{\theta}_0)$, gradient flow on the leaf is the Riemannian gradient flow of the restricted loss.

(v) By (iv) and Theorem D.3(iii), $P_0 \subseteq \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, open (dimension) and closed in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$, so P_0 is its component through $\boldsymbol{\theta}^\star$. A point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ on a leaf $\mathcal{L}'$ has a horizontal trajectory whose tail lies in $P^\star$, as in (i), so $\mathcal{L}' = \mathcal{L}(\boldsymbol{\theta}_0)$. The inclusion of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ in the leaf is an injective immersion of n-manifolds (weak embeddedness), hence open, and closed since $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is closed in $\mathfrak{B}$ and the leaf topology is finer; so $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is the component of $\mathcal{L}(\boldsymbol{\theta}_0) \cap \mathfrak{B}$ through $\boldsymbol{\theta}_0$, $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \mathcal{H}$ and $N \equiv 0$ (Proposition D.6(i)). The same argument applies to any $Y' \in \mathcal{N}$ and any loss of Remark D.1, the leaf being the same by (i). □

Lemma D.4 (Chart differential and inverse-NTK metric)**.** *Under Assumption 2.1 and integrability, let V, $C_0(\boldsymbol{\theta}_0)$ and $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ be as in Theorem D.9. For every $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0)$,*

$$Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) \;=\; J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^\top K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1} \;=\; J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{+},$$
$$G(\boldsymbol{c}) := Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})^\top Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) \;=\; K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1},$$

where J^+ denotes the Moore–Penrose pseudoinverse: the chart differential is the unique right inverse of $J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)$ with image in $\operatorname{Im} J\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^\top$, and the pullback of the induced metric of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ to $C_0(\boldsymbol{\theta}_0)$ is the inverse NTK.

Proof. Write $\boldsymbol{\theta} := g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$. Differentiating $g \circ g|_{\boldsymbol{\theta}_0}^{-1}(\cdot) = \mathrm{id}$ on $C_0(\boldsymbol{\theta}_0)$ gives $J(\boldsymbol{\theta})\, Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) = I_n$. The columns of $Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ lie in $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \operatorname{Im} J(\boldsymbol{\theta})^\top$ (Theorem D.9(v)), so $Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) = J(\boldsymbol{\theta})^\top Z$ for some $Z \in \mathbb{R}^{n \times n}$, and $I_n = J(\boldsymbol{\theta}) J(\boldsymbol{\theta})^\top Z = K(\boldsymbol{\theta}) Z$ gives $Z = K(\boldsymbol{\theta})^{-1}$; for full row rank this is the pseudoinverse $J^+ = J^\top (JJ^\top)^{-1}$. Then $G(\boldsymbol{c}) = K(\boldsymbol{\theta})^{-1} J(\boldsymbol{\theta}) J(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})^{-1} = K(\boldsymbol{\theta})^{-1}$. □

Corollary D.5 (Under integrability the flow manifold carries the inverse-NTK metric)**.** *Under Assumption 2.1 and integrability, for $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, $\dot{\boldsymbol{\theta}} \in T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ and $\dot{\boldsymbol{c}} := J(\boldsymbol{\theta})\dot{\boldsymbol{\theta}}$,*

$$\left\|\dot{\boldsymbol{\theta}}\right\|^2 \;=\; \dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta})^{-1} \dot{\boldsymbol{c}}\,:$$

the length of a path on the flow manifold is the length of its image in training-output space measured by the inverse NTK along it. Consequently the energy of the training problem, $E_Y(\boldsymbol{c}) = d^2_{\mathrm{flow}}\left(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)$ (Proposition F.1), is bounded below by the label-free leaf energy $E_{\mathcal{L}}(\boldsymbol{c}) := d^2_{\mathcal{L}}(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}))$, the squared intrinsic distance of $\mathcal{L}(\boldsymbol{\theta}_0) \cap B$, with equality wherever the

infimum over leaf paths is approached by paths in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$; on $C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}$, $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) = \tau(\boldsymbol{c})$ for the training map τ of Appendix F, and the training-cost function $E(\boldsymbol{c}) = E_{\boldsymbol{c}}(\boldsymbol{c})$ of that appendix obeys the same bound with the same equality condition, so that $E_Y = E = E_{\mathcal{L}}$ wherever both conditions hold: in the integrable class the energies of all training problems in $\mathcal{N}$ are, where these conditions hold, restrictions of one function of the outputs, determined by g and $\boldsymbol{\theta}_0$. Read through the chart $g|_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)\cap V}$, training-output space carries the Riemannian metric $\boldsymbol{c} \mapsto K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)^{-1}$, whose squared geodesic distance from $g(\boldsymbol{\theta}_0)$ to $\boldsymbol{c}$ equals $E_Y(\boldsymbol{c})$ whenever $\boldsymbol{\theta}_0$ and a minimising geodesic lie in the chart domain, and bounds it above whenever $\boldsymbol{\theta}_0$ does; and function-space gradient flow $\dot{\boldsymbol{c}} = -2K(\boldsymbol{\theta})(\boldsymbol{c} - Y)$ is the Riemannian gradient flow of $\mathcal{L}(\boldsymbol{c}, Y)$ in this metric.

Proof of Corollary D.5. Pointwise identity. $\dot{\boldsymbol{\theta}} = J(\boldsymbol{\theta})^\top \boldsymbol{u}$ with $\boldsymbol{u} = K(\boldsymbol{\theta})^{-1}\dot{\boldsymbol{c}}$ (Theorem D.9(v)), so $\left\|\dot{\boldsymbol{\theta}}\right\|^2 = \boldsymbol{u}^\top K(\boldsymbol{\theta})\boldsymbol{u} = \dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta})^{-1}\dot{\boldsymbol{c}}$.

Leaf energy. $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is leaf-open in $\mathcal{L}(\boldsymbol{\theta}_0) \cap B$, so $d_{\mathcal{L}} \le d_{\text{flow}}(\cdot,\cdot)$ on it, with the stated equality condition; $d_{\mathcal{L}}$ depends on $g, \boldsymbol{\theta}_0$ only. For $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}$, $\tau(\boldsymbol{c})$ lies in the leaf by (i) and is continuous into the leaf topology (Theorem F.2, weak embeddedness (Lee, 2012, Ch. 19)); the set of $\boldsymbol{c}$ with $\tau(\boldsymbol{c}) \in P_0$ is open, closed in the connected set $C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}$, and contains Y, so $\tau(\boldsymbol{c}) = g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ there. The basin manifold of the problem with labels $\boldsymbol{c}$ is a component of the same leaf by (v), which gives $E_{\mathcal{L}}(\boldsymbol{c}) \le E(\boldsymbol{c})$ with the same equality condition.

Chart and gradient flow. By Lemma D.4 the chart pulls the induced metric back to $K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)\right)^{-1}$, so the chart energy is the infimum defining E_Y restricted to paths in the plaque: an upper bound when $\boldsymbol{\theta}_0$ is in the chart domain, with equality when a minimising geodesic lies in it. The Riemannian gradient of $\|\boldsymbol{c} - Y\|^2$ for this metric is $2K\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)(\boldsymbol{c} - Y)$ (Theorem D.9(iv)). □

In the kernel regime the plaque may be taken to be the whole affine flow manifold within the trapping region (Corollary D.3), so the chart expression, E_Y and E agree there.

Remark D.3 (The simplest non-involutive model). Let $g(\boldsymbol{\theta}) = (\theta_3,\ \theta_1 + \theta_2\theta_3)$ on $\mathbb{R}^3$, realised by $f(x;\boldsymbol{\theta}) = \theta_3 + x(\theta_1 - \theta_3 + \theta_2\theta_3)$ at the inputs $x \in \{0,1\}$. Then $\nabla g_1 = (0,0,1)^\top$, $\nabla g_2 = (1,\theta_3,\theta_2)^\top$ and $\det K(\boldsymbol{\theta}) = 1 + \theta_3^2 \ge 1$, so the spectral bounds of Assumption 2.1 hold on every bounded ball (for the instance $\boldsymbol{\theta}_0 = (0,0,1)$, $Y = (0.95, 0.1)$ used below the whole assumption holds on $B(\boldsymbol{\theta}_0, \frac{1}{2})$ with $c = 2/\sqrt{5}$ and $C = \sqrt{13}/2$: there $\operatorname{tr} K(\boldsymbol{\theta}) < \frac{5}{4} + 2\theta_3$, so the characteristic polynomial of $K(\boldsymbol{\theta})$ exceeds $(\theta_3 - \frac{4}{5})^2 \ge 0$ at $\frac{4}{5}$ and $(\theta_3 - \frac{3}{2})(\theta_3 - 5) > 0$ at $\frac{13}{4}$, and $C\|g(\boldsymbol{\theta}_0) - Y\|/c^2 = \sqrt{65}/32 = 0.252 < \frac{1}{2}$), and $\ker J(\boldsymbol{\theta})$ is spanned by $\xi := (-\theta_3, 1, 0)^\top$. The bracket $[\nabla g_1, \nabla g_2] = (0,1,0)^\top$ has $\xi^\top[\nabla g_1, \nabla g_2] = 1$: it is transverse to $\mathcal{H}$ at every point, $\{\nabla g_1, \nabla g_2, [\nabla g_1, \nabla g_2]\}$ spans $\mathbb{R}^3$, and $\mathcal{H}$ is bracket-generating of step two – it is the kernel of the standard contact form $d\theta_2 - \theta_3\, d\theta_1$. Consequently no two-dimensional integral surface of $\mathcal{H}$ exists, so the basin manifold $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is tilted out of $\mathcal{H}$ on a dense open subset of itself and $N \ne 0$ there (Proposition D.6), and its curvature at $\boldsymbol{\theta}^\star$ is computed in Proposition D.7. The tanh neuron of Remark A.3(iii) with $x_1x_2 \ne -1$ is the network-flavoured instance of the same phenomenon.

Remark D.4 (Horizontal reachability and the sub-Riemannian distance). The set $\mathcal{O}(\boldsymbol{\theta}_0)$ of parameters reachable from $\boldsymbol{\theta}_0$ by horizontal paths in B, the orbit of the frame $\{\nabla g_i\}$ (Sussmann, 1973), contains $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$: the forward orbit of any $\boldsymbol{\theta} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, reparametrised by arc length, is a horizontal curve of finite length (Proposition D.1) with control $\boldsymbol{u} = K(\gamma)^{-1}J(\gamma)\gamma'$, $\|\boldsymbol{u}\| \le 1/c$ since $1 = \|\gamma'\|^2 = \boldsymbol{u}^\top K(\gamma)\boldsymbol{u} \ge c^2\|\boldsymbol{u}\|^2$, ending at $\boldsymbol{\theta}^\star$, and controls may be reversed. In the integrable class $\mathcal{O}(\boldsymbol{\theta}_0)$ is the leaf (Theorem D.9(i)); when $\mathcal{H}$ is bracket-generating, as in Remark D.3, it is the whole ball (Agrachev et al., 2019, Ch. 3), g has positive-dimensional fibres on it, and the least horizontal energy from $\boldsymbol{\theta}_0$, the squared sub-Riemannian distance $d_{\text{SR}}^2(\boldsymbol{\theta}_0, \cdot)$, is label-free but *not* unbiased: on Remark D.3 with $\boldsymbol{\theta}_0 = (0,0,1)$ and $Y = (0.95, 0.1)$, the interpolating line $\Theta^\star = \{(0.1 - 0.95t,\ t,\ 0.95)\}$ has gradient-flow limit $t^\star = 0.05083$, sub-Riemannian-nearest point $t_{\text{SR}} = 0.05061$ and chord-nearest point $t_{\text{AR}} = 0.04993$, and $\frac{d}{dt}d_{\text{SR}}^2(\boldsymbol{\theta}_0, \cdot)|_{\Theta^\star, t^\star} = 3.66 \ne 0$. The biases $t_{\text{SR}} - t^\star = -2.17 \cdot 10^{-4}$ and $t_{\text{AR}} - t^\star = -8.97 \cdot 10^{-4}$ are 0.247% and 1.02% of the displacement $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\| = 0.0881$ (0.34% and 1.40% measured as distances in $\mathbb{R}^3$), the sub-Riemannian one 4.13 times smaller, and halving the residual ($Y = (0.975, 0.05)$) divides them by 4.09 and 4.15: both biases are second order in the residual, as Remark D.9 proves for the chord's,

while for the sub-Riemannian one this is a numerical observation without a general bound. Thus a regulariser that is a function of the parameters and unbiased for the training problem must be anchored on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.3(iv)); label-independence is recovered at the level of the training-cost function (Appendix F), not of the carrier.

D.4 Geometry of the flow manifold and output fibres

Proposition D.10 (Output fibres are smooth submanifolds)**.** *Under Assumption 2.1, for every $\boldsymbol{c} \in g(B(\boldsymbol{\theta}_0, R))$ the output fibre $\Theta(\boldsymbol{c}) \cap B(\boldsymbol{\theta}_0, R)$ is an embedded C^∞ submanifold of $\mathbb{R}^m$ of codimension n.*

Proof of Proposition D.10. Immediate from Lemma C.4. □

Remark D.5 (Path-connectivity of output fibres). In the overparametrised regime $m \gg n$, the connected component of $\Theta(\boldsymbol{c}) \cap V$ containing $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ is path-connected via the implicit-function-theorem chart on $\Theta(\boldsymbol{c})$ over V (Lee, 2012), so the fibre distance $d_{\text{fibre}}\left(\boldsymbol{\theta}, g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)$ in Definition 4.1 is finite.

Remark D.6 (Existence and non-uniqueness of minimising geodesics). The flow manifold with its induced metric is in general not complete – it is a level set of π on the open set $\mathfrak{B}$ – so the infimum defining $d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, always finite, need not be attained. It is attained whenever the closed metric ball of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ about $\boldsymbol{\theta}_0$ of radius $d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is compact, by the Hopf–Rinow argument (do Carmo, 1992, Ch. 7); for $n = 1$ the flow manifold is a curve and the minimising geodesic is the arc between the two points. Minimisers need not be unique: conjugate points and non-trivial topology of the basin manifold can create several shortest paths, and nothing excludes $\boldsymbol{\theta}^\star$ from the cut locus of $\boldsymbol{\theta}_0$. There $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \cdot)$ is Lipschitz but not differentiable, and its Clarke subdifferential is the convex hull of $2\, d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ times the terminal unit tangents of the minimising geodesics, whose negatives are descent directions for $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \cdot)$. The arc-ridge force of Definition 4.2 never references a geodesic and is unaffected by non-uniqueness: it is determined by the realised training path alone.

D.5 Anchored ridge and the gradient-flow limit

The formal version of Theorem 3.1:

Theorem D.11 (Anchored ridge and the gradient-flow limit)**.** *Let Assumption 2.1 hold, let $\boldsymbol{\theta}^\star$ be the gradient-flow limit of Proposition 2.1, and for $\lambda > 0$ let $\boldsymbol{\theta}_\lambda$ be any minimiser of $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ over Θ. Let Λ be the set of limit points of $\{\boldsymbol{\theta}_\lambda\}$ as $\lambda \downarrow 0$.*

- *(a)* ***Localisation.*** *$\Lambda \neq \emptyset$ and $\Lambda \subseteq \Theta^\star \cap \overline{B}(\boldsymbol{\theta}_0, \ell)$ with $\ell = C\, \|g(\boldsymbol{\theta}_0) - Y\| / c^2 < R$.*
- *(b)* ***Characterisation.*** *Every $\bar{\boldsymbol{\theta}} \in \Lambda$ minimises $\|\cdot - \boldsymbol{\theta}_0\|^2$ over $\Theta^\star$ and satisfies $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \in \operatorname{Im} J(\bar{\boldsymbol{\theta}})^\top$.*
- *(c)* ***Bias.*** *If $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \notin \operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$ then $\boldsymbol{\theta}^\star \notin \Lambda$ and $\liminf_{\lambda \downarrow 0} \|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^\star\| > 0$: no sequence of anchored-ridge minimisers converges to the gradient-flow limit, however small the regularisation.*
- *(d)* ***The alignment condition is not sufficient.*** *There are models satisfying Assumption 2.1 with $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \in \operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$ and $\boldsymbol{\theta}^\star \notin \Lambda$ (Remark D.7).*

The same holds for standard ridge $\|\boldsymbol{\theta}\|^2$ with $\boldsymbol{\theta}_0$ replaced by $\mathbf{0}$ in (b)–(d), except that (a) weakens to $\Lambda \subseteq \Theta^\star \cap \overline{B}(\mathbf{0}, \|\boldsymbol{\theta}^\star\|)$ (with $\Lambda \neq \emptyset$ whenever $\overline{B}(\mathbf{0}, \|\boldsymbol{\theta}^\star\|) \subset \Theta$, as when $\Theta = \mathbb{R}^m$) and (b) holds at limit points inside $B(\boldsymbol{\theta}_0, R)$.

Remark D.7 (The alignment condition is not sufficient; the regime does not decide the bias). *(A) Constant Jacobian along the trajectory, biased anchored ridge.* Let $m = 2, n = 1, g(\boldsymbol{\theta}) = \theta_2 - h(\theta_1)$ with $h(x) = \frac{1}{2}\left(\sqrt{1 + 4x^2} - 1\right)$, $Y = 0$, $\boldsymbol{\theta}_0 = (0, 1)$. Then $J(\boldsymbol{\theta}) = (-h'(\theta_1), 1)$ with $h'(x) = 2x/\sqrt{1 + 4x^2} \in (-1, 1)$, so $1 \leq \sigma(J(\boldsymbol{\theta})) < \sqrt{2}$ everywhere and Assumption 2.1 holds with $c = 1$, $C = \sqrt{1 + h'(R)^2}$, $R = 3/2$ (then $C = 1.378 < R$ and $\|g(\boldsymbol{\theta}_0) - Y\| = 1$). The axis $\theta_1 = 0$ is invariant ($h'(0) = 0$), so $\boldsymbol{\theta}(t) = (0, e^{-2t})$, $\boldsymbol{\theta}^\star = (0, 0)$, and $J(\boldsymbol{\theta}(t)) \equiv (0, 1)$: the trajectory is in the kernel regime in the trajectory-wise sense, and $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 = (0, -1) \in \operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$. Yet on $\Theta^\star = \{\theta_2 = h(\theta_1)\}$, writing $s = \sqrt{1 + 4\theta_1^2}$, the chord is $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 = \frac{1}{2}(s^2 - 3s + 4)$, minimised at

$s = 3/2$, i.e. at $\boldsymbol{\theta} = (\pm\sqrt{5}/4,\ 1/4)$ with value $7/8 < 1 = \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\|^2$; $\boldsymbol{\theta}^\star$ is a strict local *maximum* of the chord along $\Theta^\star$. Hence $\emptyset \neq \Lambda \subseteq \{(\pm\sqrt{5}/4, 1/4)\}$ by Theorem D.11(a),(b), so $\boldsymbol{\theta}^\star \notin \Lambda$; the minimisers come in pairs exchanged by the symmetry $\theta_1 \mapsto -\theta_1$ of the model and of $\boldsymbol{\theta}_0$, so either point is the limit of a suitable selection. Geodesic ridge is unbiased here: $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is the axis (Corollary D.3), $d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = (1 - g(\boldsymbol{\theta}))^2 + \left(\int_0^{\theta_1} \sqrt{1 + h'(s)^2}\, ds\right)^2$, uniquely minimised on $\Theta^\star$ at $\boldsymbol{\theta}^\star$.

(B) Non-constant Jacobian along the trajectory, unbiased anchored ridge. In the model of Proposition A.1 take $\boldsymbol{\theta}_0 = (a, a)$ with $1 < a < 1.09$ (so that $|a^2 - 1| < a^2/6$ and Assumption 2.1 holds with $R = \|\boldsymbol{\theta}_0\|/2$). The diagonal is invariant, $J(\boldsymbol{\theta}(t)) = (s(t), s(t))$ with $s(t) \downarrow 1$ non-constant, and $\boldsymbol{\theta}^\star = (1, 1)$. The chord $(x - a)^2 + (1/x - a)^2$ on the branch $\theta_1 > 0$ has the single critical point $x = 1$ (the quartic $x^4 - ax^3 + ax - 1 = (x - 1)(x + 1)(x^2 - ax + 1)$ has no other positive root for $a < 2$), and on the branch $\theta_1 < 0$ the chord exceeds $2a^2 > 2(a - 1)^2$, so $\boldsymbol{\theta}^\star$ is the unique minimiser of the chord over $\Theta^\star$ and $\Lambda = \{\boldsymbol{\theta}^\star\}$ by Theorem D.11(a),(b).

Together with Remark D.10, (A) and (B) show that the alignment condition, not the regime, decides the bias.

Proof of Theorem D.11. Write $\Phi_\lambda(\boldsymbol{\theta}) := \mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$; by Proposition 2.1, $\boldsymbol{\theta}^\star \in \Theta^\star \cap \overline{B}(\boldsymbol{\theta}_0, \ell)$.

(a) Φ_λ attains its minimum on the compact $\overline{B}(\boldsymbol{\theta}_0, \ell) \subset \Theta$ at a value at most $\Phi_\lambda(\boldsymbol{\theta}^\star) \leq \lambda\ell^2$, while $\Phi_\lambda > \lambda\ell^2$ outside the ball; so minimisers exist, and all lie in $\overline{B}(\boldsymbol{\theta}_0, \ell)$ (Corollary D.1). Bolzano–Weierstrass gives $\Lambda \neq \emptyset$, and Lemma C.9 with $R = \|\cdot - \boldsymbol{\theta}_0\|^2$ gives $\Lambda \subseteq \Theta^\star$.

(b) Lemma C.9 gives the minimality. At $\bar{\boldsymbol{\theta}} \in B(\boldsymbol{\theta}_0, R)$, $\Theta^\star$ is a submanifold with tangent space $\ker J(\bar{\boldsymbol{\theta}})$ (Lemma C.4), so the first-order condition is $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \perp \ker J(\bar{\boldsymbol{\theta}})$, that is, $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \in \operatorname{Im} J(\bar{\boldsymbol{\theta}})^\top$ (Lemma C.5).

(c) By (b), $\boldsymbol{\theta}^\star \notin \Lambda$; Λ is closed, and any sequence realising the $\liminf$ has a subsequence converging into Λ, so $\liminf_{\lambda \downarrow 0} \|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^\star\| \geq \operatorname{dist}(\boldsymbol{\theta}^\star, \Lambda) > 0$.

(d) Remark D.7(A).

Standard ridge. Comparison with $\boldsymbol{\theta}^\star$ gives $\|\boldsymbol{\theta}_\lambda\| \leq \|\boldsymbol{\theta}^\star\|$, and (a)–(c) repeat, the first-order condition needing only $\bar{\boldsymbol{\theta}} \in B(\boldsymbol{\theta}_0, R)$; for (d), translate the witness by $-\boldsymbol{\theta}_0$, so that anchored ridge at $\mathbf{0}$ is standard ridge. □

Corollary D.6 (Kernel regime: anchored ridge is unbiased)**.** *In the kernel regime of Definition 2.1,* $\Lambda = \{\boldsymbol{\theta}^\star\}$*:* $\lim_{\lambda \downarrow 0} \boldsymbol{\theta}_\lambda = \boldsymbol{\theta}^\star$*. More generally* $\Lambda = \{\boldsymbol{\theta}^\star\}$ *whenever* $\boldsymbol{\theta}^\star$ *is the unique minimiser of* $\|\cdot - \boldsymbol{\theta}_0\|^2$ *over* $\Theta^\star$*.*

Proof of Corollary D.6. In the kernel regime $\Theta^\star \cap B(\boldsymbol{\theta}_0, R) = (\boldsymbol{\theta}^\star + \ker J) \cap B(\boldsymbol{\theta}_0, R)$ with $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0 - J^+(g(\boldsymbol{\theta}_0) - Y)$ (proof of Corollary D.3; Proposition D.15(b) at $\boldsymbol{c} = Y$). As $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \in \operatorname{Im} J^\top \perp \ker J$, $\|\boldsymbol{\theta}^\star + \boldsymbol{w} - \boldsymbol{\theta}_0\|^2 = \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\|^2 + \|\boldsymbol{w}\|^2$ for $\boldsymbol{w} \in \ker J$, and points of $\Theta^\star$ outside the ball are at distance at least $R > \ell$ from $\boldsymbol{\theta}_0$; so $\boldsymbol{\theta}^\star$ is the unique minimiser of the chord over $\Theta^\star$, and Theorem D.11(a),(b) with compactness give $\boldsymbol{\theta}_\lambda \to \boldsymbol{\theta}^\star$. The second sentence is the same last step. □

Remark D.8 (Fixed horizontal range along the trajectory). If $\operatorname{Im} J(\boldsymbol{\theta}(t))^\top$ is a fixed subspace S for all $t \geq 0$ – in particular if $J(\boldsymbol{\theta}(t)) \equiv J(\boldsymbol{\theta}_0)$ along the trajectory – then $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 = \int_0^\infty J(\boldsymbol{\theta}(s))^\top \boldsymbol{u}(s)\, ds \in S = \operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$, so the hypothesis of Theorem D.11(c) fails; by (d) this does not imply unbiasedness. Conversely, a rotating horizontal range is necessary for the bias mechanism of (c): the displacement acquires a component in $\ker J(\boldsymbol{\theta}^\star)$ only if $\operatorname{Im} J(\boldsymbol{\theta}(t))^\top$ varies along the trajectory. This is the precise sense in which anchored-ridge bias is a feature-learning phenomenon.

Lemma D.5 (Misalignment is controlled by the rotation of the horizontal space)**.** *Under Assumption 2.1, with $P_{\boldsymbol{\theta}} := J(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})^{-1} J(\boldsymbol{\theta})$ the orthogonal projector onto $\operatorname{Im} J(\boldsymbol{\theta})^\top$,*

$$\begin{aligned}\|(I - P_{\boldsymbol{\theta}^\star})(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0)\| \;&\le\; \int_0^\infty \left\|P_{\boldsymbol{\theta}(t)} - P_{\boldsymbol{\theta}^\star}\right\|_{\mathrm{op}} \left\|\dot{\boldsymbol{\theta}}(t)\right\| dt \\ &\le\; \sup_{t\ge 0} \left\|P_{\boldsymbol{\theta}(t)} - P_{\boldsymbol{\theta}^\star}\right\|_{\mathrm{op}} \frac{C\,\|g(\boldsymbol{\theta}_0) - Y\|}{c^2}.\end{aligned}$$

Proof of Lemma D.5. $\dot{\boldsymbol{\theta}}(t) \in \operatorname{Im} J(\boldsymbol{\theta}(t))^\top$, so $(I - P_{\boldsymbol{\theta}^\star})\dot{\boldsymbol{\theta}} = (P_{\boldsymbol{\theta}(t)} - P_{\boldsymbol{\theta}^\star})\dot{\boldsymbol{\theta}}$; integrate $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 = \int_0^\infty \dot{\boldsymbol{\theta}}$ and use Proposition 2.1. □

Remark D.9 (The anchored-ridge bias is second order in the displacement). Let M_P be a Lipschitz constant of $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ on $B(\boldsymbol{\theta}_0, R)$, $M_2 := \sup_B (\sum_i \left\|\nabla^2 g_i\right\|_{\mathrm{op}}^2)^{1/2}$ as in Lemma D.3, $s_\infty \le \ell := C\,\|g(\boldsymbol{\theta}_0) - Y\|/c^2$ the length of the trajectory and $s(t)$ its arc length. Since $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^\star\| \le s_\infty - s(t)$, the first bound of Lemma D.5 gives

$$\|(I - P_{\boldsymbol{\theta}^\star})(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0)\| \;\le\; M_P \int_0^{s_\infty} (s_\infty - s)\, ds \;=\; \tfrac{1}{2} M_P s_\infty^2 \;\le\; \tfrac{1}{2} M_P \ell^2,$$

and this is half the norm of the intrinsic gradient at $\boldsymbol{\theta}^\star$ of the chord $f := \|\cdot - \boldsymbol{\theta}_0\|^2$ on $\Theta^\star$. Along a unit-speed geodesic γ of $\Theta^\star \cap B$, differentiating $g_k(\gamma) = Y_k$ twice gives $\ddot{\gamma} = -J(\gamma)^+ [\dot{\gamma}^\top \nabla^2 g_k \dot{\gamma}]_k$, so $\|\ddot{\gamma}\| \le M_2/c$ and $\frac{d^2}{dt^2} f(\gamma) = 2 + 2\langle \gamma - \boldsymbol{\theta}_0, \ddot{\gamma}\rangle \ge 2(1 - \rho M_2/c)$ wherever $\|\gamma - \boldsymbol{\theta}_0\| \le \rho$. Hence, if a limit point $\bar{\boldsymbol{\theta}}$ of anchored ridge (Theorem D.11) is joined to $\boldsymbol{\theta}^\star$ by a geodesic of $\Theta^\star \cap B$ of length D along which $\|\cdot - \boldsymbol{\theta}_0\| \le \rho$ with $\rho M_2 < c$, then $f(\bar{\boldsymbol{\theta}}) \le f(\boldsymbol{\theta}^\star)$ and $f(\bar{\boldsymbol{\theta}}) \ge f(\boldsymbol{\theta}^\star) - M_P s_\infty^2 D + (1 - \rho M_2/c) D^2$ give

$$\left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\| \;\le\; D \;\le\; \frac{M_P\, \ell^2}{1 - \rho M_2/c}.$$

The bias is second order in the displacement ℓ, hence in the initial residual; both constants vanish in the kernel regime.

Remark D.10 (The alignment condition fails for every off-diagonal initialisation of the bilinear model). For $f(x; \boldsymbol{\theta}) = \theta_1 \theta_2 x$ on $\{(1,1)\}$, $\operatorname{Im} J(\boldsymbol{\theta})^\top = \operatorname{span}(\theta_2, \theta_1)$ is the tangent line of the hyperbola $\{\theta_1^2 - \theta_2^2 = \delta\}$, the flow manifold through $\boldsymbol{\theta}_0$ (Corollary D.3; δ is conserved by every horizontal motion). For $\delta \neq 0$ the tangent line at $\boldsymbol{\theta}^\star$ meets the hyperbola only at $\boldsymbol{\theta}^\star$: substituting $\boldsymbol{\theta}^\star + a(\theta_2^\star, \theta_1^\star)$ into $\theta_1^2 - \theta_2^2 - \delta$ gives $-a^2\delta$. Since $\boldsymbol{\theta}_0$ lies on the same hyperbola, the chord from any $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}^\star$ is not tangent at $\boldsymbol{\theta}^\star$; hence $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 \notin \operatorname{Im} J(\boldsymbol{\theta}^\star)^\top$ and anchored ridge is biased for every initialisation off the diagonals with $\boldsymbol{\theta}_0 \notin \Theta^\star$ (Theorem D.11(c); here $\Theta^\star$ is a smooth curve globally, so the first-order condition of (b) holds at every limit point without reference to the ball). Explicitly, alignment would require $\delta = \theta_{1,0}\theta_1^\star - \theta_{2,0}\theta_2^\star$. For $\delta = 0$ and $p_0 > 0$ the flow manifold is a diagonal ray and the chord is aligned. For the linear model $f(x; \boldsymbol{\theta}) = (\theta_1 + \theta_2)x$, $J = (1, 1)$ is constant and the chord is aligned.

The dynamical form of the bias, for every smooth regulariser ϱ (so written to keep it apart from the radius R), is a slow drift along the interpolating set. For weight decay under a Morse–Bott hypothesis this two-phase behaviour was established, without rates, by Boursier et al. (2025) (see also Musat, 2025); Theorem D.12 allows any C^2 regulariser and bounds the remainder on the ball.

Theorem D.12 (The slow drift of a regulariser along the interpolating set)**.** *Let ϱ be C^1 on B and C^2 on a neighbourhood of $\Theta^\star \cap B$, and for $\lambda > 0$ let $\boldsymbol{\theta}_\lambda(t)$ be a solution of $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}} \mathcal{L} - \lambda \nabla \varrho$ from $\boldsymbol{\theta}_0$. Let $y(s)$ solve the projected gradient flow on $\Theta^\star \cap B$,*

$$\dot{y} \;=\; -(I - P_y)\nabla\varrho(y) \;=\; -\operatorname{grad}_{\Theta^\star}(\varrho|_{\Theta^\star})(y), \qquad y(0) = \boldsymbol{\theta}^\star,$$

and suppose $y([0, S]) \subset \Theta^\star \cap B(\boldsymbol{\theta}_0, R')$ for some $S > 0$ and $R' < R$. Then there are $\lambda_0, C_1 > 0$ such that for $\lambda < \lambda_0$: $\boldsymbol{\theta}_\lambda$ exists on $[0, S/\lambda]$ and stays in $\mathfrak{B}$; $\|\pi(\boldsymbol{\theta}_\lambda(s/\lambda)) - y(s)\| \le C_1 \lambda$ for all $s \in [0, S]$; and $\|\boldsymbol{\theta}_\lambda(s/\lambda) - y(s)\| \le C_1\big(\lambda + e^{-2c^2 s/\lambda}\big)$. In particular the regularised trajectory follows the unregularised one to within $O(\lambda)$ on the time scale $t = O(1)$, comes within $O(\lambda \log(1/\lambda))$ of $\boldsymbol{\theta}^\star$ at $t = \log(1/\lambda)/(2c^2)$, and then drifts along $\Theta^\star$ at speed λ along the intrinsic gradient flow of $\varrho|_{\Theta^\star}$.

Proof. Constants. Let $\Gamma_0 := \{\boldsymbol{\theta}(t) : t \geq 0\} \cup \{\boldsymbol{\theta}^\star\}$ be the closure of the unregularised trajectory, compact in $\mathfrak{B}$ (Proposition D.1), and let $\mathcal{N}_\varrho$ be the open neighbourhood of $\Theta^\star \cap B$ on which ϱ is C^2. Since $\mathfrak{B}$ is open and $y([0,S]) \subset \Theta^\star \cap B(\boldsymbol{\theta}_0, R') \subset \mathfrak{B} \cap \mathcal{N}_\varrho$, there is $\varepsilon > 0$ such that the closed ε-neighbourhood $\mathcal{K}$ of $\Gamma_0 \cup y([0,S])$ lies in $\mathfrak{B}$ and the closed ε-neighbourhood $\mathcal{K}'$ of $y([0,S])$ lies in $\mathcal{N}_\varrho$. Put $F(\boldsymbol{\theta}) := d\pi(\boldsymbol{\theta})\nabla\varrho(\boldsymbol{\theta})$ on $\mathfrak{B}$, continuous, and C^1 on $\mathfrak{B} \cap \mathcal{N}_\varrho$ by Theorem D.2; let $M_R := \sup_{\mathcal{K}} \|\nabla\varrho\|$, $M_F := \sup_{\mathcal{K}} \|F\|$, L_X a Lipschitz constant of X on $\mathcal{K}$, and L_F a Lipschitz constant of F on $\mathcal{K}'$. On $\Theta^\star \cap B$, $F(z) = (I - P_z)\nabla\varrho(z)$ is the intrinsic gradient of $\varrho|_{\Theta^\star}$ (Theorem D.2), so $\dot{y} = -F(y)$ and $\|\dot{y}\| \leq M_F$. Fix T_0 with $\frac{C}{c^2}\|r_0\| e^{-2c^2 T_0} \leq \varepsilon/8$ and then λ_0 so small that $\lambda_0 M_R e^{L_X T_0}/L_X \leq \varepsilon/2$, $\lambda_0 C^2 M_R/(2c^4) \leq \varepsilon/8$ and $C_1'\lambda_0 \leq \varepsilon/4$, with C_1' defined below.

Bootstrap. For $\lambda < \lambda_0$ let T^* be the supremum of the $T \leq S/\lambda$ such that $\boldsymbol{\theta}_\lambda$ exists on $[0,T]$ with $\boldsymbol{\theta}_\lambda(t) \in \mathcal{K}$ for $t \leq T$ and $\|z(t) - y(\lambda t)\| \leq \varepsilon/2$ for $T_0 \leq t \leq T$, where $z(t) := \pi(\boldsymbol{\theta}_\lambda(t))$. We show that on $[0,T^*)$ the two bounds hold with $\varepsilon/2$ and $\varepsilon/4$ in place of ε and $\varepsilon/2$; since $\boldsymbol{\theta}_\lambda$ cannot leave the compact set $\mathcal{K}$ without first leaving its interior, and cannot cease to exist while in $\mathcal{K}$, this forces $T^* = S/\lambda$ and gives the theorem.

Residual. With $r(t) := g(\boldsymbol{\theta}_\lambda(t)) - Y$, on $[0,T^*)$ we have $\dot{r} = -2K(\boldsymbol{\theta}_\lambda)\, r - \lambda J(\boldsymbol{\theta}_\lambda)\nabla\varrho(\boldsymbol{\theta}_\lambda)$, hence $\frac{d}{dt}\|r\| \leq -2c^2\|r\| + \lambda C M_R$ and

$$\|r(t)\| \;\leq\; \|r_0\| e^{-2c^2 t} + \frac{\lambda C M_R}{2c^2}, \qquad \|\boldsymbol{\theta}_\lambda(t) - z(t)\| \;\leq\; \rho(t) := \frac{C\,\|r(t)\|}{c^2},$$

the second by Proposition D.1 applied at the point $\boldsymbol{\theta}_\lambda(t) \in \mathfrak{B}$. For $t \geq T_0$ the choices above give $\rho(t) \leq \varepsilon/4$.

Slow equation. $z = \pi \circ \boldsymbol{\theta}_\lambda$ is C^1 with $\dot{z} = d\pi(\boldsymbol{\theta}_\lambda)\big(X(\boldsymbol{\theta}_\lambda) - \lambda\nabla\varrho(\boldsymbol{\theta}_\lambda)\big) = -\lambda F(\boldsymbol{\theta}_\lambda)$, because $d\pi X = 0$ (Theorem D.2); so $\|z(t) - z(t')\| \leq \lambda M_F |t - t'|$, and $z(0) = \pi(\boldsymbol{\theta}_0) = \boldsymbol{\theta}^\star = y(0)$.

Transient. For $t \leq \min(T_0, T^*)$, comparing with the unregularised flow, $\frac{d}{dt}\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}\| \leq L_X\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}\| + \lambda M_R$, so $\|\boldsymbol{\theta}_\lambda(t) - \boldsymbol{\theta}(t)\| \leq \lambda M_R(e^{L_X t} - 1)/L_X \leq \varepsilon/2$: $\boldsymbol{\theta}_\lambda(t)$ lies within $\varepsilon/2$ of Γ_0. Also $\|z(t) - y(\lambda t)\| \leq \|z(t) - z(0)\| + \|y(0) - y(\lambda t)\| \leq 2\lambda t M_F \leq 2\lambda T_0 M_F$.

Slow phase. For $T_0 \leq t < T^*$ write $s := \lambda t$ and $\zeta(s) := z(s/\lambda)$, so that $\zeta'(s) = -F(\zeta(s)) + e(s)$ with $e(s) := F(\zeta(s)) - F(\boldsymbol{\theta}_\lambda(s/\lambda))$. Both $\zeta(s)$, within $\varepsilon/2$ of $y(s)$, and $\boldsymbol{\theta}_\lambda(s/\lambda)$, within a further $\rho \leq \varepsilon/4$, lie in $\mathcal{K}'$, so $\|e(s)\| \leq L_F\rho(s/\lambda)$ and $\|F(\zeta) - F(y)\| \leq L_F\|\zeta - y\|$. Grönwall's inequality from $s_0 := \lambda T_0$ gives, for $s \in [s_0, \lambda T^*)$,

$$\begin{aligned}\|\zeta(s) - y(s)\| \;&\leq\; e^{L_F(s - s_0)}\Big(\|\zeta(s_0) - y(s_0)\| + \int_{s_0}^{s} L_F\,\rho(\sigma/\lambda)\,d\sigma\Big) \\ &\leq\; e^{L_F S}\Big(2\lambda T_0 M_F + L_F\frac{C}{c^2}\Big(\frac{\lambda\|r_0\|}{2c^2} + \frac{S\lambda C M_R}{2c^2}\Big)\Big) \;=:\; C_1'\lambda,\end{aligned}$$

using $\int_0^\infty e^{-2c^2\sigma/\lambda}d\sigma = \lambda/(2c^2)$. By the choice of λ_0, $\|z(t) - y(\lambda t)\| \leq \varepsilon/4$ for $T_0 \leq t < T^*$, and then $\|\boldsymbol{\theta}_\lambda(t) - y(\lambda t)\| \leq \rho(t) + \varepsilon/4 \leq \varepsilon/2$, so $\boldsymbol{\theta}_\lambda(t)$ lies within $\varepsilon/2$ of $y([0,S])$. This closes the bootstrap.

Conclusions. The transient and slow-phase bounds, plus $\rho(s/\lambda)$, give the two bounds with $C_1 := \max(2T_0 M_F, C_1') + \frac{C}{c^2}\max(\|r_0\|, \frac{CM_R}{2c^2})$; at $t = \log(1/\lambda)/(2c^2)$, $\|y(\lambda t) - \boldsymbol{\theta}^\star\| \leq M_F\lambda t$ gives the $O(\lambda\log(1/\lambda))$ statement. This is the deterministic case of the two-timescale theorem of Katzenberger (1991), with π in the role of the limit map and $-\nabla\varrho$ as the slow drift, used for a manifold of minimisers by Li et al. (2022a) and, for weight decay, by Boursier et al. (2025); the proof above is self-contained because π is smooth on $\mathfrak{B}$. □

Corollary D.7 (Drift criterion)**.** *A regulariser ϱ as in Theorem D.12 is drift-free at $\boldsymbol{\theta}^\star$, in the sense that $y \equiv \boldsymbol{\theta}^\star$, if and only if $(I - P_{\boldsymbol{\theta}^\star})\nabla\varrho(\boldsymbol{\theta}^\star) = \mathbf{0}$, that is, if and only if $\boldsymbol{\theta}^\star$ is a critical point of $\varrho|_{\Theta^\star}$; otherwise the regularised trajectory at slow time s is within $o(1)$, as $\lambda \downarrow 0$, of $y(s) \neq \boldsymbol{\theta}^\star$. For anchored ridge the condition reads $(I - P_{\boldsymbol{\theta}^\star})(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0) = \mathbf{0}$, the alignment condition of Theorem D.11; its insufficiency for unbiasedness (Remark D.7) is the difference between a critical point and a minimiser of $\varrho|_{\Theta^\star}$. Geodesic ridge is finite only on the set $\mathcal{W} \subseteq V$ of Proposition D.17, so it is not C^1 on B and enters Theorem D.12 through its localisations $\varrho := \chi\, d^2(\cdot, \boldsymbol{\theta}_0)$, χ a smooth cut-off equal to 1 near $\boldsymbol{\theta}^\star$ and compactly supported in a neighbourhood V_1 of $\boldsymbol{\theta}^\star$ on which geodesic*

ridge is C^2, so that ϱ is C^2 on B and equals geodesic ridge near $\boldsymbol{\theta}^\star$, where the slow flow of the drift-free case lives. Such a V_1 exists whenever E_Y is C^2 near Y: the fibre term is C^∞ near $\boldsymbol{\theta}^\star$ (Proposition D.17(ii)) and the flow term is $E_Y \circ g$ there; a sufficient condition is that minimising geodesics of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$ to points near $\boldsymbol{\theta}^\star$ be unique, depend continuously on the endpoint, and have no conjugate point at $\boldsymbol{\theta}^\star$ (Proposition E.5). For such ϱ, $\varrho|_{\Theta^\star} = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star) + d^2_{\mathrm{fibre}}(\boldsymbol{\theta}^\star, \cdot)$ near $\boldsymbol{\theta}^\star$, whose intrinsic gradient at $\boldsymbol{\theta}^\star$ vanishes and whose Hessian there is $2\,\mathrm{Id}$, so $\boldsymbol{\theta}^\star$ is a strict minimum of $\varrho|_{\Theta^\star}$ and an asymptotically stable equilibrium of the slow flow: under that condition geodesic ridge is drift-free. The ambient dynamics of geodesic ridge itself, started on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V_1$ rather than at $\boldsymbol{\theta}_0$, is Proposition E.5(iii).

Proof. The slow flow $\dot{y} = -F(y)$ with $F = \mathrm{grad}_{\Theta^\star}(\varrho|_{\Theta^\star})$ has the constant solution $y \equiv \boldsymbol{\theta}^\star$ if and only if $F(\boldsymbol{\theta}^\star) = \mathbf{0}$, and solutions of a C^1 vector field are unique, so this is the drift-free case; otherwise $y(s) \neq \boldsymbol{\theta}^\star$ for small $s > 0$ and Theorem D.12 places $\boldsymbol{\theta}_\lambda(s/\lambda)$ within $C_1(\lambda + e^{-2c^2 s/\lambda}) = o(1)$ of it. For anchored ridge $\nabla\varrho = 2(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, and the condition is the alignment condition; Remark D.7(A) is a critical point of the chord on $\Theta^\star$ that is a strict local maximum, so the slow flow is stationary at $\boldsymbol{\theta}^\star$ but unstable there, and the minimisers of Theorem D.11 lie elsewhere. For geodesic ridge on $\Theta^\star \cap V = \Theta(Y) \cap V$, $g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta})) = g|^{-1}_{\boldsymbol{\theta}_0}(Y) = \boldsymbol{\theta}^\star$ (Corollary D.4), so the flow term is the constant $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star)$ and the fibre term is the squared intrinsic distance from $\boldsymbol{\theta}^\star$ on the submanifold $\Theta(Y) \cap V$ (Proposition D.10); the squared Riemannian distance from a point is C^∞ near that point with vanishing gradient and Hessian $2\,\mathrm{Id}$ there, which is read off in normal coordinates about $\boldsymbol{\theta}^\star$ (do Carmo, 1992, Ch. 3). A strict non-degenerate minimum of $\varrho|_{\Theta^\star}$ is an asymptotically stable equilibrium of its gradient flow. □

D.6 Failure of non-canonical regularisation in the feature-learning regime

Remark D.11 (Standard ridge fails doubly in feature-learning networks). Standard ridge $\|\boldsymbol{\theta}\|^2$ fails for two independent reasons. *First*, it ignores the initialisation: when $\boldsymbol{\theta}_0 \notin \mathrm{Im}\, J^\top$ it is biased already *in the kernel regime*, where anchored ridge is canonical (He et al., 2020; Calvo-Ordoñez et al., 2026); for linear models anchored ridge is Tikhonov regularisation (Tikhonov, 1963) centred at $\boldsymbol{\theta}_0$, MAP under $\mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{2\lambda} I)$ (Pearce et al., 2020; Osband et al., 2018), and ridge regression (Hoerl & Kennard, 1970) is the case $\boldsymbol{\theta}_0 = \mathbf{0}$. *Second*, as anchored ridge with anchor $\mathbf{0}$ it is subject to Theorem D.11: whenever $\boldsymbol{\theta}^\star \notin \mathrm{Im}\, J(\boldsymbol{\theta}^\star)^\top$, $\boldsymbol{\theta}^\star$ is not a limit point of its minimisers.

Proposition D.13 (Static-anchor regularisers bias gradient flow)**.** *Under Assumption 2.1, let $\Sigma \in \mathbb{R}^{m\times m}$ be symmetric positive definite, let $\boldsymbol{v} \in \mathbb{R}^m$, let $R_{\Sigma,\boldsymbol{v}}(\boldsymbol{\theta}) := (\boldsymbol{\theta} - \boldsymbol{v})^\top \Sigma (\boldsymbol{\theta} - \boldsymbol{v})$, and for $\lambda > 0$ let $\boldsymbol{\theta}_\lambda$ be any minimiser of $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda R_{\Sigma,\boldsymbol{v}}(\boldsymbol{\theta})$ over Θ. Every limit point $\bar{\boldsymbol{\theta}}$ of $\{\boldsymbol{\theta}_\lambda\}$ as $\lambda \downarrow 0$ lies in $\Theta^\star$, minimises $R_{\Sigma,\boldsymbol{v}}$ over $\Theta^\star$, and, if $\bar{\boldsymbol{\theta}} \in B(\boldsymbol{\theta}_0, R)$, satisfies $\Sigma(\bar{\boldsymbol{\theta}} - \boldsymbol{v}) \in \mathrm{Im}\, J(\bar{\boldsymbol{\theta}})^\top$. In particular, the gradient-flow limit $\boldsymbol{\theta}^\star$ is a limit point of $\{\boldsymbol{\theta}_\lambda\}$ only if*

$$\Sigma(\boldsymbol{\theta}^\star - \boldsymbol{v}) \;\in\; \mathrm{Im}\, J(\boldsymbol{\theta}^\star)^\top .$$

Proof of Proposition D.13. Comparison with $\boldsymbol{\theta}^\star$ gives $R_{\Sigma,\boldsymbol{v}}(\boldsymbol{\theta}_\lambda) \leq R_{\Sigma,\boldsymbol{v}}(\boldsymbol{\theta}^\star)$, a bounded set as $\Sigma \succ 0$; Lemma C.9 gives the first two claims, and the argument of Theorem D.11(b) with gradient $2\Sigma(\boldsymbol{\theta} - \boldsymbol{v})$ the third; $\boldsymbol{\theta}^\star \in B(\boldsymbol{\theta}_0, R)$ gives the displayed condition. □

Proposition D.14 (Output-norm regularisers fail to select the representative)**.** *Under Assumption 2.1, write $C_0(\boldsymbol{\theta}_0) = B(Y, \rho)$ for the chart domain of Theorem D.4, and let $q : \mathbb{R}^n \to [0, \infty)$ be continuous with $q(g(\boldsymbol{\theta}_0)) = 0$ and strictly convex on $C_0(\boldsymbol{\theta}_0)$. The pure output-shaped regulariser $R(\boldsymbol{\theta}) := q(g(\boldsymbol{\theta}))$ satisfies, for every $\lambda > 0$ with $\lambda q(Y) < \rho^2/4$,*

$$\arg\min_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, R)} \big[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda R(\boldsymbol{\theta})\big] \;=\; \Theta(\boldsymbol{c}_\lambda) \cap B(\boldsymbol{\theta}_0, R)$$

for a single output value $\boldsymbol{c}_\lambda \in C_0(\boldsymbol{\theta}_0)$, the entire output fibre at the optimal output value. The vanishing-regularisation limit is $\Theta^\star \cap B(\boldsymbol{\theta}_0, R)$, not $\{\boldsymbol{\theta}^\star\}$, so a non-trivial fibre term is necessary in geodesic ridge (Proposition D.17).

Proof of Proposition D.14. The objective is $\Phi(g(\boldsymbol{\theta}))$ with $\Phi(\boldsymbol{c}) := \mathcal{L}(\boldsymbol{c}, Y) + \lambda q(\boldsymbol{c})$. For $\|\boldsymbol{c} - Y\| \geq \rho/2$, $\Phi(\boldsymbol{c}) \geq \rho^2/4 > \Phi(Y)$, so Φ attains its infimum in $B(Y, \rho/2)$, at a unique $\boldsymbol{c}_\lambda$

by strict convexity; the minimising set is $\Theta(\boldsymbol{c}_\lambda) \cap B(\boldsymbol{\theta}_0, R)$, non-empty (it contains $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}_\lambda)$) and an $(m-n)$-dimensional submanifold (Proposition D.10). As $\lambda \downarrow 0$, $\|\boldsymbol{c}_\lambda - Y\|^2 \le \lambda q(Y) \to 0$. □

D.7 Kernel-regime reduction of the canonical objects

Proposition D.15 (Kernel-regime reduction). *In the kernel regime of Definition 2.1, write $J := J(\boldsymbol{\theta}_0)$, the constant Jacobian on $B(\boldsymbol{\theta}_0, R)$, and $K := JJ^\top \succ 0$. Then the canonical objects of Section 4 reduce as follows:*

- *(a)* Affine basin manifold: $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \left(\boldsymbol{\theta}_0 + \text{Im}(J^\top)\right) \cap \mathfrak{B}$, *a convex set (Corollary D.3);*
- *(b)* Closed-form representative: $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) = \boldsymbol{\theta}_0 + J^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$;
- *(c)* RKHS energy: $E_Y(\boldsymbol{c}) = E(\boldsymbol{c}) = (\boldsymbol{c} - g(\boldsymbol{\theta}_0))^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0)) = \|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\|_{\text{RKHS}}^2$ *for every label vector Y, on the common domain of the two functions;*
- *(d)* Pythagorean form: *with V of Theorem D.4 convex, geodesic ridge satisfies $d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ (anchored ridge) wherever it is finite;*
- *(e)* Geodesic distance: $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ *for every $\boldsymbol{\theta} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$: K is constant on the flow manifold and its geodesics are straight segments;*
- *(f)* NTK Gaussian-process prior: *the prior $p(\boldsymbol{c} \mid X) \propto \exp(-\beta E(\boldsymbol{c}))$ of Remark 4.1 reduces to $g(\boldsymbol{\theta}) \sim \mathcal{N}(g(\boldsymbol{\theta}_0), (2\beta)^{-1}K)$, the marginal at X of the NTK Gaussian process $\mathcal{GP}(f(\cdot; \boldsymbol{\theta}_0), (2\beta)^{-1}k_{\text{NTK}})$, and it is the exact pushforward of $\mathcal{N}(\boldsymbol{\theta}_0, (2\beta)^{-1}I)$ under the affine extension $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ of g from the ball.*

Proof of Proposition D.15. Throughout, $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ on $B(\boldsymbol{\theta}_0, R)$. *(a)* is Corollary D.3. *(b)* $\boldsymbol{\theta} - \boldsymbol{\theta}_0 = J^\top \boldsymbol{a}$ and $g(\boldsymbol{\theta}) = \boldsymbol{c}$ force $\boldsymbol{a} = K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$, and Corollary D.4 identifies this point with $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$. *(e)* Segments from $\boldsymbol{\theta}_0$ lie in the convex $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, so the intrinsic distance is at most the chord, and at least it by Lemma C.6.

(c) By Proposition F.1(i), (e) and (b), $E_Y(\boldsymbol{c}) = \left\|J^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))\right\|^2 = (\boldsymbol{c} - g(\boldsymbol{\theta}_0))^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$, free of Y; equivalently the chart metric is the constant $G = K^{-1}$, whose geodesics are straight lines (do Carmo, 1992, Ch. 2). For $\boldsymbol{c} \in \mathcal{N}$ the problem with labels $\boldsymbol{c}$ has the same J, so Corollary D.3 gives $\tau(\boldsymbol{c}) = \boldsymbol{\theta}_0 + J^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$ on the same plane, and $E(\boldsymbol{c})$ is the same quadratic form.

(d) In the kernel regime transversality and injectivity of g hold on all of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, so V may be any convex neighbourhood of $\boldsymbol{\theta}^\star$ in $\mathfrak{B}$. By Lemma C.5, $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 = \left\|g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta})) - \boldsymbol{\theta}_0\right\|^2 + \left\|\boldsymbol{\theta} - g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))\right\|^2$; the first term is $E_Y(g(\boldsymbol{\theta}))$ by (c), and the second is the fibre term, the fibre being the affine space $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta})) + \ker J$ and V convex (Lemma C.6).

(f) By (c), $\exp(-\beta E(\boldsymbol{c}))$ is the unnormalised density of $\mathcal{N}(g(\boldsymbol{\theta}_0), (2\beta)^{-1}K)$, $K = k_{\text{NTK}}(X, X)$, and the affine image of $\mathcal{N}(\boldsymbol{\theta}_0, (2\beta)^{-1}I)$ is $\mathcal{N}(g(\boldsymbol{\theta}_0), (2\beta)^{-1}JJ^\top)$: the pushforward is exact at the level of measures. □

Corollary D.8 (Anchored ridge is geodesic ridge in the kernel regime). *Under the hypotheses of Proposition D.15, geodesic ridge of Definition 4.1 coincides with anchored ridge: $d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ for every $\boldsymbol{\theta} \in V$ with $g(\boldsymbol{\theta}) \in C_0(\boldsymbol{\theta}_0)$.*

Proof of Corollary D.8. Proposition D.15(e) identifies the flow term of Definition 4.1 with the squared chord from $\boldsymbol{\theta}_0$ to $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$, the affine fibre identifies the fibre term with the squared chord from $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ to $\boldsymbol{\theta}$, and the two chords are orthogonal (Lemma C.5); Proposition D.15(d) reassembles them into $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$. □

D.8 Uniqueness of the canonical regulariser

The canonical regulariser is determined by three derived inputs – the rule that prices motion, the class of admissible motions, and the anchor – and one principle, least energy, which the paper uses in the same form for geodesic ridge, for arc ridge (Theorem E.7(d)) and for anchored ridge. Least energy, and the additive pricing of the vertical leg by the same rule, are modelling choices rather than

consequences of training: the second does not affect the regularised estimator (Proposition D.17), and the first is made precise by the maximality characterisation (U2) of Theorem D.16. The class is derived in Lemma E.2 (Appendix E).

Lemma D.6 (The measurement rule is forced)**.** *A* measurement rule *assigns to each piecewise-C^1 path $\gamma : [0,1] \to \Theta$ the energy $\mathfrak{E}(\gamma) = \int_0^1 \varrho\big(\gamma(t), \dot\gamma(t)\big)\, dt$ for a continuous density $\varrho : \Theta \times \mathbb{R}^m \to [0,\infty)$. Suppose that (N)* naturality*: $\varrho(\varphi(\boldsymbol\theta), D\varphi\, \boldsymbol v) = \varrho(\boldsymbol\theta, \boldsymbol v)$ for every isometry φ of the Euclidean metric of effective-parameter space and every $\boldsymbol\theta$ with $\boldsymbol\theta, \varphi(\boldsymbol\theta) \in \Theta$; and (C)* calibration*: for every label vector $Y' \in \mathbb{R}^n$ and every $\boldsymbol\theta \in B$, writing $X_{Y'}(\boldsymbol\theta) := -\nabla_{\boldsymbol\theta} \mathcal{L}\left(g\left(\boldsymbol\theta\right), Y'\right)$ and $\phi_t^{Y'}$ for the gradient flow of $\mathcal{L}\left(g\left(\cdot\right), Y'\right)$,*

$$\varrho\big(\boldsymbol\theta, X_{Y'}(\boldsymbol\theta)\big) \;=\; -\frac{d}{dt}\Big|_{t=0} \mathcal{L}\left(g\left(\phi_t^{Y'}(\boldsymbol\theta)\right), Y'\right),$$

the rate at which training on the labels Y' decreases its loss as it passes through $\boldsymbol\theta$. Then $\varrho(\boldsymbol\theta, \boldsymbol v) = \|\boldsymbol v\|^2$ for all $\boldsymbol\theta \in B$ and all $\boldsymbol v \in \mathbb{R}^m$, for every $n \geq 1$, with no free scale. Under a fixed preconditioner $D \succ 0$ – gradient flow $\dot{\boldsymbol\theta} = -D\nabla_{\boldsymbol\theta}\mathcal{L}$, whose metric is $\langle \boldsymbol v, D^{-1}\boldsymbol v\rangle$ – the same argument gives $\varrho(\boldsymbol\theta, \boldsymbol v) = \boldsymbol v^\top D^{-1} \boldsymbol v$.

Proof of Lemma D.6. Step 1. By (N) with translations, ϱ is independent of $\boldsymbol\theta \in B$; with the rotations $\boldsymbol\theta'' \mapsto \boldsymbol\theta_0 + Q(\boldsymbol\theta'' - \boldsymbol\theta_0)$, $Q \in O(m)$, which preserve B, it is $f(\|\boldsymbol v\|)$ for a continuous f.

Step 2. By Lemma C.2, (C) reads $f(\|X_{Y'}(\boldsymbol\theta)\|) = \|X_{Y'}(\boldsymbol\theta)\|^2$, and $X_{Y'}(\boldsymbol\theta) = 2J\left(\boldsymbol\theta\right)^\top \left(Y' - g\left(\boldsymbol\theta\right)\right)$ ranges over the n-dimensional $\operatorname{Im} J\left(\boldsymbol\theta\right)^\top$ (Lemma C.1), so its norm takes every value in $[0,\infty)$ and $f(s) = s^2$, with no free scale.

Preconditioner. The loss decreases at rate $\dot{\boldsymbol\theta}^\top D^{-1}\dot{\boldsymbol\theta}$, the isometries $\boldsymbol\theta'' \mapsto Q\boldsymbol\theta'' + \boldsymbol a$ with $Q^\top D^{-1} Q = D^{-1}$ act transitively on D^{-1}-spheres, and the velocities fill $D \operatorname{Im} J\left(\boldsymbol\theta\right)^\top$; Steps 1–2 repeat. □

Calibration on the realised trajectory alone suffices when the learning rate $\eta > 0$ is free: the speeds $\eta\,\|\nabla_{\boldsymbol\theta}\mathcal{L}(\boldsymbol\theta(t))\|$ exhaust $(0,\infty)$, the gradient norm being continuous, positive at finite times when $\boldsymbol\theta_0 \notin \Theta^\star$ and tending to zero. The lemma excludes the natural rivals. The length density $\|\boldsymbol v\|$ satisfies (N) but fails (C). The density $\|\boldsymbol v\|^2 + (\boldsymbol w(\boldsymbol\theta)^\top \boldsymbol v)^2$, $\boldsymbol w$ a continuous section of $\ker J\left(\cdot\right)$, satisfies (C), the calibrating velocities being horizontal, yet fails (N), which tests every velocity. The constant quadratic densities $a\,\|\boldsymbol v\|^2 + (J\boldsymbol v)^\top B (J\boldsymbol v)$ of the kernel regime, $a > 0$, $B \in \operatorname{Sym}_n^+$, whose least energies over all paths are the regularisers $R_{a,B}$ of Proposition D.18, satisfy (N) only when $aI + J^\top B J$ is a multiple of the identity, which forces $B = 0$ because $J^\top B J$ has rank at most $n < m$; (N) alone therefore removes every $B \neq 0$, after which (C) fixes $a = 1$. Calibration alone leaves the one-parameter subfamily $B = (1-a)K^{-1}$, $0 < a \leq 1$, on which the two terms compensate exactly along horizontal velocities, $a\,\boldsymbol u^\top K \boldsymbol u + \boldsymbol u^\top K B K \boldsymbol u = \boldsymbol u^\top K \boldsymbol u$; naturality, which tests the vertical velocities that calibration never sees, is what removes it. Naturality carries more than the vertical leg. Calibration tests horizontal velocities only, whereas away from $\boldsymbol\theta^\star$ the flow manifold is tilted out of the horizontal space (Lemma D.3), so the paths over which Theorem D.16 minimises have velocities that no gradient flow realises; the price of those velocities, and with it the flow term itself, is fixed by (N), not by calibration. Admitting only the velocities calibration prices, the horizontal paths, gives instead the sub-Riemannian energy, which is biased (Remark D.4).

Remark D.12 (An alternative derivation of the rule). Let $\varrho : \Theta \times \mathbb{R}^m \to [0,\infty)$ be continuous, with $\varrho(\boldsymbol\theta, \cdot)$ convex and $\varrho(\boldsymbol\theta, \boldsymbol 0) = 0$, let $\varrho^*(\boldsymbol\theta, \boldsymbol w) := \sup_{\boldsymbol v}[\langle \boldsymbol v, \boldsymbol w\rangle - \varrho(\boldsymbol\theta, \boldsymbol v)] \in [0,\infty]$ be its Legendre transform, and for a C^1 function $\mathcal{L}$ on Θ and a C^1 curve $\gamma : [0,T] \to \Theta$ define the action

$$\mathcal{A}(\gamma) \;:=\; \int_0^T \Big[\varrho\big(\gamma, \dot\gamma\big) + \varrho^*\big(\gamma, -\nabla\mathcal{L}(\gamma)\big)\Big] dt \;+\; \mathcal{L}(\gamma(T)) - \mathcal{L}(\gamma(0)).$$

By Fenchel–Young the integrand is non-negative, with equality iff $\dot\gamma \in \partial_{\boldsymbol w}\varrho^*(\gamma, -\nabla\mathcal{L}(\gamma))$; for $\varrho = \frac{1}{2}\|\boldsymbol v\|^2$ the zero-action curves are the gradient-flow curves. Conversely, if every gradient-flow solution of every C^1 loss has zero action, then $\varrho(\boldsymbol\theta, \boldsymbol v) = \frac{1}{2}\|\boldsymbol v\|^2$ (the factor, the usual normalisation of a dissipation potential, is absorbed into λ). Indeed, the linear losses $-\langle \boldsymbol w, \cdot\rangle$ make $h := \varrho(\cdot, \boldsymbol w) + \varrho^*(\cdot, \boldsymbol w) - \|\boldsymbol w\|^2 \geq 0$ vanish almost everywhere (Fubini), hence everywhere by lower semicontinuity; so $\boldsymbol w \in \partial\varrho^*(\boldsymbol\theta, \boldsymbol w)$ for all $\boldsymbol w$, and the subgradient inequality in both directions gives $|\psi(\boldsymbol w') - \psi(\boldsymbol w)| \leq$

$\frac{1}{2}\|\boldsymbol{w}'-\boldsymbol{w}\|^2$ for $\psi := \varrho^*(\boldsymbol{\theta},\cdot)-\frac{1}{2}\|\cdot\|^2$, so ψ is constant; Fenchel–Moreau and $\varrho(\boldsymbol{\theta},\mathbf{0})=0$ conclude. This trades the naturality of Lemma D.6 for convexity of the density and loss-agnosticism of the optimiser: the same optimiser is run on every loss it is given, so its dissipation potential cannot depend on the loss, and the losses whose gradients are vertical pin the potential on the vertical velocities that calibration never tests. Restricting the hypothesis to the squared-error losses of all label vectors pins $\varrho(\boldsymbol{\theta},\cdot)$ on the horizontal velocities $\operatorname{Im} J(\boldsymbol{\theta})^\top$ only, the velocities those losses realise.

Theorem D.16 (Three characterisations of the flow term)**.** *Let $\mathcal{A}$ be the class of piecewise-C^1 paths $\gamma:[0,1]\to\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ with $\gamma(0)=\boldsymbol{\theta}_0$ – the runs of admissible forces from $\boldsymbol{\theta}_0$ (Lemma E.2) – and for $\boldsymbol{\theta}\in\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ let*

$$R_{\mathrm{flow}}(\boldsymbol{\theta}) \;:=\; \inf\Big\{\int_0^1 \|\dot\gamma(t)\|^2\,dt:\ \gamma\in\mathcal{A},\ \gamma(1)=\boldsymbol{\theta}\Big\},$$

the least energy, in the rule of Lemma D.6, of an outcome-preserving motion from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$. Then $R_{\mathrm{flow}}=d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0,\cdot)$, finite on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, and $u_0:=d_{\mathrm{flow}}(\boldsymbol{\theta}_0,\cdot)$ is characterised in each of the following ways.

- *(U1)* Dynamic programming. *$u_0(\boldsymbol{\theta}_0)=0$ and $u_0(\boldsymbol{\theta})=\inf_{\boldsymbol{\theta}'\in\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}[u_0(\boldsymbol{\theta}')+d_{\mathrm{flow}}(\boldsymbol{\theta}',\boldsymbol{\theta})]$; and u_0 is the unique $u:\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)\to[0,\infty)$ with $u(\boldsymbol{\theta}_0)=0$ that is dominated by every run, $u(\gamma(1))\le L(\gamma)$ for all $\gamma\in\mathcal{A}$, and approached by some run: for every $\boldsymbol{\theta}$ and $\varepsilon>0$ some $\gamma\in\mathcal{A}$ with $\gamma(1)=\boldsymbol{\theta}$ has $L(\gamma)\le u(\boldsymbol{\theta})+\varepsilon$.*
- *(U2)* Maximality. *u_0 is the largest $u:\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)\to[0,\infty)$ with $u(\boldsymbol{\theta}_0)=0$ and $u(\gamma(1))-u(\gamma(0))\le L(\gamma)$ for every piecewise-C^1 path γ in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$: the largest function vanishing at $\boldsymbol{\theta}_0$ that never grows faster than the length of any outcome-preserving motion. For locally Lipschitz u the growth condition is equivalent to $\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}u\right\|\le 1$ almost everywhere.*
- *(U3)* Eikonal. *u_0 is a viscosity solution of $\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}u\right\|=1$ on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)\setminus\{\boldsymbol{\theta}_0\}$, and it is the maximal non-negative viscosity solution with $u(\boldsymbol{\theta}_0)=0$: every continuous viscosity subsolution vanishing at $\boldsymbol{\theta}_0$ lies below it.*

Equivalently, $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0,\cdot)$ is the greatest function of $\boldsymbol{\theta}$ that never exceeds the early-stopping penalty s^2 of any outcome-preserving run reaching $\boldsymbol{\theta}$; Theorem E.7(d) is the dual statement, arc ridge being the least path-determined penalty that dominates the intrinsic distance of every geometry consistent with the run.

Proof of Theorem D.16. Value function. $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is connected (Theorem D.3(i)), so the paths in $\mathcal{A}$ ending at $\boldsymbol{\theta}$ form a non-empty, reparametrisation-closed family whose infimal length is $d_{\mathrm{flow}}(\boldsymbol{\theta}_0,\boldsymbol{\theta})$ (corners are smoothable at arbitrarily small cost); Lemma C.7 gives $R_{\mathrm{flow}}=d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0,\cdot)$.

(U2). If u has the growth condition, $u(\boldsymbol{\theta})\le L(\gamma)$ for every run ending at $\boldsymbol{\theta}$, so $u\le u_0$; u_0 has it by the triangle inequality (and the conclusion persists if the growth condition is imposed on runs from $\boldsymbol{\theta}_0$ only). For locally Lipschitz u: the growth condition bounds $|du(\boldsymbol{v})|\le 1$ at points of differentiability (Rademacher in charts (Evans & Gariepy, 2015, Ch. 3)), and conversely, if $\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}u\right\|\le 1$ a.e., chart mollifications $u_\delta\to u$ satisfy $\|\operatorname{grad}u_\delta\|_g\le 1+\varepsilon$ for small δ (continuity of the metric, convexity of the norm), which gives the growth condition along each C^1 piece as $\delta,\varepsilon\downarrow 0$.

(U1). The identity holds for every function 1-Lipschitz for the intrinsic distance, so it alone fixes nothing; domination by every run gives $u\le u_0$, approach by some run $u\ge u_0$, and u_0 has both properties by definition.

(U3). Viscosity sub- and supersolutions are as in (Crandall & Lions, 1983; Lions, 1982). *Subsolution:* if $u_0-\varphi$ has a local maximum at $\boldsymbol{\theta}$, then along unit-speed geodesics $\varphi(\sigma(s))-\varphi(\boldsymbol{\theta})\ge u_0(\sigma(s))-u_0(\boldsymbol{\theta})\ge -s$, so $\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}\varphi(\boldsymbol{\theta})\right\|\le 1$. *Supersolution at $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$:* if $u_0-\varphi$ has a local minimum at $\boldsymbol{\theta}$, take a run γ ending at $\boldsymbol{\theta}$ with $L(\gamma)\le u_0(\boldsymbol{\theta})+r^2$ and its point $\boldsymbol{\theta}'$ at arc length r before the end; then $\varphi(\boldsymbol{\theta}')-\varphi(\boldsymbol{\theta})\le u_0(\boldsymbol{\theta}')-u_0(\boldsymbol{\theta})\le r^2-r$, while integrating along a minimising geodesic gives $\varphi(\boldsymbol{\theta}')-\varphi(\boldsymbol{\theta})\ge -(\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}\varphi(\boldsymbol{\theta})\right\|+\omega(r))r$, so $\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}\varphi(\boldsymbol{\theta})\right\|\ge 1-r-\omega(r)\to 1$, as for Riemannian distance functions (Mantegazza & Mennucci, 2002). *Maximality:* a continuous

viscosity subsolution of this coercive equation is locally Lipschitz and satisfies it classically almost everywhere (Lions, 1982; Crandall & Lions, 1983), so (U2) gives $u \le u_0$. On the incomplete $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ maximality, not comparison, is the correct formulation: on the punctured disc $\{0 < |x| < \rho\}$ the functions $\min(|x|, a+\rho-|x|)$, $0 \le a < \rho$, are further non-negative solutions vanishing at the origin.

Last sentence. A function lies below $L(\gamma)^2 = s^2$ for every run ending at $\boldsymbol{\theta}$ iff it lies below $R_{\mathrm{flow}}(\boldsymbol{\theta})$, which itself does. □

Proposition D.17 (Extension off the basin manifold, and what the estimator sees)**.** *Let V, $C_0(\boldsymbol{\theta}_0) = B(Y,\rho)$ and $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ be as in Theorem D.4, let $\mathcal{W}$ be the set of $\boldsymbol{\theta} \in V$ with $g(\boldsymbol{\theta}) \in C_0(\boldsymbol{\theta}_0)$ lying in the connected component of $\Theta(g(\boldsymbol{\theta})) \cap V$ containing $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ (Remark D.5), and let $E_Y(\boldsymbol{c}) := d^2_{\mathrm{flow}}\left(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)$ on $C_0(\boldsymbol{\theta}_0)$.*

(i) Let $S : \mathcal{W} \to [0,\infty)$ satisfy $S(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$, and let $R' := E_Y \circ g + S$ on $\mathcal{W}$, $R' := +\infty$ elsewhere. Then for every $\lambda > 0$,

$$\inf_{\boldsymbol{\theta}\in\Theta}\left[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda R'(\boldsymbol{\theta})\right] \;=\; \inf_{\boldsymbol{c}\in C_0(\boldsymbol{\theta}_0)}\left[\|\boldsymbol{c}-Y\|^2 + \lambda E_Y(\boldsymbol{c})\right].$$

For $\lambda E_Y(Y) < \rho^2/4$ the right-hand side is attained, the minimisers of $\mathcal{L} + \lambda R'$ are exactly the points $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}_\lambda)$ with $\boldsymbol{c}_\lambda$ a minimiser of the right-hand side, and every selection $\lambda \mapsto \boldsymbol{c}_\lambda$ has $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}_\lambda) \to \boldsymbol{\theta}^\star$ as $\lambda \downarrow 0$.

(ii) Geodesic ridge is the case $S = d^2_{\mathrm{fibre}}\left(g|_{\boldsymbol{\theta}_0}^{-1}(g(\cdot)), \cdot\right)$: the vertical leg priced by the rule of Lemma D.6, this S being the least energy $\inf_\psi \int_0^1 \left\|\dot\psi(t)\right\|^2 dt$ over piecewise-C^1 paths ψ in $\Theta(g(\boldsymbol{\theta})) \cap V$ from $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ to $\boldsymbol{\theta}$ (Lemma C.7, Proposition D.10). It is C^∞ on a neighbourhood of $\boldsymbol{\theta}^\star$ in Θ, on which it vanishes to second order along $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$, and in the kernel regime it makes R' anchored ridge (Proposition D.15(d)).

Proof of Proposition D.17. (i) On $\mathcal{W} \cap \Theta(\boldsymbol{c})$ the loss is $\|\boldsymbol{c}-Y\|^2$ and $R' \ge E_Y(\boldsymbol{c})$ with equality exactly at $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) \in \mathcal{W}$; minimising over $\boldsymbol{c}$ gives the identity and the description of the minimisers. E_Y is finite ($\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is connected) and continuous on $C_0(\boldsymbol{\theta}_0)$, $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ being 1-Lipschitz for the intrinsic distance and $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ a diffeomorphism onto the open $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ (Theorem D.4(iii)). For $\|\boldsymbol{c}-Y\| \ge \rho/2$ the objective exceeds $\lambda E_Y(Y)$, so the minimum is attained in $\overline{B}(Y,\rho/2)$, with $\|\boldsymbol{c}_\lambda - Y\|^2 \le \lambda E_Y(Y) \to 0$ and $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}_\lambda) \to \boldsymbol{\theta}^\star$.

(ii) The paths in $\Theta(\boldsymbol{c}) \cap V$ from $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ to $\boldsymbol{\theta} \in \mathcal{W}$ form a non-empty (Remark D.5), reparametrisation-closed family, so Lemma C.7 applies. For regularity, straighten the submersion near $\boldsymbol{\theta}^\star$ by a diffeomorphism $\Phi : V' \to C_0' \times D$ with $g = \mathrm{pr}_1 \circ \Phi$ and $\Phi(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})) = (\boldsymbol{c}, \mathbf{0})$; each fibre becomes $\{\boldsymbol{c}\} \times D$ with a metric $h_{\boldsymbol{c}}$ smooth in $(\boldsymbol{c}, \boldsymbol{w})$. By the inverse function theorem $(\boldsymbol{c}, \boldsymbol{v}) \mapsto (\boldsymbol{c}, \exp_{\mathbf{0}}^{h_{\boldsymbol{c}}}(\boldsymbol{v}))$ is a diffeomorphism near $(Y, \mathbf{0})$ onto uniform normal balls, in which the radial geodesic is the unique shortest curve (do Carmo, 1992, Ch. 3); so $S = h_{\boldsymbol{c}}(\boldsymbol{v}, \boldsymbol{v})$ is C^∞ near $\boldsymbol{\theta}^\star$, vanishing to second order exactly on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V'$. The kernel case is Proposition D.15(d). □

The regularised estimator therefore depends on the extension only through the selection $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$, never through the shape of S: it is the representative on the basin manifold of the shrunk outputs $\boldsymbol{c}_\lambda$, whatever the price of leaving that manifold. The additive form and the vertical rule are conventions without effect on the estimator, retained because they price the channel invisible to the loss by the same rule as the channel the loss sees, give an objective that is C^2 across $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ near $\boldsymbol{\theta}^\star$ whenever E_Y is C^2 near Y, and are what the ambient gradient flow of $\mathcal{L} + \lambda R$ responds to (Proposition E.5). What is not a convention is the anchor $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$: Theorem 4.2(b) forces it for the training problem, and Theorem F.4 shows that unbiasedness for every label vector at once forces the training map τ in its place.

Proposition D.18 (Unbiasedness alone does not pin the regulariser)**.** *Let $\mathcal{W}$ be as in Proposition D.17, let $\Psi : C_0(\boldsymbol{\theta}_0) \to [0,\infty)$ be continuous and $b > 0$, and let $R_{\Psi,b}(\boldsymbol{\theta}) := \Psi(g(\boldsymbol{\theta})) + b\, d^2_{\mathrm{fibre}}\left(g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta})), \boldsymbol{\theta}\right)$ on $\mathcal{W}$, $R_{\Psi,b} := +\infty$ elsewhere. Then every selection of minimisers of*

$\mathcal{L}(g(\cdot), Y) + \lambda R_{\Psi,b}$ *over* Θ *– or of* ϵ_λ*-minimisers with* $\epsilon_\lambda/\lambda \to 0$*, which always exist – converges to* $\boldsymbol{\theta}^\star$ *as* $\lambda \downarrow 0$*. The family is infinite-dimensional,* Ψ *being free. In the kernel regime, with* V *convex as in Proposition D.15(d), it contains the restrictions to* $\mathcal{W}$ *of the trajectory-blind quadratic regularisers*

$$R_{a,B}(\boldsymbol{\theta}) \;:=\; a\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \big(J(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big)^\top B\,\big(J(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big), \qquad a>0,\; B\in \mathrm{Sym}_n^+,$$

Sym_n^+ *the cone of symmetric positive semi-definite* $n\times n$ *matrices: with* $\boldsymbol{c} = g(\boldsymbol{\theta})$ *and* $P := J^\top K^{-1} J$*,*

$$\begin{aligned} R_{a,B}(\boldsymbol{\theta}) \;&=\; \big[a\,(\boldsymbol{c}-g(\boldsymbol{\theta}_0))^\top K^{-1}(\boldsymbol{c}-g(\boldsymbol{\theta}_0)) + (\boldsymbol{c}-g(\boldsymbol{\theta}_0))^\top B\,(\boldsymbol{c}-g(\boldsymbol{\theta}_0))\big] \\ &\quad + a\,\|(I-P)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 \end{aligned}$$

(Proposition D.15(b),(d)), and distinct (a,B) *give distinct regularisers and distinct Gaussian MAP parameter priors* $\mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{2\lambda}(aI_m + J^\top B J)^{-1})$*. Unbiasedness in the limit* $\lambda\downarrow 0$ *therefore constrains the fibre channel alone; the horizontal channel is fixed by Theorem D.16: a flow term* $\Psi \neq E_Y$ *either exceeds the energy of some outcome-preserving motion or is not the greatest function dominated by all of them.*

Proof of Proposition D.18. Let $\Phi_\lambda(\boldsymbol{\theta}_\lambda) \le \inf\Phi_\lambda + \epsilon_\lambda$, $\Phi_\lambda := \mathcal{L}(g(\cdot), Y) + \lambda R_{\Psi,b}$. Comparison with $\boldsymbol{\theta}^\star = g|_{\boldsymbol{\theta}_0}^{-1}(Y) \in \mathcal{W}$ gives $\mathcal{L}(g(\boldsymbol{\theta}_\lambda), Y) \le \lambda\Psi(Y) + \epsilon_\lambda \to 0$ and $R_{\Psi,b}(\boldsymbol{\theta}_\lambda) \le \Psi(Y) + \epsilon_\lambda/\lambda$; so $\boldsymbol{\theta}_\lambda \in \mathcal{W}$, $g(\boldsymbol{\theta}_\lambda) \to Y$, and by Lemma C.6

$$b\,\big\|\boldsymbol{\theta}_\lambda - g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}_\lambda))\big\|^2 \;\le\; R_{\Psi,b}(\boldsymbol{\theta}_\lambda) - \Psi(g(\boldsymbol{\theta}_\lambda)) \;\le\; \Psi(Y) - \Psi(g(\boldsymbol{\theta}_\lambda)) + \epsilon_\lambda/\lambda \;\to\; 0,$$

while $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}_\lambda)) \to \boldsymbol{\theta}^\star$; so $\boldsymbol{\theta}_\lambda \to \boldsymbol{\theta}^\star$, using of Ψ only finiteness at Y and continuity.

In the kernel regime $J(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \boldsymbol{c} - g(\boldsymbol{\theta}_0)$, and Lemma C.5 with Proposition D.15(b),(d) splits $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$ into $(\boldsymbol{c}-g(\boldsymbol{\theta}_0))^\top K^{-1}(\boldsymbol{c}-g(\boldsymbol{\theta}_0))$ and $\big\|\boldsymbol{\theta} - g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\big\|^2 = d^2_{\mathrm{fibre}}\big(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}), \boldsymbol{\theta}\big)$, giving the display with $b=a$. If $aI_m + J^\top B J = a' I_m + J^\top B' J$, rank forces $a=a'$, and $K(B-B')K=0$ forces $B=B'$. The last sentence is Theorem D.16(U2) read through the chart. □

Remark D.13 (What each input excludes). *(i) The admissible class.* Replacing the paths in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ by all paths in Θ gives the least energy $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$, attained by the segment (Lemmas C.7 and C.6): anchored ridge, biased whenever the alignment condition fails (Theorem D.11). Replacing them by the training curve alone gives arc ridge, s^2 (Proposition E.2), a function of the run rather than of a point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$: the surrogate of Appendix E. Replacing them by the horizontal paths gives the squared sub-Riemannian distance $d^2_{\mathrm{SR}}(\boldsymbol{\theta}_0, \cdot)$, label-free but biased (Remark D.4). *(ii) Maximality and calibration.* Dropping maximality in (U2) admits $\alpha\, d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ for every $\alpha \in (0,1)$, and the chord $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$ restricted to $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, whose square root is 1-Lipschitz for the intrinsic distance and lies below $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ (Lemma C.6); dropping calibration admits $\alpha\, d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ for every $\alpha > 0$, the least energies of the rules $\alpha\,\|\boldsymbol{v}\|^2$. *(iii) Naturality.* Dropping (N) admits the density $\|\boldsymbol{v}\|^2 + (\boldsymbol{w}(\boldsymbol{\theta})^\top \boldsymbol{v})^2$ with $\boldsymbol{w}$ a section of $\ker J(\cdot)$, calibrated on every horizontal velocity yet pricing vertical motion, and with it the fibre term, differently. *(iv) The extension.* An anchor $a(\boldsymbol{c}) \in \Theta(\boldsymbol{c}) \cap V$, continuous in $\boldsymbol{c}$, with $a \neq g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ gives a fibre term that does not vanish along the trajectory, failing Theorem 4.2(b), and the argument of Proposition D.17(i) with a in place of $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ makes the estimator $a(\boldsymbol{c}_\lambda) \to a(Y)$, biased unless $a(Y) = \boldsymbol{\theta}^\star$. The ambient chord $\big\|\boldsymbol{\theta} - g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))\big\|^2$ drops vertical admissibility, pricing a motion that leaves the fibre; $2\, d^2_{\mathrm{fibre}}\big(g|_{\boldsymbol{\theta}_0}^{-1}(g(\cdot)), \cdot\big)$ drops the single rule; and the single-clock alternative $\big(d_{\mathrm{flow}}\big(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))\big) + d_{\mathrm{fibre}}\big(g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta})), \boldsymbol{\theta}\big)\big)^2$ fails the kernel-regime Pythagorean identity of Proposition D.15(d) and, through its cross term $2\, d_{\mathrm{flow}}\big(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))\big)\; d_{\mathrm{fibre}}\big(g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta})), \boldsymbol{\theta}\big)$, has a cone singularity along the fibre at $g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ wherever the flow term is positive, so that $\mathcal{L} + \lambda R$ is not differentiable at its own minimiser. None of these last three changes the estimator, by Proposition D.17(i): they change the regularity of the objective and the ambient dynamics, not the selection.

Theorem D.19 (Uniqueness of the canonical regulariser)**.** *Under Assumption 2.1, for every* $n \ge 1$*, the following are determined by the training problem* $(g, \boldsymbol{\theta}_0, Y)$ *and the conditioning ball of Assumption 2.1, with no further choice (on the dependence on the ball, see the discussion after Theorem D.3): the carrier* $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$*, the unique closed connected invariant* n*-dimensional surface*

in $\mathfrak{B}$ meeting $\Theta^\star$ only at $\boldsymbol{\theta}^\star$ (Theorem D.3(iv)); the measurement rule $\|\boldsymbol{v}\|^2$, the unique natural rule calibrated to the rate of loss decrease (Lemma D.6); the admissible class, the paths in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, which are the runs of the forces that preserve the training outcome (Lemma E.2); and the flow term $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \cdot)$, the least energy over that class, equivalently the greatest penalty vanishing at $\boldsymbol{\theta}_0$ whose square root never grows faster than the length of any outcome-preserving motion, equivalently the square of the maximal eikonal subsolution (Theorem D.16), once the regulariser is required to charge the least energy an outcome-preserving run must spend, the one principle not derived from training. Given the additive convention with the anchor $g|^{-1}_{\boldsymbol{\theta}_0}(\cdot)$ and the same rule on vertical paths, the extension is $d^2(\cdot, \boldsymbol{\theta}_0)$ of Definition 4.1 (Proposition D.17(ii)), and no scale survives except the regularisation weight λ. In the kernel regime this is anchored ridge (Corollary D.8); for $n = 1$ its flow term on the training curve is arc ridge, $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}(t)) = s(t)^2$.

Proof of Theorem D.19. The carrier is Theorem D.3(iv), the rule Lemma D.6, the class Lemma E.2, the flow term Theorem D.16, the extension Proposition D.17(ii); calibration fixes the only scale, and the kernel case is Corollary D.8. For $n = 1$, $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is an embedded arc (Corollary D.3) along which g is strictly monotone on the trajectory, so the trajectory is injective and its length $s(t)$ is the intrinsic distance of the arc (Proposition E.2). □

Three things are not claimed: uniqueness among all functions on Θ with the $\lambda \downarrow 0$ property, which is impossible (Proposition D.18); uniqueness of solutions of the eikonal equation on the incomplete $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, of which only maximality holds (Theorem D.16(U3)); and a derivation of the additive convention, which Proposition D.17(i) shows to be without effect on the estimator.

D.9 Proof of Theorem 4.2 (canonical regulariser characterisation)

Proof of Theorem 4.2. **(a)** On $\Theta(\boldsymbol{c}) \cap V$ the flow term of Definition 4.1 is constant and the fibre term is at least the squared chord to $g|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c})$ (Lemma C.6), zero only there.

(b) Whenever the trajectory lies in V (in particular for all large t), $\boldsymbol{\theta}(t) = g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta}(t)))$ (Corollary D.4), the fibre term vanishes, and $d^2_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}(t)) \le s(t)^2$, the trajectory being a path in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ of length $s(t)$.

(c) This is Proposition D.18 with $\Psi = E_Y$, continuous by the proof of Proposition D.17(i), and $b = 1$: every selection of (ϵ_λ-)minimisers of $\mathcal{L}(g(\cdot), Y) + \lambda d^2(\cdot, \boldsymbol{\theta}_0)$ converges to $\boldsymbol{\theta}^\star$.

(d) $\nabla_{\boldsymbol{\theta}}\mathcal{L} \in T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.3(ii)), so it is the intrinsic gradient of $\mathcal{L}|_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)}$; a critical point therefore has $\nabla_{\boldsymbol{\theta}}\mathcal{L} = 0$, hence lies in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap \Theta^\star = \{\boldsymbol{\theta}^\star\}$ (Lemma C.3, Theorem D.3(iii)), a strict minimum since $\mathcal{L} \ge 0$. The drift clause is Theorem D.12 with Corollary D.7: geodesic ridge restricted to $\Theta^\star \cap V$ is $E_Y(Y) + d^2_{\text{fibre}}(\boldsymbol{\theta}^\star, \cdot)$, stationary with Hessian $2\,\mathrm{Id}$ at $\boldsymbol{\theta}^\star$, and enters the theorem through its localisations to a neighbourhood on which it is C^2, which exists whenever E_Y is C^2 near Y (Corollary D.7, Propositions D.17(ii) and E.5). □

E Arc Ridge

Throughout, $\gamma_{\text{gf}}(t) := \phi_{\text{gf}}(\boldsymbol{\theta}_0, t)$ is the gradient-flow trajectory, $\Gamma_{\text{gf}} := \gamma_{\text{gf}}([0, \infty))$ its image, the *training curve*, $s_{\text{gf}}(t) := \int_0^t \|\dot\gamma_{\text{gf}}(u)\|\, du$ its arc length, and $s_\infty := \lim_{t\to\infty} s_{\text{gf}}(t) \le \ell := C\|g(\boldsymbol{\theta}_0) - Y\|/c^2 < R$ (Proposition 2.1). The loss is the squared error of A.3, $\mathcal{L} = \|g(\boldsymbol{\theta}) - Y\|^2$, so that $\nabla_{\boldsymbol{\theta}}\mathcal{L} = 2J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y)$ and $2c\|g(\boldsymbol{\theta}) - Y\| \le \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| \le 2C\|g(\boldsymbol{\theta}) - Y\|$ on $B(\boldsymbol{\theta}_0, R)$ (Lemma C.1, Assumption 2.1). We work in effective-parameter space, the stored parameters multiplied by their μP width scalings as in the implementation of Appendix H, and suppress the subscript *eff* for readability.

E.1 THE UPPER BOUND

Proposition E.1 (Upper bound via the length infimum, expanded)**.** *Under Assumption 2.1, let $\gamma : [0, T] \to \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ be any piecewise-C^1 path in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$. Then*

$$d^2_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \;\le\; L^2(\gamma) \;=\; \Big(\int_0^T \|\dot\gamma(t)\|\, dt\Big)^2.$$

Equality holds if and only if γ is a minimising geodesic of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, traversed at any speed.

Proof. The bound holds because $d_{\text{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is the infimum of lengths of such paths (Definition 3.1); equality means γ realises it, and length-minimising paths are, up to parametrisation, the minimising geodesics (do Carmo, 1992, Ch. 3). *Discrete steps.* Under gradient descent with step η, the iterates $\boldsymbol{\theta}_K$ and polygonal length s_K, $K\eta = T$, are Euler approximations to $\gamma_{\text{gf}}(T)$ and $s_{\text{gf}}(T)$ with $O(\eta)$ errors; geodesic ridge is continuous near $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$, its fibre term being $O(\|\boldsymbol{\theta} - g|^{-1}_{\boldsymbol{\theta}_0}(g(\boldsymbol{\theta}))\|^2)$ (Proposition D.10, Theorem D.4), so whenever $\gamma_{\text{gf}}(T) \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$ and η is small, $d^2(\boldsymbol{\theta}_K, \boldsymbol{\theta}_0) \le s_K^2 + O(\eta)$. □

E.2 FIRST VARIATION OF LENGTH AND THE TRAINING CURVE

Arc length is a functional of a path, not a function of its endpoint; the lemma records what the endpoint controls.

Lemma E.1 (First variation of length; endpoint term)**.** *Let $\gamma : [0, T] \to \mathbb{R}^m$ be piecewise-C^2 with corners at $t_1 < \cdots < t_k$ and $\dot\gamma \neq 0$, with unit tangent $\boldsymbol{t} := \dot\gamma / \|\dot\gamma\|$ and curvature vector $\kappa := \dot{\boldsymbol{t}} / \|\dot\gamma\|$. Let $\Gamma : (-\varepsilon_0, \varepsilon_0) \times [0, T] \to \mathbb{R}^m$ be continuous, C^2 on each $(-\varepsilon_0, \varepsilon_0) \times [t_i, t_{i+1}]$, with $\Gamma(0, \cdot) = \gamma$ and variation field $V := \partial_\varepsilon \Gamma|_{\varepsilon=0}$. Then*

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} L\big(\Gamma(\varepsilon, \cdot)\big) \;= \langle \boldsymbol{t}(T^-), V(T)\rangle - \langle \boldsymbol{t}(0^+), V(0)\rangle - \int_0^T \langle \kappa, V\rangle \, \|\dot\gamma\|\, dt + \sum_{i=1}^{k} \langle \boldsymbol{t}(t_i^-) - \boldsymbol{t}(t_i^+), V(t_i)\rangle.$$

If γ and Γ take values in a C^2 submanifold $M \subseteq \mathbb{R}^m$, the integrand may be replaced by $\langle \kappa_g, V\rangle$, where the geodesic curvature vector κ_g is the orthogonal projection of κ onto TM; on a C^2 piece, $\kappa_g \equiv 0$ if and only if that piece is a geodesic of M. In particular:

- *(i) if γ is a minimising geodesic of M from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$, then $\operatorname{grad}_M d_M(\boldsymbol{\theta}_0, \cdot)(\boldsymbol{\theta}) = \boldsymbol{t}(T)$ whenever $d_M(\boldsymbol{\theta}_0, \cdot)$ is differentiable at $\boldsymbol{\theta}$;*
- *(ii) for an arbitrary γ, the endpoint term $\langle \boldsymbol{t}(T^-), V(T)\rangle$ is the response of $L(\gamma)$ to variations with V supported in $[T - \delta, T]$, up to an error $O(\delta)$; and if $\kappa_g \not\equiv 0$ there is no function φ on M, nor on $\mathbb{R}^m$, with $L(\gamma') = \varphi(\gamma'(T))$ for all paths γ' in M near γ with $\gamma'(0) = \boldsymbol{\theta}_0$, so the endpoint term is not the gradient of any function of the endpoint.*

Proof. Differentiating under the integral (C^2 on each piece, mixed partials commuting) gives $\partial_\varepsilon \|\partial_t \Gamma\|\,|_{\varepsilon=0} = \langle \boldsymbol{t}, \dot V\rangle$, and integrating by parts on each piece, with $\dot{\boldsymbol{t}} = \|\dot\gamma\|\,\kappa$, gives the endpoint, corner and interior terms. In M, V is tangent, so $\langle \kappa, V\rangle = \langle \kappa_g, V\rangle$, and $\kappa_g \equiv 0$ characterises geodesics (do Carmo, 1992, Ch. 3).

(i) For $w \in T_{\boldsymbol{\theta}} M$ take the variation $\Gamma(\varepsilon, t) := \exp_{\gamma(t)}(\varepsilon V(t))$ with $V(0) = 0, V(T) = w$ (do Carmo, 1992, Ch. 9). $f(\varepsilon) := L(\Gamma(\varepsilon, \cdot)) - d_M(\boldsymbol{\theta}_0, \Gamma(\varepsilon, T)) \ge 0$ vanishes at 0, so $0 = f'(0) = \langle \boldsymbol{t}(T) - \operatorname{grad}_M d_M(\boldsymbol{\theta}_0, \cdot)(\boldsymbol{\theta}), w\rangle$.

(ii) For V supported in $[T - \delta, T]$, $\|V\| \le \|V(T)\|$, the interior term is at most $\sup\|\kappa\|\,\|V(T)\|\,L(\gamma|_{[T-\delta,T]}) = O(\delta)$. If such a φ existed, a tangent field V vanishing at 0, T and the corners with $\int_0^T \langle \kappa_g, V\rangle\,\|\dot\gamma\|\, dt \neq 0$ would give an endpoint-fixing variation in M along which L is constant yet has non-zero derivative. □

Proposition E.2 (Arc ridge is geodesic ridge of the training curve)**.** *Under A.1, A.3 and Assumption 2.1 with $\boldsymbol{\theta}_0 \notin \Theta^\star$, the map $t \mapsto \gamma_{\mathrm{gf}}(t)$ is an injective C^∞ immersion of $[0,\infty)$ and a homeomorphism onto its image, so the training curve Γ_{gf} is an embedded one-dimensional C^∞ submanifold of Θ with boundary point $\boldsymbol{\theta}_0$, and it lies in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.3(i)). Its arc length,*

$$s : \Gamma_{\mathrm{gf}} \to [0, s_\infty), \qquad s\big(\gamma_{\mathrm{gf}}(t)\big) := s_{\mathrm{gf}}(t),$$

is a C^∞ function on Γ_{gf} equal to the intrinsic distance $d_{\Gamma_{\mathrm{gf}}}(\boldsymbol{\theta}_0, \cdot)$, and the intrinsic gradient of its square is the arc-ridge force,

$$\mathrm{grad}_{\Gamma_{\mathrm{gf}}} s^2 \;=\; 2s\,\frac{\dot\gamma_{\mathrm{gf}}}{\|\dot\gamma_{\mathrm{gf}}\|} \;=\; \boldsymbol{F}_{\mathrm{arc}}.$$

Consequently, under Assumption 2.1, for every $\boldsymbol{\theta} \in \Gamma_{\mathrm{gf}}$,

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 = d^2_\Theta(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \;\le\; d^2_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \;\le\; d^2_{\Gamma_{\mathrm{gf}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = s(\boldsymbol{\theta})^2,$$

and $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ on $\Gamma_{\mathrm{gf}} \cap V$ (Corollary D.4): anchored ridge, geodesic ridge, and arc ridge are the squared intrinsic distances from $\boldsymbol{\theta}_0$ on $\Theta \supset \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \supset \Gamma_{\mathrm{gf}}$, and the sandwich is the monotonicity of intrinsic distance under inclusion.

Proof. $\dot\gamma_{\mathrm{gf}}(t) \neq 0$ for all t (a trajectory point in $\Theta^\star$ would be an equilibrium, forcing $\boldsymbol{\theta}_0 \in \Theta^\star$ by uniqueness), the curve is injective since the loss strictly decreases (Lemma C.2), and it is a homeomorphism onto its image since $\gamma_{\mathrm{gf}}(t) \to \boldsymbol{\theta}^\star \notin \{\gamma_{\mathrm{gf}}(t_0)\}$; it lies in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ by Theorem D.3(i). s is a unit-speed C^∞ coordinate, and any path in Γ_{gf} from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$ has length at least the total variation $s(\boldsymbol{\theta})$ of s along it; so $d_{\Gamma_{\mathrm{gf}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = s(\boldsymbol{\theta})$ and $\mathrm{grad}_{\Gamma_{\mathrm{gf}}} s^2 = 2s\,\boldsymbol{t} = \boldsymbol{F}_{\mathrm{arc}}$. The chain of inequalities is the shrinking of the path families along $\Theta \supset \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \supset \Gamma_{\mathrm{gf}}$, and the last identity is Corollary D.4 with Definition 4.1. □

Why the surrogate. $T\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ at the current iterate is the space of perturbations that *will* be forgotten (Lemma D.2); at time T a run knows one vector of it, $X(\boldsymbol{\theta}(T))$, and one invariant submanifold, its own path, on which arc ridge is the value function of Theorem D.16.

E.3 Proof of Theorem 4.3 (arc ridge is an admissible upper bound)

Proof of Theorem 4.3. (a) is Proposition E.1, with $d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$, where the fibre term vanishes (Corollary D.4).

(b) $\boldsymbol{F}_{\mathrm{arc}}(T)$ is a positive multiple of $\dot\gamma(T^-) \in T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, and under gradient flow it is also horizontal (Lemma C.1, Theorem D.3(ii)). The gradient $2(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ of anchored ridge has a component along $\ker J(\boldsymbol{\theta})$ exactly when alignment fails (Lemma C.5), transverse to $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ at $\boldsymbol{\theta}^\star$.

(c) On $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ geodesic ridge is $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$, whose intrinsic gradient, where it exists, is $2\,d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ times that of the distance. The latter has norm at most 1 (Lipschitz) and at least 1: along a near-minimising path β with $L(\beta) \le d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + h^2$, the point q at chord distance h from $\boldsymbol{\theta}$ has $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, q) \le d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + h^2 - h$ (Lemma C.6) and $d_{\mathrm{flow}}(\boldsymbol{\theta}, q) = h(1 + O(h^2))$ (Lemma C.8). Hence the norm is $2\,d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \le 2L(\gamma) = \|\boldsymbol{F}_{\mathrm{arc}}(T)\|$; equality of the vectors holds iff γ is a minimising geodesic, by Lemma E.1(i) and (a). The ambient gradient is treated in Proposition E.5. □

E.4 Admissible forces

A regulariser that is a function on Θ acts through its gradient; arc ridge acts through a force, and the right notion of unbiasedness for a force is dynamical.

Definition E.1 (Admissible regularisation force and force-regularised dynamics)**.** A *regularisation force* is a map $\boldsymbol{F} : (\boldsymbol{\theta}, s) \mapsto \boldsymbol{F}(\boldsymbol{\theta}, s) \in \mathbb{R}^m$, defined and locally Lipschitz on an open subset of $B(\boldsymbol{\theta}_0, R) \times [0,\infty)$. It is *admissible* if $\boldsymbol{F}(\boldsymbol{\theta}, s) \in T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ whenever $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ and $(\boldsymbol{\theta}, s)$ lies in its domain: the force never pushes off the basin manifold. The *force-regularised dynamics* with weight $\lambda > 0$ is the autonomous system

$$\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}} \mathcal{L}\left(g(\boldsymbol{\theta}), Y\right) - \lambda\, \boldsymbol{F}(\boldsymbol{\theta}, s), \qquad \dot s = \left\|\dot{\boldsymbol{\theta}}\right\|, \qquad (\boldsymbol{\theta}, s)(0) = (\boldsymbol{\theta}_0, 0),$$

whose right-hand side is locally Lipschitz, so that its solutions are unique; s is a memory variable, the length of the path traversed, and not a coordinate on Θ. The arc-ridge force of Definition 4.2, $\boldsymbol{F}_{\mathrm{arc}}(\boldsymbol{\theta}, s) = -2s\,\nabla_{\boldsymbol{\theta}}\mathcal{L}/\,\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|$, is defined and C^∞ on $(B(\boldsymbol{\theta}_0, R) \setminus \Theta^\star) \times [0,\infty)$ (Lemma C.3) and admissible, since $\boldsymbol{F}_{\mathrm{arc}} \propto X = -\nabla_{\boldsymbol{\theta}}\mathcal{L} \in T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ at every point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.3(ii)); it is moreover horizontal, since $\nabla_{\boldsymbol{\theta}}\mathcal{L} = -J(\boldsymbol{\theta})^\top \boldsymbol{u} \in \operatorname{Im} J(\boldsymbol{\theta})^\top$ (Lemma C.1). The intrinsic gradient of geodesic ridge is admissible on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$; the gradients of anchored ridge and of the ambient extension of geodesic ridge are not in general (Proposition E.5). A horizontal force need not be admissible: away from $\boldsymbol{\theta}^\star$, $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \neq \operatorname{Im} J(\boldsymbol{\theta})^\top$ in general (Theorem D.5), and a horizontal force moves the training limit at the rate $-\lambda\, d\pi(\boldsymbol{\theta})\boldsymbol{F}$, which vanishes on $\Theta^\star$ and is of the order of the residual near it (Proposition E.5(iii)). When $\mathcal{H}$ is bracket-generating, a formal Stokes-type expansion of that rate around a small horizontal loop near $\Theta^\star$ suggests a net displacement of the limit of the order of the enclosed area; we do not prove this.

The admissible class is not a choice:

Lemma E.2 (Admissible forces are exactly the outcome-preserving ones, and their runs are the paths in the basin manifold)**.** *Under A.1–A.3 with squared-error loss and Assumption 2.1, let $\boldsymbol{F}$ be a regularisation force whose domain contains $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \times [0,\infty)$.*

- *(i)* Outcome preservation is tangency. *If $\boldsymbol{F}$ is admissible, then for every $\lambda > 0$ and every $(\boldsymbol{\theta}', s') \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \times [0,\infty)$ the force-regularised trajectory from $(\boldsymbol{\theta}', s')$ remains in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ for as long as it exists in $\mathfrak{B}$; in particular the trajectory from $(\boldsymbol{\theta}_0, 0)$ does. Conversely, if for some $\lambda > 0$ the trajectory from every $(\boldsymbol{\theta}', s') \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \times [0,\infty)$ remains in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ on some initial interval, then $\boldsymbol{F}$ is admissible; and a trajectory in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ that reaches a point $(\boldsymbol{\theta}(t_1), s(t_1))$ with $\boldsymbol{F}(\boldsymbol{\theta}(t_1), s(t_1)) \notin T_{\boldsymbol{\theta}(t_1)}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ leaves $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ immediately after t_1.*
- *(ii)* The runs are the paths. *Every force-regularised trajectory of an admissible force from $(\boldsymbol{\theta}_0, 0)$ is a C^1 path in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$. Conversely, let $\gamma : [0, \sigma_1] \to \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ be a C^2 path from $\boldsymbol{\theta}_0$ with $\dot\gamma \neq 0$, parametrised by arc length. For every $\lambda > 0$ there is an admissible force whose trajectory from $(\boldsymbol{\theta}_0, 0)$ is $\boldsymbol{\theta}(t) = \gamma(t)$, $s(t) = t$, for $0 \le t \le \sigma_1$. Every piecewise-C^1 path in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$ is the uniform limit of such C^2 paths with the same endpoints and lengths converging to its length.*

Proof. (i) The field $\Phi(\boldsymbol{\theta}, s) := \big(-\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda\boldsymbol{F},\ \|\nabla_{\boldsymbol{\theta}}\mathcal{L} + \lambda\boldsymbol{F}\|\big)$ is locally Lipschitz and, by Theorem D.3(ii) and admissibility, tangent to the embedded submanifold $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \times [0,\infty)$; its maximal integral curve there is, by uniqueness, the ambient trajectory, which therefore stays in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ as long as the restricted curve exists. The restricted curve cannot stop earlier: $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is closed in $\mathfrak{B}$ (Theorem D.3(i)), so the escape lemma (Lee, 2012, Ch. 9) would be violated. Conversely, if $\boldsymbol{F}(\boldsymbol{\theta}', s') \notin T_{\boldsymbol{\theta}'}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, the initial velocity has a non-zero normal component, while a sequence $\boldsymbol{\theta}(t_k) \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, $t_k \downarrow 0$, would force it tangent in a slice chart; so the trajectory leaves at once, and the same argument applies at t_1.

(ii) Admissible trajectories are C^1 and lie in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ by (i). Conversely, with $\dot\gamma$ extended constantly beyond σ_1, ξ the nearest-point retraction of a tubular neighbourhood U of $\gamma([0,\sigma_1])$ in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (Lee, 2012, Ch. 6) and $\Pi_{\boldsymbol{\theta}}$ the orthogonal projection onto $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, the force

$$\boldsymbol{F}(\boldsymbol{\theta}, s) := -\frac{1}{\lambda}\Big(\Pi_{\xi(\boldsymbol{\theta})}\,\dot\gamma(s) + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\Big)$$

is locally Lipschitz and admissible, and $(\gamma(t), t)$ solves the regularised system, hence is its trajectory. Piecewise-C^1 paths are approximated, uniformly and in length, by C^2 ones with the same endpoints (inscribed polygons in charts with rounded corners). □

E.5 Formal equivalence to early stopping under gradient flow

Proposition E.3 (Arc ridge is early stopping under gradient flow)**.** *Under A.1–A.3 with squared-error loss and Assumption 2.1, suppose $\boldsymbol{\theta}_0 \notin \Theta^\star$ and set*

$$\psi_\lambda(t) := 2\lambda\, s_{\mathrm{gf}}(t) - \big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\big(\gamma_{\mathrm{gf}}(t)\big)\big\|, \qquad t \ge 0.$$

Then $\psi_\lambda(0) < 0$ *and* $\psi_\lambda(t) \to 2\lambda s_\infty > 0$ *as* $t \to \infty$*, so the* stopping time $T^\star_\lambda := \min\{t \geq 0 : \psi_\lambda(t) = 0\}$ *exists and lies in* $(0, \infty)$*. The arc-ridge dynamics of Definition E.1,* $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda \boldsymbol{F}_{\mathrm{arc}}(\boldsymbol{\theta}, s)$*,* $\dot{s} = \left\|\dot{\boldsymbol{\theta}}\right\|$*, from* $(\boldsymbol{\theta}_0, 0)$*, has a unique solution defined for all* $t \geq 0$*, and:*

- *(i)* Reparametrisation. $\boldsymbol{\theta}(t) = \gamma_{\mathrm{gf}}(\tilde{t}(t))$ *and* $s(t) = s_{\mathrm{gf}}(\tilde{t}(t))$*, where* $\tilde{t}(0) = 0$ *and* $\dot{\tilde{t}} = -\psi_\lambda(\tilde{t}) / \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\gamma_{\mathrm{gf}}(\tilde{t}))\right\| \in (0, 1]$ *while* $\tilde{t} < T^\star_\lambda$*.*
- *(ii)* Same path, truncated. *The image of the regularised trajectory is* $\gamma_{\mathrm{gf}}([0, T^\star_\lambda)) \subset \Gamma_{\mathrm{gf}} \subseteq \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$*.*
- *(iii)* Stopping. $\tilde{t}(t) \uparrow T^\star_\lambda$ *and* $\boldsymbol{\theta}(t) \to \gamma_{\mathrm{gf}}(T^\star_\lambda)$ *as* $t \to \infty$*; the limit satisfies the stopping condition*
$$2\lambda\, s_{\mathrm{gf}}(T^\star_\lambda) \;=\; \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\big(\gamma_{\mathrm{gf}}(T^\star_\lambda)\big)\right\|, \tag{1}$$
and $(\gamma_{\mathrm{gf}}(T^\star_\lambda), s_{\mathrm{gf}}(T^\star_\lambda))$ *is an equilibrium of the* $(\boldsymbol{\theta}, s)$ *system, approached asymptotically and reached at no finite time.*
- *(iv)* Non-interpolation. $\gamma_{\mathrm{gf}}(T^\star_\lambda) \notin \Theta^\star$*, and* $\|g(\gamma_{\mathrm{gf}}(T^\star_\lambda)) - Y\| \leq \lambda s_{\mathrm{gf}}(T^\star_\lambda)/c \leq \lambda\ell/c$*.*
- *(v)* Consistency. $T^\star_\lambda \to \infty$ *and* $\gamma_{\mathrm{gf}}(T^\star_\lambda) \to \boldsymbol{\theta}^\star$ *as* $\lambda \downarrow 0$*.*

Write σ *for the arc-length coordinate along* Γ_{gf} *and* $\sigma^\star_\lambda := s_{\mathrm{gf}}(T^\star_\lambda)$*. If* $t \mapsto \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\gamma_{\mathrm{gf}}(t))\|$ *is non-increasing near* $T^\star_\lambda$*, as it is when* $\nabla^2_{\boldsymbol{\theta}}\mathcal{L} \succeq 0$ *along the arc, the equilibrium is exponentially stable:* $\sigma^\star_\lambda - s(t)$ *decays at the rate* $2\lambda - \frac{d}{d\sigma}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|\big|_{\sigma=\sigma^\star_\lambda} > 0$*. If it is non-increasing on all of* $[0, \infty)$*, then* ψ_λ *is strictly increasing and* $T^\star_\lambda$ *is its unique zero.*

Proof. $\psi_\lambda(0) < 0$ (Lemma C.3), ψ_λ is C^∞ since $\gamma_{\mathrm{gf}}(t) \notin \Theta^\star$ (Proposition E.2), and $\psi_\lambda(t) \to 2\lambda s_\infty > 0$ (Proposition 2.1); its zero set is closed, so $T^\star_\lambda$ exists.

(i) With $\varphi := -\psi_\lambda / \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\gamma_{\mathrm{gf}})\| = 1 - 2\lambda s_{\mathrm{gf}} / \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\gamma_{\mathrm{gf}})\| \in (0, 1]$ on $[0, T^\star_\lambda)$ and $\dot{\tilde{t}} = \varphi(\tilde{t})$, the pair $(\gamma_{\mathrm{gf}}(\tilde{t}), s_{\mathrm{gf}}(\tilde{t}))$ satisfies $\dot{\boldsymbol{\theta}} = -(1 - 2\lambda s / \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|)\nabla_{\boldsymbol{\theta}}\mathcal{L} = -\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda \boldsymbol{F}_{\mathrm{arc}}$ and $\dot{s} = \left\|\dot{\boldsymbol{\theta}}\right\|$, so by uniqueness it is the solution.

(iii) $\tilde{t} \equiv T^\star_\lambda$ solves the scalar equation, so $\tilde{t}(t)$ increases to a limit $\tilde{t}_\infty \leq T^\star_\lambda$ with $\varphi(\tilde{t}_\infty) = 0$, hence $\tilde{t}_\infty = T^\star_\lambda$; the limit state has zero velocity. (ii) follows.

(iv) At the stopping point $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| = 2\lambda s_{\mathrm{gf}}(T^\star_\lambda) > 0$ and $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| \geq 2c\,\|g(\boldsymbol{\theta}) - Y\|$.

(v) By (iv) the residual tends to 0; as the loss strictly decreases along γ_{gf} and is positive at finite times, $T^\star_\lambda \to \infty$.

Stability and uniqueness. In arc length, $\dot{s} = \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| - 2\lambda s = -\tilde{\psi}_\lambda(s)$, whose linearisation at $\sigma^\star_\lambda$ has rate $\tilde{\psi}'_\lambda(\sigma^\star_\lambda)$ as stated, at least 2λ when the gradient norm is non-increasing, as it is when $\nabla^2_{\boldsymbol{\theta}}\mathcal{L} \succeq 0$ since $\frac{d}{dt}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|^2 = -2\nabla_{\boldsymbol{\theta}}\mathcal{L}^\top \nabla^2_{\boldsymbol{\theta}}\mathcal{L}\nabla_{\boldsymbol{\theta}}\mathcal{L}$; if it is non-increasing throughout, ψ_λ is strictly increasing. □

Readings of the stopping condition. Condition (1) balances the prior force $2\lambda s$ against the data force $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|$. It is also MAP estimation on the training curve: in the arc-length coordinate σ, $\phi(\sigma) := \mathcal{L}(g(\boldsymbol{\theta}(\sigma)), Y) + \lambda\sigma^2$ has $\phi' = -\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| + 2\lambda\sigma = \tilde{\psi}_\lambda$, and the arc-ridge dynamics is the gradient flow of ϕ on Γ_{gf}, stopping at its first critical point (Proposition E.8). Unlike classical early stopping it needs no validation set, only the accumulated length, the current gradient norm and λ. Under a Gaussian likelihood with noise variance σ^2 and the prior $p(\boldsymbol{c}) \propto \exp(-\beta E_Y(\boldsymbol{c}))$ of Remark 4.1, the MAP objective is $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda E_Y(g(\boldsymbol{\theta}))$ with $\lambda = 2\beta\sigma^2$ (no factor of n, $\mathcal{L}$ being a sum), so training stops where $s_{\mathrm{gf}}(T^\star_\lambda) = \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\gamma_{\mathrm{gf}}(T^\star_\lambda))\| / (4\beta\sigma^2)$.

E.6 No fibre bias for admissible forces

Theorem E.4 (No fibre bias for admissible forces)**.** *Under A.1–A.3 with squared-error loss and Assumption 2.1, let* $\boldsymbol{F}$ *be admissible and* $B : [0, \infty) \to [0, \infty)$ *non-decreasing, and suppose the force-regularised trajectory exists for all* $t \geq 0$*, remains in* $\mathfrak{B}$*, has* $s(t) \uparrow s_\lambda < \infty$*, and converges,* $\boldsymbol{\theta}(t) \to \omega_\lambda$*, with* $(\omega_\lambda, s_\lambda)$ *in the domain of* $\boldsymbol{F}$*,* $\|\boldsymbol{F}(\omega_\lambda, s_\lambda)\| \leq B(s_\lambda)$*, and* ω_λ *in the neighbourhood* V *of Theorem D.4, taken inside* $\mathfrak{B}$ *and so small that* $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ *is connected. Then:*

(a) ***Basin invariance.*** *$\boldsymbol{\theta}(t) \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ for all $t \geq 0$.*
(b) ***Residual bound.*** *$\|g(\omega_\lambda) - Y\| \leq \lambda B(s_\lambda)/(2c)$.*
(c) ***Consistency.*** *$\omega_\lambda = g|_{\boldsymbol{\theta}_0}^{-1}(g(\omega_\lambda))$, and $\omega_\lambda \to \boldsymbol{\theta}^\star$ as $\lambda B(s_\lambda) \to 0$, in particular as $\lambda \downarrow 0$ whenever $\sup_\lambda B(s_\lambda) < \infty$.*

For $\boldsymbol{F} = \boldsymbol{F}_{\mathrm{arc}}$ the hypotheses hold for every sufficiently small λ, with $B(s) = 2s \leq 2\ell$; the trajectory lies in Γ_{gf}; and $\omega_\lambda = \gamma_{\mathrm{gf}}(T_\lambda^\star)$ with $\|g(\omega_\lambda) - Y\| \leq \lambda s_{\mathrm{gf}}(T_\lambda^\star)/c \leq \lambda\ell/c$ and $\omega_\lambda \to \boldsymbol{\theta}^\star$ as $\lambda \downarrow 0$, where $T_\lambda^\star$ is the stopping time of Proposition E.3.

Proof. (a) is Lemma E.2(i). *(b)* Convergence of $\boldsymbol{\theta}(t)$ and continuity of the right-hand side force the limiting velocity to vanish, so $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\omega_\lambda)\| = \lambda\|\boldsymbol{F}(\omega_\lambda, s_\lambda)\| \leq \lambda B(s_\lambda)$, and $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\| \geq 2c\|g(\boldsymbol{\theta}) - Y\|$. *(c)* $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is closed in $\mathfrak{B}$, so $\omega_\lambda \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ and $\omega_\lambda = g|_{\boldsymbol{\theta}_0}^{-1}(g(\omega_\lambda))$ (Corollary D.4); continuity of $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ and (b) give the limit. *Arc ridge.* $\|\boldsymbol{F}_{\mathrm{arc}}\| = 2s$, and Proposition E.3 supplies the remaining hypotheses (the limit lies in V for small λ since $T_\lambda^\star \to \infty$) and the stated bounds; the trajectory lies in $\Gamma_{\mathrm{gf}} \subseteq \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. □

A force has no objective, so unbiasedness means consistency of the limit of the regularised dynamics. For anchored ridge the same equilibrium argument gives $\|g(\omega_\lambda) - Y\| \to 0$, but ω_λ leaves $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, so (c) is unavailable, and any limit is a critical point of the chord on $\Theta^\star$, different from $\boldsymbol{\theta}^\star$ whenever alignment fails. The intrinsic gradient of geodesic ridge satisfies (a)–(c) but leaves the training curve; its ambient gradient generates a different, non-admissible dynamics (Proposition E.5).

Proposition E.5 (Intrinsic and ambient regularised dynamics)**.** *Under A.1–A.3 with squared-error loss and Assumption 2.1, let $R := d^2(\cdot, \boldsymbol{\theta}_0)$ be geodesic ridge (Definition 4.1), defined on the set $\mathcal{W} \subseteq V$ of Proposition D.17; write $R_{\mathrm{flow}} := d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \cdot)$ for its flow term, a function on all of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ with $R = R_{\mathrm{flow}}$ on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ (Corollary D.4), and let $M_1 \subseteq \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ be the open set on which R_{flow} is C^2, which contains a neighbourhood of $\boldsymbol{\theta}_0$ in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. Suppose that E_Y is C^2 on a neighbourhood of Y; this holds, for instance, when every point of a neighbourhood of $\boldsymbol{\theta}^\star$ in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is joined to $\boldsymbol{\theta}_0$ by a unique minimising geodesic of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, depending continuously on the endpoint, along which $\boldsymbol{\theta}^\star$ is not conjugate to $\boldsymbol{\theta}_0$ (Remark D.6), and in the kernel regime, where E_Y is quadratic (Proposition D.15). Let $V_1 \subseteq V \cap \mathfrak{B}$ be an open neighbourhood of $\boldsymbol{\theta}^\star$ with compact closure in $\mathfrak{B}$ on which R is C^2 (Proposition D.17(ii)), so that $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V_1 \subseteq M_1$, let $\Pi_{\boldsymbol{\theta}}$ be the orthogonal projection onto $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, and let $N(\boldsymbol{c}) = (I - P_{g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})})Dg|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ and $G = K^{-1} + N^\top N$ be the tilt and the pullback metric of Theorem D.5.*

(i) ***Two gradients.*** *At $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V_1$ with $\boldsymbol{c} := g(\boldsymbol{\theta})$, the ambient gradient is $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = J(\boldsymbol{\theta})^\top \nabla E_Y(\boldsymbol{c}) \in \operatorname{Im} J(\boldsymbol{\theta})^\top$, the unique horizontal vector whose pairings with $T_{\boldsymbol{\theta}}\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ are those of the intrinsic gradient; the intrinsic gradient is $\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)} R(\boldsymbol{\theta}) = \Pi_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$; and*

$$\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)} R(\boldsymbol{\theta})\right\|^2 = \nabla E_Y(\boldsymbol{c})^\top G(\boldsymbol{c})^{-1} \nabla E_Y(\boldsymbol{c}) \leq \nabla E_Y(\boldsymbol{c})^\top K(\boldsymbol{\theta}) \nabla E_Y(\boldsymbol{c}) = \|\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})\|^2,$$

with equality if and only if $\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)} R(\boldsymbol{\theta})$ is horizontal, in which case the two gradients coincide. They coincide at $\boldsymbol{\theta}^\star$, and at every point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V_1$ in the integrable class (Theorem D.9).

(ii) ***Intrinsic dynamics.*** *On M_1 the* intrinsic *dynamics $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda \operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)} R_{\mathrm{flow}}(\boldsymbol{\theta})$ is the gradient flow of $(\mathcal{L} + \lambda R_{\mathrm{flow}})|_{M_1}$ for the induced metric. Its force is admissible, so its trajectories from $\boldsymbol{\theta}_0$ are paths in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (Lemma E.2), along which*

$$\left\|\operatorname{grad}_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)} R_{\mathrm{flow}}(\boldsymbol{\theta}(t))\right\| = 2\, d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}(t)) \leq 2s(t) = \|\boldsymbol{F}_{\mathrm{arc}}(\boldsymbol{\theta}(t), s(t))\|,$$

with equality if and only if the path traversed so far is a minimising geodesic of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$; and Theorem E.4 applies to any of its trajectories that satisfies that theorem's remaining hypotheses, with $B(s) = 2s$.

(iii) ***Ambient dynamics.*** *The* ambient *dynamics $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda \nabla_{\boldsymbol{\theta}} R$ on V_1, what gradient descent on $\mathcal{L} + \lambda R$ does, has horizontal velocity at every point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V_1$, and its force is admissible at such a point if and only if the equality case of (i) holds*

there; wherever $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \neq \operatorname{Im} J(\boldsymbol{\theta})^\top$ the admissible horizontal vectors form a proper subspace of $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, and an ambient trajectory that reaches a point of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V_1$ at which its force is not admissible leaves $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ immediately. Its training limit moves at the rate $\frac{d}{dt}\pi(\boldsymbol{\theta}(t)) = -\lambda\, d\pi(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$, which at points of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V_1$ is $O\big(\lambda \|g(\boldsymbol{\theta}) - Y\| \|\nabla E_Y(g(\boldsymbol{\theta}))\|\big)$; the training limit moves at a rate $O(\lambda \|g(\boldsymbol{\theta}) - Y\| \|\boldsymbol{F}\|)$ under every horizontal force $\boldsymbol{F}$. For λ small, $\mathcal{L} + \lambda R$ has a unique critical point $\boldsymbol{\theta}_\lambda$ in a neighbourhood of $\boldsymbol{\theta}^\star$ that does not depend on λ, with $\boldsymbol{\theta}_\lambda = \boldsymbol{\theta}^\star + O(\lambda)$, and $\boldsymbol{\theta}_\lambda$ is a strict local minimiser; the slow flow of Theorem D.12 is stationary at $\boldsymbol{\theta}^\star$ (Corollary D.7). Finally, if for every small λ the ambient trajectory from a point $\boldsymbol{\theta}_{\text{in}} \in \mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V_1$ (from $\boldsymbol{\theta}_0$ itself when $\boldsymbol{\theta}_0 \in V_1$) remains for all $t \geq 0$ in a fixed compact subset $\mathcal{K}$ of $V_1 \cap \mathcal{W}$ containing $\boldsymbol{\theta}^\star$ and converges to ω_λ, then $\omega_\lambda \to \boldsymbol{\theta}^\star$ as $\lambda \downarrow 0$.

Proof. Regularity. $E_Y \circ g$ is C^2 where $g(\cdot)$ is near Y, and $S := d^2_{\text{fibre}}\big(g|^{-1}_{\boldsymbol{\theta}_0}(g(\cdot)), \cdot\big)$ is C^2 near $\boldsymbol{\theta}^\star$ (Proposition D.17(ii)), so V_1 exists. Near $\boldsymbol{\theta}_0$, $R_{\text{flow}} = \left\|\exp^{-1}_{\boldsymbol{\theta}_0}(\cdot)\right\|^2$ is C^∞ (do Carmo, 1992, Ch. 3). Under the sufficient condition, non-conjugacy makes $\exp_{\boldsymbol{\theta}_0}$ a local diffeomorphism near the initial velocity $\boldsymbol{v}(\boldsymbol{\theta}^\star)$ of the minimising geodesic, continuity identifies $q \mapsto \boldsymbol{v}(q)$ with its local inverse, and $R_{\text{flow}} = \|\boldsymbol{v}\|^2$ and $E_Y = R_{\text{flow}} \circ g|^{-1}_{\boldsymbol{\theta}_0}(\cdot)$ are C^∞; in the kernel regime E_Y is quadratic (Proposition D.15(c)).

(i) $S \geq 0$ is C^2 and vanishes on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V_1$, so $\nabla_{\boldsymbol{\theta}} R = J(\boldsymbol{\theta})^\top \nabla E_Y(g(\boldsymbol{\theta}))$ there, and $\operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} R = \Pi_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} R$. A horizontal vector orthogonal to $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is orthogonal to $T_{\boldsymbol{\theta}}\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \oplus \ker J(\boldsymbol{\theta}) = \mathbb{R}^m$ (Theorem D.4(ii), Lemma C.5), which gives uniqueness. In the chart, the intrinsic gradient is $Dg|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c})\, \boldsymbol{x}$ with $\boldsymbol{x} := G^{-1}\nabla E_Y$, so $\left\|\operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} R\right\|^2 = \boldsymbol{x}^\top G \boldsymbol{x} = \boldsymbol{x}^\top K^{-1}\boldsymbol{x} + \|N\boldsymbol{x}\|^2$, while $\nabla E_Y = K^{-1}\boldsymbol{x} + N^\top N \boldsymbol{x}$ gives $\|\nabla_{\boldsymbol{\theta}} R\|^2 = \nabla E_Y^\top K \nabla E_Y = \boldsymbol{x}^\top K^{-1}\boldsymbol{x} + 2\|N\boldsymbol{x}\|^2 + \left\|K^{1/2} N^\top N \boldsymbol{x}\right\|^2$. The difference vanishes iff $N\boldsymbol{x} = (I - P_{\boldsymbol{\theta}}) \operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} R = 0$, and then the two horizontal vectors with equal pairings coincide. At $\boldsymbol{\theta}^\star$ and in the integrable class every tangent vector is horizontal.

(ii) As $\nabla_{\boldsymbol{\theta}} \mathcal{L} \in T\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$, the intrinsic dynamics is the gradient flow of the restricted objective. With ξ the nearest-point retraction of a tubular neighbourhood of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (Lee, 2012, Ch. 6), $\boldsymbol{F} := \nabla_{\boldsymbol{\theta}}(R_{\text{flow}} \circ \xi)$ is C^1 on $\xi^{-1}(M_1)$ and equals $\operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} R_{\text{flow}}$ on M_1, so it is an admissible force (Lemma E.2(i)). By the proof of Theorem 4.3(c), $\left\|\operatorname{grad}_{\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)} R_{\text{flow}}\right\| = 2\, d_{\text{flow}}(\boldsymbol{\theta}_0, \cdot) \leq 2s(t)$ along a trajectory, with equality iff the path so far is a minimising geodesic (Proposition E.1); at the limit state this gives the bound of Theorem E.4 with $B(s) = 2s$.

(iii) Both $\nabla_{\boldsymbol{\theta}} \mathcal{L}$ and $\nabla_{\boldsymbol{\theta}} R$ are horizontal on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V_1$; admissibility of $\nabla_{\boldsymbol{\theta}} R$ is $\Pi_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} R = \nabla_{\boldsymbol{\theta}} R$, the equality case of (i); two distinct n-dimensional subspaces meet in a proper subspace; departure is Lemma E.2(i). *Training limit.* $\frac{d}{dt}\pi(\boldsymbol{\theta}) = -\lambda\, d\pi(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} R$ (Theorem D.2), and for horizontal $\boldsymbol{v} = J(\boldsymbol{\theta})^\top \boldsymbol{a}$ and $p := \pi(\boldsymbol{\theta})$, with $\|\boldsymbol{\theta} - p\| \leq C\|g(\boldsymbol{\theta}) - Y\|/c^2$ and $d\pi(p) J(p)^\top = (I - P_p) J(p)^\top = 0$, Lipschitz continuity of $d\pi$ and J gives $d\pi(\boldsymbol{\theta})\boldsymbol{v} = O(\|\boldsymbol{\theta} - p\| \|\boldsymbol{a}\|) = O(\|g(\boldsymbol{\theta}) - Y\| \|\boldsymbol{v}\|)$.

Critical point. Coordinates $(p, \boldsymbol{w}) \mapsto p + \boldsymbol{w}$, $p \in \Theta^\star \cap V_1$, $\boldsymbol{w} \in \operatorname{Im} J(p)^\top$, cover a neighbourhood of $\boldsymbol{\theta}^\star$. The horizontal component $P_p \nabla \Phi_\lambda(p + \boldsymbol{w})$, $\Phi_\lambda := \mathcal{L} + \lambda R$, has $\boldsymbol{w}$-derivative $2J(p)^\top J(p)$ at $(p, \mathbf{0}, 0)$, an automorphism of $\operatorname{Im} J(p)^\top$, so the implicit function theorem gives its zero set as $\boldsymbol{w} = \boldsymbol{w}(p, \lambda) = O(\lambda)$, C^1. The vertical component at that point is $\lambda h(p, \lambda)$ with

$$h(p, \lambda) := \frac{F(p, \boldsymbol{w}(p, \lambda))}{\lambda} + (I - P_p)\nabla_{\boldsymbol{\theta}} R\big(p + \boldsymbol{w}(p, \lambda)\big), \qquad F(p, \boldsymbol{w}) := (I - P_p)\nabla_{\boldsymbol{\theta}} \mathcal{L}(p + \boldsymbol{w}),$$

C^1 near $(\boldsymbol{\theta}^\star, 0)$ because F vanishes to second order in $\boldsymbol{w}$ ($D_{\boldsymbol{w}} F(p, \mathbf{0}) = 2(I - P_p) J(p)^\top J(p) = 0$). $h(\cdot, 0)$ is the intrinsic gradient of $R|_{\Theta^\star \cap V_1} = E_Y(Y) + d^2_{\text{fibre}}(\boldsymbol{\theta}^\star, \cdot)$, zero at $\boldsymbol{\theta}^\star$ with derivative $2\,\mathrm{Id}$ (do Carmo, 1992, Ch. 3); the implicit function theorem gives a unique zero $p_\lambda = \boldsymbol{\theta}^\star + O(\lambda)$, hence a unique critical point $\boldsymbol{\theta}_\lambda = \boldsymbol{\theta}^\star + O(\lambda)$ near $\boldsymbol{\theta}^\star$.

Strictness. The critical-point equation gives $\boldsymbol{r} := g(\boldsymbol{\theta}_\lambda) - Y = -\frac{\lambda}{2}\nabla E_Y(Y) + O(\lambda^2)$. In blocks along $\operatorname{Im} J(\boldsymbol{\theta}^\star)^\top \oplus \ker J(\boldsymbol{\theta}^\star)$, the horizontal block of $\nabla^2 \Phi_\lambda(\boldsymbol{\theta}_\lambda)$ is at least $(2c^4/C^2 - O(\lambda))\mathrm{Id}$,

the off-diagonal block is $O(\lambda)$ since $J(\boldsymbol{\theta}_\lambda)\,\boldsymbol{z} = O(\lambda)\,\|\boldsymbol{z}\|$ on $\ker J(\boldsymbol{\theta}^\star)$, and the vertical block is

$$\boldsymbol{z}^\top \nabla^2 \Phi_\lambda \boldsymbol{z} = 2\,\|J(\boldsymbol{\theta}_\lambda)\,\boldsymbol{z}\|^2 - \lambda \sum_k \partial_k E_Y(Y)\,\boldsymbol{z}^\top \nabla^2 g_k(\boldsymbol{\theta}^\star)\boldsymbol{z} + \lambda\,\boldsymbol{z}^\top \nabla^2_{\boldsymbol{\theta}} R(\boldsymbol{\theta}^\star)\boldsymbol{z} + o(\lambda)\,\|\boldsymbol{z}\|^2 .$$

Along a geodesic γ of $\Theta^\star$ through $\boldsymbol{\theta}^\star$ with $\dot\gamma(0) = \boldsymbol{z}$, $2\,\|\boldsymbol{z}\|^2 = \frac{d^2}{dt^2} R(\gamma)|_0 = \boldsymbol{z}^\top \nabla^2_{\boldsymbol{\theta}} R\,\boldsymbol{z} + \langle \nabla_{\boldsymbol{\theta}} R, \ddot\gamma(0)\rangle$ and $\langle \nabla g_k, \ddot\gamma(0)\rangle = -\boldsymbol{z}^\top \nabla^2 g_k \boldsymbol{z}$, so the two sums cancel and the vertical block is $2\lambda\,\|\boldsymbol{z}\|^2 + o(\lambda)\,\|\boldsymbol{z}\|^2$; the Schur complement makes the Hessian positive definite for small λ. *Slow flow.* $(I - P_{\boldsymbol{\theta}^\star})\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}^\star) = \mathbf{0}$ by (i) (Corollary D.7).

Limit. With M_R, M_π bounds for $\nabla_{\boldsymbol{\theta}} R$ and $d\pi$ and L_R a Lipschitz constant of R on $\mathcal{K}$, $\|\boldsymbol{r}(t)\| \le \|\boldsymbol{r}(0)\|\, e^{-2c^2 t} + \lambda C M_R/(2c^2)$, and at $t_1 := \log(1/\lambda)/(2c^2)$, $\|\boldsymbol{r}(t_1)\| = O(\lambda)$, $\|\pi(\boldsymbol{\theta}(t_1)) - \boldsymbol{\theta}^\star\| \le \lambda M_\pi M_R t_1$ and $\|\boldsymbol{\theta}(t_1) - \pi(\boldsymbol{\theta}(t_1))\| \le C\,\|\boldsymbol{r}(t_1)\|/c^2$ (Proposition D.1), so $\Phi_\lambda(\omega_\lambda) \le \Phi_\lambda(\boldsymbol{\theta}(t_1)) \le \Phi_\lambda(\boldsymbol{\theta}^\star) + \epsilon_\lambda$ with $\epsilon_\lambda = O(\lambda^2) + \lambda L_R\, O(\lambda \log(1/\lambda)) = o(\lambda)$. The argument of Proposition D.18 uses only this comparison and gives $\omega_\lambda \to \boldsymbol{\theta}^\star$. □

The intrinsic dynamics is canonical: admissible, unbiased by basin invariance, and obeying the force bound. The ambient dynamics, gradient descent on $\mathcal{L} + \lambda R$, is horizontal but not admissible where the basin manifold is tilted, moves the training limit at a rate proportional to the residual, and is unbiased in the limit through its critical point rather than through invariance; the two coincide at $\boldsymbol{\theta}^\star$ and, in the integrable class, throughout the chart. The same computation is a prediction for any regulariser applied through the horizontal projection $P_{\boldsymbol{\theta}}$ of its gradient: the projection removes the fibre component but does not make the force admissible, and the training limit drifts at a rate proportional to the current residual, whereas under the arc-ridge force it does not move at all.

E.7 The sandwich

Proposition E.6 (Anchored ridge lower-bounds the canonical regulariser)**.** *Under Assumption 2.1, for every $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$,*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \;\le\; d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \;=\; d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\,.$$

Anchored ridge is the least energy over all paths, geodesic ridge over paths confined to $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ (relaxation).

Proof of Proposition E.6. The fibre term vanishes at $\boldsymbol{\theta} = g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ (Corollary D.4), and every path in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is at least as long as the chord (Lemmas C.6, C.7). □

Corollary E.1 (Sandwich bound and shared vanishing-regularisation limit)**.** *Under Assumption 2.1, for every $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ reached by a piecewise-C^1 path γ_{train} in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$,*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \;\le\; d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \;=\; d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \;\le\; L^2(\gamma_{\mathrm{train}}).$$

Geodesic ridge attains the gradient-flow interpolator as its vanishing-regularisation MAP limit (Theorem 4.2(c)). Arc ridge is not the gradient of a function on Θ, and the corresponding statement is dynamical: for every $\lambda > 0$ the arc-ridge-regularised flow (Definition E.1) runs along the unregularised trajectory and halts at $\phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, T^\star_\lambda)$, which tends to $\boldsymbol{\theta}^\star$ as $\lambda \downarrow 0$ (Proposition E.3, Theorem E.4). Both are unbiased in the sense that their regularised dynamics (the intrinsic one for geodesic ridge, Proposition E.5) never leave $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ and their $\lambda \downarrow 0$ limits coincide with $\boldsymbol{\theta}^\star$; anchored ridge fails the first property, and fails the second whenever the alignment condition of Theorem D.11(c) fails. Arc ridge is the minimax-robust choice within the sandwich (Theorem E.7).

Proof of Corollary E.1. The lower bound is Proposition E.6; the identity $d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap V$ is the vanishing of the fibre term of Definition 4.1 at $\boldsymbol{\theta} = g|_{\boldsymbol{\theta}_0}^{-1}(g(\boldsymbol{\theta}))$ (Corollary D.4); the upper bound is Theorem 4.3(a). The limit statement for geodesic ridge is Theorem 4.2(c); for arc ridge it is Proposition E.3(iii) and (v). Basin invariance of both regularised dynamics is Theorem E.4(a), whose admissibility hypothesis holds for the intrinsic gradient of geodesic ridge by Proposition E.5(ii) and for the arc-ridge force by Definition E.1. The statements on anchored ridge are Theorem D.11, and the minimax statement is Theorem E.7(d). □

E.8 MINIMAX-ROBUST MOTIVATION

At each time the accumulated arc length is known exactly while the canonical regulariser is not; arc ridge is the least upper bound over the geometries a run cannot distinguish.

Theorem E.7 (Least-favourable geometry; minimax optimality of arc ridge)**.** *Let $\gamma : [0, T] \to \mathbb{R}^m$ be a C^∞ embedded curve from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta} := \gamma(T) \neq \boldsymbol{\theta}_0$ with length $L := L(\gamma)$, and for $1 \leq n \leq m-1$ let $\mathfrak{M}_n(\gamma)$ be the class of injectively immersed C^2 n-dimensional submanifolds $M \subset \mathbb{R}^m$ (embedded ones included) with $\gamma([0, T]) \subset M$, each carrying the metric induced from $\mathbb{R}^m$ and its intrinsic distance $d_M(\cdot, \cdot)$: the* geometries consistent with the observed path*. Then:*

- *(a)* ***Sandwich in every geometry.*** *$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq d_M(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq L$ for every $M \in \mathfrak{M}_n(\gamma)$.*
- *(b)* ***The upper end is attained in the limit.*** *$\sup_{M \in \mathfrak{M}_n(\gamma)} d_M(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = L$.*
- *(c)* ***The lower end.*** *Whenever some $M \in \mathfrak{M}_n(\gamma)$ contains the segment $[\boldsymbol{\theta}_0, \boldsymbol{\theta}]$, $\inf_{M \in \mathfrak{M}_n(\gamma)} d_M(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$, attained at that M.*
- *(d)* ***Minimax optimality.*** *Call a regulariser* path-determined *if it is a functional $\Phi(\gamma)$ of the observed path alone, and* robust *if $\Phi(\gamma) \geq d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ for every $M \in \mathfrak{M}_n(\gamma)$. Every robust path-determined regulariser satisfies $\Phi(\gamma) \geq L^2(\gamma)$, and L^2 is robust: arc ridge is the pointwise least robust path-determined regulariser, the unique minimiser among them of the worst-case over-regularisation $\sup_M \big[\Phi(\gamma) - d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})\big]$, and the robust objective is*

$$\sup_{M \in \mathfrak{M}_n(\gamma)} \big[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda\, d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})\big] \;=\; \mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda\, L^2(\gamma).$$

- *(e)* ***Tightness.*** *Under Assumption 2.1, if γ lies in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ then $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \in \mathfrak{M}_n(\gamma)$, and $L^2(\gamma) - d_{\text{flow}}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0$ if and only if γ is a minimising geodesic of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$; the excess never exceeds the width $L^2(\gamma) - \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ of the sandwich.*

Proof. (a) The upper bound holds because γ is a path in M from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$; the lower bound is Lemma C.6, whose proof uses only that paths in M are paths in $\mathbb{R}^m$.

(b) Thin tube. Extend γ to a C^∞ embedded arc $\tilde{\gamma}$ on $(-\eta, T + \eta)$, with compact sub-arc $A := \tilde{\gamma}([-\eta/2, T + \eta/2])$, curvature bound $\bar{\kappa}$, and reach $\rho > 0$ (Federer, 1959, §4) (positive since $\dot{\tilde{\gamma}}$ is bounded away from zero, $\ddot{\tilde{\gamma}}$ is bounded, and A is compact and embedded), so that the nearest-point map ξ onto A is $\rho/(\rho - \delta)$-Lipschitz on the closed δ-neighbourhood, $\delta < \rho$. With parallel orthonormal normal fields $\nu_1, \ldots, \nu_{n-1}$ along the interval and $0 < \delta < \min\{\rho, 1/\bar{\kappa}\}$, set

$$\Psi(t, v) := \tilde{\gamma}(t) + \sum_{i=1}^{n-1} v_i\, \nu_i(t), \qquad M_\delta := \Psi\big((-\eta/2, T + \eta/2) \times \{\|v\| < \delta\}\big).$$

$\xi \circ \Psi(t, v) = \tilde{\gamma}(t)$, so Ψ is injective with continuous inverse, and the tangential component $\big\|\dot{\tilde{\gamma}}\big\|(1 - \sum_i v_i \langle \nu_i, \kappa \rangle) > 0$ of $\partial_t \Psi$ makes it an immersion; so $M_\delta \in \mathfrak{M}_n(\gamma)$. For a path σ in M_δ from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$, $\xi \circ \sigma$ runs along A from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$, so $L \leq L(\xi \circ \sigma) \leq \frac{\rho}{\rho - \delta} L(\sigma)$, and $d_{M_\delta}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq (1 - \delta/\rho) L \to L$.

(c) If $M \supset [\boldsymbol{\theta}_0, \boldsymbol{\theta}]$ then the segment is a path in M, so $d_M(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$, and Lemma C.6 gives equality.

(d) Robustness at M_δ and (b) give $\Phi(\gamma) \geq \sup_M d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = L^2$; and L^2 is robust by (a). For any robust Φ, $\sup_M [\Phi(\gamma) - d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})] = \Phi(\gamma) - \inf_M d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq L^2 - \inf_M d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, with equality only if $\Phi(\gamma) = L^2$; and the identity for the robust objective is (b) with $\lambda > 0$.

(e) $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is an embedded C^∞ n-dimensional submanifold containing γ (Theorem D.3(i)); the equality case is Proposition E.1, and the last claim is (a) applied to $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$. □

Three points delimit the theorem. $\mathfrak{M}_n(\gamma)$ is a geometric relaxation: it contains thin tubes that are the flow manifold of no network, and restricting to network-realisable geometries could in principle admit tighter path-determined bounds, of which none is known; for $n = 1$ the relaxation is exact, the flow manifold being the gradient line through $\boldsymbol{\theta}_0$ (Theorem D.19; Figure 2). The criterion is robust upper-bounding; a symmetric criterion $\sup_M |\Phi(\gamma) - d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})|$ would select the midpoint of the range. And the theorem concerns the value L^2; the arc-ridge force is minimax only as the intrinsic gradient of s^2 on the training curve (Proposition E.2).

Proposition E.8 (Minimax-robust regulariser)**.** *Under the hypotheses of Proposition E.3, fix $\lambda > 0$ and, for $\boldsymbol{\theta} = \gamma_{\mathrm{gf}}(t) \in \Gamma_{\mathrm{gf}}$, let the canonical regulariser be known only through the geometries consistent with the path observed up to $\boldsymbol{\theta}$, so that its value ranges over $\{d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}) : M \in \mathfrak{M}_n(\gamma_{\mathrm{gf}}|_{[0,t]})\} \subseteq [\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2, s(\boldsymbol{\theta})^2]$. The robust MAP problem on the training curve is the arc-ridge MAP problem on the training curve,*

$$\min_{\boldsymbol{\theta} \in \Gamma_{\mathrm{gf}}} \sup_{M \in \mathfrak{M}_n(\gamma_{\mathrm{gf}}|_{[0,t]})} \left[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda\, d_M^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})\right] \;=\; \min_{\boldsymbol{\theta} \in \Gamma_{\mathrm{gf}}} \left[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda\, s(\boldsymbol{\theta})^2\right];$$

its minimum is attained on Γ_{gf} at a point satisfying the stopping condition (1)*, and when the zero of ψ_λ is unique that point is the stopping point $\gamma_{\mathrm{gf}}(T_\lambda^\star)$ of the arc-ridge dynamics.*

Proof. By Proposition E.2 and Theorem E.7(d) the inner supremum is $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda s(\boldsymbol{\theta})^2$. In arc length this is ϕ, continuous on $[0, s_\infty]$ with $\phi' = \psi_\lambda$, negative at 0 and tending to $2\lambda s_\infty > 0$, so its minimum is at an interior zero of ψ_λ, which is $T_\lambda^\star$ when unique. □

For a loss read as $\mathcal{L} = \alpha\big(-\log p(Y \mid \boldsymbol{c})\big)$ up to a constant, the computation after Proposition E.3 gives $\lambda = \alpha\beta$ ($\alpha = 2\sigma^2$ for the Gaussian likelihood).

E.9 Tightness in the kernel regime

Proposition E.9 (Tightness of arc ridge in the kernel regime)**.** *In the kernel regime of Definition 2.1, write $K = JJ^\top = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ for the constant NTK Gram matrix and $g(\boldsymbol{\theta}_0) - Y = \sum_{i=1}^n \rho_i \boldsymbol{v}_i$. The gradient-flow trajectory has total arc length*

$$s_\infty = \int_0^\infty 2\Big(\sum_{i=1}^n \lambda_i \rho_i^2 e^{-4\lambda_i t}\Big)^{1/2} dt \;\geq\; \Big(\sum_{i=1}^n \rho_i^2/\lambda_i\Big)^{1/2} = \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\| = d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star),$$

with equality if and only if $g(\boldsymbol{\theta}_0) - Y$ is an eigenvector of K.

Proof. $\boldsymbol{r}(t) = e^{-2Kt}\boldsymbol{r}(0)$, so $\left\|\dot{\boldsymbol{\theta}}\right\|^2 = 4\boldsymbol{r}^\top K \boldsymbol{r} = \|\boldsymbol{a}(t)\|^2$ with $a_i := 2\sqrt{\lambda_i}|\rho_i|e^{-2\lambda_i t}$, $\int_0^\infty a_i = |\rho_i|/\sqrt{\lambda_i}$; the triangle inequality for integrals gives the bound, with equality iff $\boldsymbol{a}$ has constant direction, iff all λ_i with $\rho_i \neq 0$ coincide. $\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0 = -J^\top K^{-1}\boldsymbol{r}(0)$ has squared norm $\sum_i \rho_i^2/\lambda_i$, and the geodesic distance is the chord (Proposition D.15(e)). □

Arc ridge is therefore loose even in the kernel regime unless the initial residual is an NTK eigenvector, and the slack is a spectral-anisotropy functional of K on the residual: the trajectory direction rotates through the eigenspaces of K, and the training path is a geodesic of the affine flow manifold exactly when it does not. Under natural-gradient preconditioning by K^{-1} the kernel-regime trajectory is the straight segment $\boldsymbol{\theta}(t) = \boldsymbol{\theta}_0 - J^\top K^{-1}(1 - e^{-2t})(g(\boldsymbol{\theta}_0) - Y)$, and arc ridge, anchored ridge, and geodesic ridge coincide exactly.

E.10 Implementation, and comparison with early stopping

Remark E.1 (Implementation under momentum and minibatch noise). The exact statements of this appendix are gradient-flow statements; the implementation used in Section 5 retains horizontality exactly, admissibility under full-batch steps, and the value bound approximately. At each step the arc-ridge force is $-2s_k\, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{B_k} / \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{B_k}\|$, where $\mathcal{L}_{B_k}$ is the loss on the current minibatch B_k and $s_k := \sum_{j \leq k} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}\|$ is the polygonal length of the realised path in effective-parameter space, accumulated as a single scalar. The direction is the current minibatch data gradient, which lies in $\operatorname{Im} J_{B_k}(\boldsymbol{\theta}_k)^\top \subseteq \operatorname{Im} J(\boldsymbol{\theta}_k)^\top$, J_{B_k} being the rows of J indexed by B_k, so the force is horizontal at every step for every batch; with the full batch its direction is that of $X(\boldsymbol{\theta}_k)$, tangent to the flow manifold through $\boldsymbol{\theta}_k$ (Theorem D.2), whereas a minibatch direction need not be, the tangent spaces of flow manifolds differing from $\operatorname{Im} J(\cdot)^\top$ away from $\Theta^\star$ (Theorem D.5). The momentum buffer is not admissible: it sums data gradients taken at past iterates, and $\operatorname{Im} J(\boldsymbol{\theta}_{k-j})^\top \neq \operatorname{Im} J(\boldsymbol{\theta}_k)^\top$ in the feature-learning regime, the rotating-subspace mechanism of Lemma D.5; a force aligned with the unit update direction would therefore acquire a fibre component under momentum, and we do not use

one. The force is added to the data gradient before the update enters the momentum buffer, so that the buffer integrates the regularised velocity; the fibre leakage of momentum SGD itself, heuristically of order $\eta/(1-\mu)$ per unit time for momentum μ and step η (the effective step of heavy-ball momentum; we do not derive it), is a property of the optimiser and lies outside the exact statements. Under full-batch gradient descent the value bound holds up to $O(\eta)$ on finite horizons (proof of Proposition E.1). Under minibatch noise the polygonal length over-counts the length of the underlying flow, since $\mathbb{E}\,\|\Delta\boldsymbol{\theta}\| \geq \|\mathbb{E}\,\Delta\boldsymbol{\theta}\|$, which loosens the bound in the conservative direction and acts as additional regularisation whose size grows with the noise level and with $\mu/(1-\mu)$. In discrete time the factor $1 - 2\lambda s_k/\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{B_k}\right\|$ can overshoot zero, the iterate then steps uphill, s ratchets upward under any motion, and the balance point $2\lambda s = \|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|$ drifts away from interpolation; the validation-based early stopping of the experiments caps this drift. For the same reason the early-stopping equivalence of Proposition E.3 is exact only under gradient flow: with momentum the update is the buffer, which carries the history of the force, and the discrete path is not a truncation of the unregularised one. The experiments report λ in the convention of the released code, stated in Appendix H.

Remark E.2 (Arc ridge and early stopping). Proposition E.3 gives early stopping its theory in the feature-learning regime and separates the two: the criterion (1) needs no validation split, and early stopping returns a time, whereas arc ridge returns a regularisation value, s^2, comparable across runs and sandwiched against the canonical regulariser (Corollary E.1). And the equivalence explains why early stopping transfers to feature-learning networks while anchored ridge does not: its action is confined to the realised flow direction, and its stopping point converges to the gradient-flow limit as the regularisation vanishes (Theorem E.4).

F Pushforward Energies and the Training-Cost Function

Every function-space statement of the paper is the contraction of a parameter-space object under g, the infimum over the fibre. Throughout, A.1–A.3 and Assumption 2.1 hold, with the notation of Appendix D.2: $B := B(\boldsymbol{\theta}_0, R)$, $\ell := C\,\|g(\boldsymbol{\theta}) - Y\|/c^2$, the flow ϕ_t, the trapping region $\mathfrak{B}$, the training-limit map π, and V, $C_0(\boldsymbol{\theta}_0) = B(Y,\rho)$ and $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ of Theorem D.4.

F.1 The energy of the training problem

Read as a MAP objective, $\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is a Gaussian likelihood of the training outputs with noise variance σ^2 and a prior $\exp(-\beta d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0))$ on parameters, with $\lambda = 2\beta\sigma^2$ (Proposition E.8). The likelihood sees $\boldsymbol{\theta}$ only through $g(\boldsymbol{\theta})$, so what the data can act on is the least value of the regulariser on each output fibre.

Proposition F.1 (Energy of the training problem)**.** *For $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0)$ let $E_Y(\boldsymbol{c}) := \inf\{d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) : g(\boldsymbol{\theta}) = \boldsymbol{c}\}$, the contraction of geodesic ridge under g. Then:*

- *(i) $E_Y(\boldsymbol{c}) = d^2_{\mathrm{flow}}\left(\boldsymbol{\theta}_0, g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})\right)$, the infimum being attained exactly at $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$; E_Y is continuous, $E_Y(Y) = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star)$, and $E_Y(\boldsymbol{c}) \geq C^{-2}\,\|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\|^2$.*
- *(ii) For every $\lambda > 0$,*

$$\inf_{\boldsymbol{\theta}\in\Theta}\left[\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\right] \;=\; \inf_{\boldsymbol{c}\in C_0(\boldsymbol{\theta}_0)}\left[\|\boldsymbol{c} - Y\|^2 + \lambda E_Y(\boldsymbol{c})\right];$$

  for $\lambda E_Y(Y) < \rho^2/4$ both infima are attained, and the minimisers of the left-hand side are exactly the representatives $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}_\lambda)$ of the minimisers $\boldsymbol{c}_\lambda$ of the right-hand side.
- *(iii) Whenever the trajectory $\boldsymbol{\theta}(t) = \phi_{\mathrm{gf}}(\boldsymbol{\theta}_0, t)$ lies in V, $E_Y(g(\boldsymbol{\theta}(t))) = d^2_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}(t)) \leq s(t)^2$. The length of any horizontal motion $\boldsymbol{\theta}(\cdot)$ in B is the inverse-NTK length $\int(\dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta})^{-1}\dot{\boldsymbol{c}})^{1/2}\,dt$ of its output path $\boldsymbol{c}(\cdot) = g(\boldsymbol{\theta}(\cdot))$, the NTK evaluated along the motion; in particular $s(t) = \int_0^t(\dot{\boldsymbol{c}}^\top K(\boldsymbol{\theta})^{-1}\dot{\boldsymbol{c}})^{1/2}\,dt'$.*
- *(iv) Through the chart, $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)\cap V$ is isometric to $C_0(\boldsymbol{\theta}_0)$ with the metric $G = K^{-1} + N^\top N$ of Theorem D.5: $\sqrt{E_Y}$ is 1-Lipschitz for the G-distance of $C_0(\boldsymbol{\theta}_0)$, and $E_Y(\boldsymbol{c})$ is the squared G-distance from $g(\boldsymbol{\theta}_0)$ to $\boldsymbol{c}$ whenever $\boldsymbol{\theta}_0$ and a minimising geodesic of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ from $\boldsymbol{\theta}_0$ to $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ lie in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)\cap V$. Along the trajectory G measures the velocity as the inverse NTK does, $\dot{\boldsymbol{c}}^\top G\dot{\boldsymbol{c}} = \dot{\boldsymbol{c}}^\top K^{-1}\dot{\boldsymbol{c}}$, and along its tail $G = K^{-1} + O(\|g(\boldsymbol{\theta}(t)) - Y\|^2)$.*
- *(v) In the kernel regime $E_Y(\boldsymbol{c}) = (\boldsymbol{c} - g(\boldsymbol{\theta}_0))^\top K^{-1}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$, for every label vector Y.*

Proof. (i) Geodesic ridge is finite only on V and minimised on $\Theta(\boldsymbol{c}) \cap V$ exactly at $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$ (Theorem 4.2(a)); continuity is in the proof of Proposition D.17(i). For a path γ in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \subseteq B$ from $\boldsymbol{\theta}_0$ to $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})$, $\|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\| = \left\|\int J(\gamma)\,\dot{\gamma}\right\| \le C\,L(\gamma)$.

(ii) Geodesic ridge is the R' of Proposition D.17 with S the squared fibre distance, and (i) of that proposition applies.

(iii) The first statement is Theorem 4.2(b); for $\dot{\boldsymbol{\theta}} = J(\boldsymbol{\theta})^{\top}\boldsymbol{u}$, $\dot{\boldsymbol{c}} = K(\boldsymbol{\theta})\boldsymbol{u}$ and $\left\|\dot{\boldsymbol{\theta}}\right\|^2 = \dot{\boldsymbol{c}}^{\top}K(\boldsymbol{\theta})^{-1}\dot{\boldsymbol{c}}$.

(iv) $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ is an isometry of $(C_0(\boldsymbol{\theta}_0), G)$ onto $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$, so $d_{\text{flow}}\left(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}), g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}')\right) \le d_G(\boldsymbol{c}, \boldsymbol{c}')$ and the triangle inequality gives the Lipschitz bound; when $\boldsymbol{\theta}_0$ and a minimising geodesic lie in $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) \cap V$, its image is a G-shortest path. The trajectory statements are Theorem D.5 and Lemma D.3. *(v)* is Proposition D.15(c). □

By (ii) the regularised estimator is the representative of the shrunk outputs, the prior acting in output space through E_Y; by (iii) arc ridge is an output-space quantity, the inverse-NTK length of the output path; by (iv)–(v) the feature-learning content of E_Y is the tilt term $N^{\top}N$ of its metric.

Remark F.1 (Pushforward without a volume measure). Proposition F.1(ii) is a statement about energies, and it is all the MAP reading needs; it is also the leading exponential order of a pushforward of measures, in which no Riemannian volume enters. Let $U_\star$ be the neighbourhood of $\boldsymbol{\theta}^\star$ on which the proof of Proposition D.17(ii) makes the fibre term smooth: $g(U_\star) =: C_0'$ is an open ball about Y, $g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c}) \in U_\star$ for every $\boldsymbol{c} \in C_0'$, and geodesic ridge $F := d^2(\cdot, \boldsymbol{\theta}_0)$ is continuous on the bounded open set $U_\star$. For $\beta \to \infty$ the probability measures $\mu_\beta(d\boldsymbol{\theta}) \propto \exp(-\beta F(\boldsymbol{\theta}))\mathbf{1}_{U_\star}(\boldsymbol{\theta})\,d\boldsymbol{\theta}$, with Lebesgue reference measure on Θ, satisfy a large-deviation principle with speed β and good rate function I equal to $F - \inf_{U_\star} F$ on $U_\star$, extended to $\overline{U_\star}$ by lower semicontinuity and by $+\infty$ elsewhere. This is Laplace's method. By the contraction principle (Dembo & Zeitouni, 2010, §4.2) the images $g_{\#}\mu_\beta$ satisfy a large-deviation principle with rate $\boldsymbol{c} \mapsto \inf\{I(\boldsymbol{\theta}) : g(\boldsymbol{\theta}) = \boldsymbol{c}\}$, which on C_0' is $E_Y - \inf_{C_0'} E_Y$: at points of $\overline{U_\star}$ over $\boldsymbol{c} \in C_0'$ the envelope of F is at least $E_Y(\boldsymbol{c})$, because $F \ge E_Y \circ g$ and E_Y is continuous, and $F(g|_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{c})) = E_Y(\boldsymbol{c})$ (Proposition F.1(i)). Rate functions are insensitive to a continuous positive reference density, and no Riemannian volume on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ or on output space enters: $p(\boldsymbol{c} \mid X) \propto \exp(-\beta E_Y(\boldsymbol{c}))$ is the leading exponential order. In the kernel regime the pushforward is exact at the level of measures (Proposition D.15(f)).

F.2 The training map and the reachable manifold

E_Y is the energy of one training problem, and it depends on the labels through $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ (Theorem D.8). What does not depend on them is built from the family of all training problems that share g and $\boldsymbol{\theta}_0$. Write

$$\ell_{\boldsymbol{c}} := C\,\|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\|/c^2, \qquad \mathcal{N} := \left\{\boldsymbol{c} \in \mathbb{R}^n : \ell_{\boldsymbol{c}} < R\right\},$$

an open ball about $g(\boldsymbol{\theta}_0)$ determined by g, $\boldsymbol{\theta}_0$ and the conditioning ball of Assumption 2.1, whose radius condition is exactly $Y \in \mathcal{N}$. Every $\boldsymbol{c} \in \mathcal{N}$ is therefore a label vector for which Assumption 2.1 holds with the same ball and constants, and Appendix D.2 applies to it verbatim. For $\boldsymbol{c} \in \mathcal{N}$ we write $\phi_t^{\boldsymbol{c}}$, $\mathfrak{B}_{\boldsymbol{c}}$, $\pi_{\boldsymbol{c}}$ and $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$ for the flow, trapping region, training-limit map and flow manifold of the problem with labels $\boldsymbol{c}$, and $d_{\boldsymbol{c}}$ for the intrinsic distance of $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$; for $\boldsymbol{c} = Y$ these are ϕ_t, $\mathfrak{B}$, π, $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ and $d_{\text{flow}}(\cdot,\cdot)$, and $\ell_Y = \ell$.

Theorem F.2 (The training map and the reachable manifold)**.** *(a) For every $\boldsymbol{c} \in \mathcal{N}$ the gradient flow $\phi_t^{\boldsymbol{c}}(\boldsymbol{\theta}_0)$ of $\|g(\boldsymbol{\theta}) - \boldsymbol{c}\|^2$ from $\boldsymbol{\theta}_0$ stays in B and converges to an interpolator $\tau(\boldsymbol{c}) \in g^{-1}(\boldsymbol{c}) \cap B$, with $\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\| \le \ell_{\boldsymbol{c}}$; $\tau(g(\boldsymbol{\theta}_0)) = \boldsymbol{\theta}_0$ and $\tau(Y) = \boldsymbol{\theta}^\star$.*

(b) The training map $\tau : \mathcal{N} \to B$ is C^∞, and $g \circ \tau = \mathrm{id}_{\mathcal{N}}$.

(c) The reachable manifold *$\mathcal{R}(\boldsymbol{\theta}_0) := \tau(\mathcal{N})$ is an embedded C^∞ n-dimensional submanifold of B through $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}^\star$, on which g is a diffeomorphism onto $\mathcal{N}$ with inverse τ. It is determined by g, $\boldsymbol{\theta}_0$, the conditioning ball and the loss form, and by no label vector.*

(d) $D\tau(g(\boldsymbol{\theta}_0)) = J(\boldsymbol{\theta}_0)^{\top}K(\boldsymbol{\theta}_0)^{-1} = J(\boldsymbol{\theta}_0)^{+}$, so $T_{\boldsymbol{\theta}_0}\mathcal{R}(\boldsymbol{\theta}_0) = \mathrm{Im}\,J(\boldsymbol{\theta}_0)^{\top}$. At every $\boldsymbol{c} \in \mathcal{N}$, $D\tau(\boldsymbol{c}) = J(\tau(\boldsymbol{c}))^{\top}K(\tau(\boldsymbol{c}))^{-1} + N_{\mathcal{R}}(\boldsymbol{c})$ with $N_{\mathcal{R}}(\boldsymbol{c}) := (I - P_{\tau(\boldsymbol{c})})D\tau(\boldsymbol{c})$

vertical, and the metric induced on $\mathcal{N}$ through τ is $G_{\mathcal{R}} = K^{-1} + N_{\mathcal{R}}^{\top} N_{\mathcal{R}}$, with K evaluated at $\tau(\boldsymbol{c})$; $N_{\mathcal{R}}(g(\boldsymbol{\theta}_0)) = 0$.

(e) The flow manifolds $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$, $\boldsymbol{c} \in \mathcal{N}$, all contain $\boldsymbol{\theta}_0$, and $\tau(\boldsymbol{c})$ is the unique interpolator of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$: $\mathcal{R}(\boldsymbol{\theta}_0)$ is the set of networks that training from $\boldsymbol{\theta}_0$ produces. A point $\boldsymbol{\theta} \in B$ with $g(\boldsymbol{\theta}) \in \mathcal{N}$ lies on $\mathcal{R}(\boldsymbol{\theta}_0)$ if and only if $\boldsymbol{\theta} = \tau(g(\boldsymbol{\theta}))$, that is, if and only if training from $\boldsymbol{\theta}_0$ with labels $g(\boldsymbol{\theta})$ ends at $\boldsymbol{\theta}$. If $\mathcal{R}(\boldsymbol{\theta}_0)$ is invariant under the gradient flows of an open set of label vectors, it is tangent to $\mathcal{H}$ at each of its points.

(f) In the integrable class, $\mathcal{R}(\boldsymbol{\theta}_0)$ is an open subset of the leaf $\mathcal{L}(\boldsymbol{\theta}_0)$ of $\mathcal{H}$ through $\boldsymbol{\theta}_0$, tangent to $\mathcal{H}$ everywhere, $\tau = g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ on $\mathcal{N} \cap C_0(\boldsymbol{\theta}_0)$, and the trajectory of every problem with labels in $\mathcal{N}$ lies on $\mathcal{R}(\boldsymbol{\theta}_0)$ for as long as its outputs remain in $\mathcal{N}$. In the kernel regime $\tau(\boldsymbol{c}) = \boldsymbol{\theta}_0 + J^{+}(\boldsymbol{c} - g(\boldsymbol{\theta}_0))$, and $\mathcal{R}(\boldsymbol{\theta}_0)$ is the part of the plane $\boldsymbol{\theta}_0 + \operatorname{Im} J^{\top}$ over $\mathcal{N}$.

Proof. *(a)* is Proposition 2.1 with labels $\boldsymbol{c}$; $\boldsymbol{\theta}_0$ is an equilibrium for the labels $g(\boldsymbol{\theta}_0)$.

(b) The augmented system. On $\mathcal{O} := B \times \mathcal{N}$ let $\hat{X}(\boldsymbol{\theta}, \boldsymbol{c}) := \left(-2J(\boldsymbol{\theta})^{\top}(g(\boldsymbol{\theta}) - \boldsymbol{c}),\ \mathbf{0}\right)$, with flow $\hat{\phi}_t(\boldsymbol{\theta}, \boldsymbol{c}) = (\phi_t^{\boldsymbol{c}}(\boldsymbol{\theta}), \boldsymbol{c})$. Its zero set is the graph $\Sigma := \{g(\boldsymbol{\theta}) = \boldsymbol{c}\}$, an embedded m-manifold with $T\Sigma = \{(\boldsymbol{v}, J\boldsymbol{v})\}$, and on Σ

$$D\hat{X}(\boldsymbol{\theta}, \boldsymbol{c}) \;=\; \begin{pmatrix} -2J^{\top}J & 2J^{\top} \\ 0 & 0 \end{pmatrix},$$

with kernel $T\Sigma$ and invariant complement $E^s := \operatorname{Im} J^{\top} \times \{\mathbf{0}\}$ carrying the eigenvalues $-2\lambda_i(K) \in [-2C^2, -2c^2]$ (Lemma C.5). So Lemma D.1 applies with $\mu := 2c^2$, and the stable fibre at $(\tau(\boldsymbol{c}), \boldsymbol{c})$, read in the slice $\mathbb{R}^m \times \{\boldsymbol{c}\}$, is a piece of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$.

(b) Smoothness. For $\boldsymbol{c}_1 \in \mathcal{N}$, a closed ball $\bar{N}_1 \subset \mathcal{N}$ about it and $r \in \mathbb{N}$, the lemma applied to the compact $K_1 := \{(\boldsymbol{\theta}, \boldsymbol{c}) \in \Sigma : \boldsymbol{c} \in \bar{N}_1,\ \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \max_{\bar{N}_1} \ell_{\boldsymbol{c}'}\}$ gives a C^r submersion $\Pi(\boldsymbol{\theta}, \boldsymbol{c}) = (\pi_{\boldsymbol{c}}(\boldsymbol{\theta}), \boldsymbol{c})$ on a neighbourhood U' of K_1. As in the proof of Theorem D.2, $\tau(\boldsymbol{c}) = \mathrm{pr}_{\boldsymbol{\theta}}\, \Pi(\hat{\phi}_T(\boldsymbol{\theta}_0, \boldsymbol{c}))$ near $\boldsymbol{c}_1$, so τ is C^{∞}, and $g \circ \tau = \mathrm{id}$.

(c) $J(\tau(\boldsymbol{c}))\, D\tau(\boldsymbol{c}) = I_n$, so τ is an injective immersion and a homeomorphism onto its image (inverse g), hence an embedding (Lee, 2012, Ch. 5). Only the losses $\|g(\boldsymbol{\theta}) - \boldsymbol{c}\|^2$, $\boldsymbol{c} \in \mathcal{N}$, enter; Y enters only through $Y \in \mathcal{N}$.

(d) As in the proof of Theorem D.5, $PD\tau = J^{\top}K^{-1}$ and $D\tau^{\top}D\tau = K^{-1} + N_{\mathcal{R}}^{\top}N_{\mathcal{R}}$. At $\boldsymbol{c}_0 := g(\boldsymbol{\theta}_0)$, $(\boldsymbol{\theta}_0, \boldsymbol{c}_0) \in \Sigma$, so $\tau = \mathrm{pr}_{\boldsymbol{\theta}} \circ \Pi(\boldsymbol{\theta}_0, \cdot)$ nearby and $D\Pi$ is the projection onto $T\Sigma$ along E^s ((iv) of the lemma): $(\mathbf{0}, \delta\boldsymbol{c}) = (\boldsymbol{v}, J_0\boldsymbol{v}) + (J_0^{\top}\boldsymbol{a}, \mathbf{0})$ gives $\boldsymbol{v} = J_0^{\top}K_0^{-1}\delta\boldsymbol{c}$, so $D\tau(\boldsymbol{c}_0) = J_0^{+}$ and $N_{\mathcal{R}}(\boldsymbol{c}_0) = 0$.

(e) Theorem D.3(i),(iii) for the labels $\boldsymbol{c}$, and Theorem D.8 for the last statement.

(f) By Theorem D.9(i) for each problem, $\tau(\boldsymbol{c}) \in \mathcal{L}(\boldsymbol{\theta}_0)$, and τ is smooth into the weakly embedded leaf (Lee, 2012, Ch. 19), an immersion of n-manifolds, hence open. $\tau = g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$ on $\mathcal{N} \cap C_0(\boldsymbol{\theta}_0)$ is in Corollary D.5. For a trajectory with labels in $\mathcal{N}$, the set of times at which $\boldsymbol{\theta}(t) = \tau(g(\boldsymbol{\theta}(t)))$ contains 0, is closed, and is open because g is injective on leaf-neighbourhoods (Theorem D.9(ii)) and both curves lie in the leaf with equal outputs. The kernel case is Corollary D.3. □

The reachable manifold and the flow manifold are dual. The flow manifold fixes the target and varies the start: it contains the trajectory, is tangent to the horizontal space at the trained network, and depends on the labels. The reachable manifold fixes the start and varies the target: it contains every trained network, is tangent to the horizontal space at the initialisation, and depends on no label vector. Both are n-dimensional charts of output space whose induced metrics exceed the inverse NTK by the square of their tilt out of the horizontal space, $N^{\top}N$ and $N_{\mathcal{R}}^{\top}N_{\mathcal{R}}$; they meet at $\boldsymbol{\theta}^{\star} = \tau(Y)$, and they share the trajectory only as far as training is route-independent, by (e). On the model of Remark D.3 with $\boldsymbol{\theta}_0 = (0, 0, 1)$, the finite-difference derivative of τ at $g(\boldsymbol{\theta}_0)$ agrees with $J(\boldsymbol{\theta}_0)^{+}$ to within $2 \cdot 10^{-9}$; at $Y = (0.95, 0.1)$, $J(\tau(Y))\, D\tau(Y) = I$ holds to within $3 \cdot 10^{-10}$ and $\|N_{\mathcal{R}}(Y)\|_F = 0.0178$; and along the trajectory to Y, $\|\boldsymbol{\theta}(t) - \tau(g(\boldsymbol{\theta}(t)))\| = 4.86{\cdot}10^{-6}, 4.37{\cdot}10^{-5}, 8.92 \cdot 10^{-5}, 6.25 \cdot 10^{-5}$ and $1.59 \cdot 10^{-6}$ at $t = 0.05, 0.2, 0.5, 1$ and 3, against $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|$ between 0.0136 and 0.0880: the trajectory is off $\mathcal{R}(\boldsymbol{\theta}_0)$ at every intermediate time tested, and meets it at its two ends.

F.3 THE TRAINING-COST FUNCTION

Each problem of the family carries its own energy $E_{\boldsymbol{c}}$, the energy of Proposition F.1 for the problem with labels $\boldsymbol{c}$, whose representative of its own labels is its trained network $\tau(\boldsymbol{c})$. Evaluating each at its own labels gives one function of the outputs:

Definition F.1 (Training-cost function)**.** For $\boldsymbol{c} \in \mathcal{N}$,

$$E(\boldsymbol{c}) \;:=\; E_{\boldsymbol{c}}(\boldsymbol{c}) \;=\; d^2_{\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}}\left(\boldsymbol{\theta}_0, \tau(\boldsymbol{c})\right),$$

the value of the regulariser of the problem with labels $\boldsymbol{c}$ at its own trained network: the least energy that outcome-preserving training must spend to produce the outputs $\boldsymbol{c}$ from $\boldsymbol{\theta}_0$ (Theorem D.16). It depends on $(g, \boldsymbol{\theta}_0)$, the conditioning ball and the loss form only.

The function E is the cost of training from $\boldsymbol{\theta}_0$ to each output vector, and $\exp(-\beta E)$ is its prior reading: a distribution over what the network can be trained to, fixed before any label is seen. Near the initial outputs it is Gaussian with the NTK covariance to second order, in every regime:

Theorem F.3 (The training-cost function at initialisation)**.** *Let $\mathcal{N}' := \{\boldsymbol{c} \in \mathbb{R}^n :\ (2 + C^2/c^2)\,\ell_{\boldsymbol{c}} < R\}$, a ball about $g(\boldsymbol{\theta}_0)$ contained in $\mathcal{N}$.*

- *(i) E is finite and upper semicontinuous on $\mathcal{N}$ and continuous on $\mathcal{N}'$, in particular at $g(\boldsymbol{\theta}_0)$; $E(Y) = E_Y(Y)$; and $\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\|^2 \le E(\boldsymbol{c}) \le s_\infty(\boldsymbol{c})^2 \le \ell^2_{\boldsymbol{c}}$, with $s_\infty(\boldsymbol{c})$ the length of the run from $\boldsymbol{\theta}_0$ to $\tau(\boldsymbol{c})$.*
- *(ii) With $\delta\boldsymbol{c} := \boldsymbol{c} - g(\boldsymbol{\theta}_0)$, $J_0 := J(\boldsymbol{\theta}_0)$, $K_0 := K(\boldsymbol{\theta}_0)$, $\boldsymbol{a} := K_0^{-1}\delta\boldsymbol{c}$, $\boldsymbol{v} := J_0^\top \boldsymbol{a} = J_0^+ \delta\boldsymbol{c}$, and $D_{\boldsymbol{v}}K$ the derivative of K at $\boldsymbol{\theta}_0$ along $\boldsymbol{v}$,*

$$E(\boldsymbol{c}) \;=\; \delta\boldsymbol{c}^\top K_0^{-1}\delta\boldsymbol{c} \;-\; \tfrac{1}{2}\,\delta\boldsymbol{c}^\top K_0^{-1}(D_{\boldsymbol{v}}K)K_0^{-1}\delta\boldsymbol{c} \;+\; O(\|\delta\boldsymbol{c}\|^4),$$

  and the same expansion holds for $\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\|^2$, for $d^2_{\mathcal{R}(\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0, \tau(\boldsymbol{c}))$ and for $\mathrm{dist}^2(\boldsymbol{\theta}_0, g^{-1}(\boldsymbol{c}))$: each of these label-independent energies is the inverse-NTK energy to second order at initialisation, with the same third-order term, the velocity of the NTK along the training direction, which vanishes when J is constant.
- *(iii) In the integrable class, if $Y \in \mathcal{N}'$ then $E = E_Y$ on $C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}'$. In the kernel regime $E(\boldsymbol{c}) = \delta\boldsymbol{c}^\top K^{-1}\delta\boldsymbol{c}$ for every $\boldsymbol{c} \in \mathcal{N}$.*

Proof. (i) Bounds. By Theorem D.3(i) for the problem with labels $\boldsymbol{c}$, the run $\phi^{\boldsymbol{c}}_t(\boldsymbol{\theta}_0)$ lies in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$ and converges to $\tau(\boldsymbol{c})$. Its restriction to $[0,t]$ is a path in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$ of length at most $s_\infty(\boldsymbol{c})$, and $d_{\boldsymbol{c}}(\phi^{\boldsymbol{c}}_t(\boldsymbol{\theta}_0), \tau(\boldsymbol{c})) \to 0$, the intrinsic distance of a connected embedded submanifold inducing its topology; hence $d_{\boldsymbol{c}}(\boldsymbol{\theta}_0, \tau(\boldsymbol{c})) \le s_\infty(\boldsymbol{c}) \le \ell_{\boldsymbol{c}}$ (Proposition 2.1), and Lemma C.6 gives the lower bound. $E(Y) = E_Y(Y)$ is the definition.

(i) The family of flow manifolds. Let $\hat{\mathfrak{B}} := \{(\boldsymbol{\theta}, \boldsymbol{c}) :\ \boldsymbol{c} \in \mathcal{N},\ \boldsymbol{\theta} \in \mathfrak{B}_{\boldsymbol{c}}\}$. It is open, its defining function $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + C\,\|g(\boldsymbol{\theta}) - \boldsymbol{c}\|/c^2$ being continuous on $B \times \mathcal{N}$; it is invariant under the augmented flow of the proof of Theorem F.2, and contains Σ (Proposition D.1 for each $\boldsymbol{c}$). Every trajectory in it converges to a point of Σ, and the proof of Theorem D.2, with Lemma D.1 applied to Σ as in the proof of Theorem F.2(b), shows that $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{c}) := (\pi_{\boldsymbol{c}}(\boldsymbol{\theta}), \boldsymbol{c})$ is a C^∞ submersion of $\hat{\mathfrak{B}}$ onto Σ. The graph of τ is an embedded n-dimensional submanifold of Σ, so its preimage

$$\mathcal{F} \;:=\; \{(\boldsymbol{\theta}, \boldsymbol{c}) \in \hat{\mathfrak{B}} :\ \pi_{\boldsymbol{c}}(\boldsymbol{\theta}) = \tau(\boldsymbol{c})\} \;=\; \bigcup_{\boldsymbol{c} \in \mathcal{N}} \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}} \times \{\boldsymbol{c}\}$$

is an embedded C^∞ submanifold of dimension $2n$ (Lee, 2012, Ch. 6), closed in $\hat{\mathfrak{B}}$. The projection $\mathrm{pr}(\boldsymbol{\theta}, \boldsymbol{c}) := \boldsymbol{c}$ is a submersion of $\mathcal{F}$ onto $\mathcal{N}$: for $z = (\boldsymbol{\theta}, \boldsymbol{c}) \in \mathcal{F}$ and $\boldsymbol{e} \in \mathbb{R}^n$ there is $\boldsymbol{u}$ with $d\hat{\pi}_z \boldsymbol{u} = (D\tau(\boldsymbol{c})\boldsymbol{e}, \boldsymbol{e})$, which lies in $T_z\mathcal{F}$ and has $\boldsymbol{c}$-component $\boldsymbol{e}$, since $\hat{\pi}$ preserves $\boldsymbol{c}$. Its fibres are the flow manifolds $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}} \times \{\boldsymbol{c}\}$, isometric to $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$.

(i) Transport between flow manifolds. For $\boldsymbol{e} \in \mathbb{R}^n$ let $\tilde{\boldsymbol{e}}(z)$ be the vector of $T_z\mathcal{F}$ orthogonal to $\ker d\,\mathrm{pr}_z$ with $d\,\mathrm{pr}_z \tilde{\boldsymbol{e}}(z) = \boldsymbol{e}$, smooth in $(z, \boldsymbol{e})$ and linear in $\boldsymbol{e}$, and let $\Phi_{\boldsymbol{e}}$ be the time-one map of the flow of $(z, \boldsymbol{e}) \mapsto (\tilde{\boldsymbol{e}}(z), \boldsymbol{0})$ on $\mathcal{F} \times \mathbb{R}^n$, a smooth field vanishing where $\boldsymbol{e} = \boldsymbol{0}$. For a compact $\mathcal{K} \subset \mathcal{F}$ the openness of the domain of a flow gives $\delta > 0$ such that $\Phi_{\boldsymbol{e}}$ is defined on $\mathcal{K}$ for $\|\boldsymbol{e}\| < \delta$, and as $\boldsymbol{e} \to \boldsymbol{0}$, uniformly on $\mathcal{K}$, $\Phi_{\boldsymbol{e}} \to \mathrm{id}$ and the operator norm of $D\Phi_{\boldsymbol{e}}$ on $T\mathcal{F}$ tends to 1. Since

$\mathrm{pr}\circ\Phi_{\boldsymbol{e}}=\mathrm{pr}+\boldsymbol{e}$, $\Phi_{\boldsymbol{e}}$ carries a piecewise-C^1 path in $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}}$ lying in $\mathcal{K}$ to one in $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}+\boldsymbol{e}}$ whose length is at most $1+\omega(\boldsymbol{e})$ times its own, $\omega(\boldsymbol{e})\to 0$. Two points of the same flow manifold near a point $z_1\in\mathcal{F}$ are joined in it by a path of length at most a constant times their distance: in a submersion chart $\psi: W\to U_1\times U_2$ of pr about z_1, bi-Lipschitz on W, with U_2 a ball and $\mathrm{pr}\circ\psi^{-1}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}$, take the image of a segment of $\{\boldsymbol{c}\}\times U_2$.

(i) Upper semicontinuity. Let $\boldsymbol{c}_1\in\mathcal{N}$, $\varepsilon>0$, and γ a path in $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}_1}$ from $\boldsymbol{\theta}_0$ to $\tau\left(\boldsymbol{c}_1\right)$ of length at most $E(\boldsymbol{c}_1)^{1/2}+\varepsilon$. For $\boldsymbol{c}$ near $\boldsymbol{c}_1$, $\Phi_{\boldsymbol{c}-\boldsymbol{c}_1}$ carries γ to a path in $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}}$ of length at most $(1+\omega)(E(\boldsymbol{c}_1)^{1/2}+\varepsilon)$ whose endpoints are close to $(\boldsymbol{\theta}_0,\boldsymbol{c}_1)$ and $\left(\tau\left(\boldsymbol{c}_1\right),\boldsymbol{c}_1\right)$; so are $(\boldsymbol{\theta}_0,\boldsymbol{c})$ and $\left(\tau\left(\boldsymbol{c}\right),\boldsymbol{c}\right)$, τ being continuous, and chart paths join the corresponding endpoints within $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}}$ at a cost tending to 0. Hence $\limsup_{\boldsymbol{c}\to\boldsymbol{c}_1}E(\boldsymbol{c})^{1/2}\le E(\boldsymbol{c}_1)^{1/2}+\varepsilon$ for every $\varepsilon>0$.

(i) Lower semicontinuity on $\mathcal{N}'$. Let $\boldsymbol{c}_k\to\boldsymbol{c}_1\in\mathcal{N}'$, $\bar{N}_1\subset\mathcal{N}'$ a closed ball about $\boldsymbol{c}_1$ containing every $\boldsymbol{c}_k$, and γ_k a path in $\mathcal{M}_{\mathrm{flow}}\left(\boldsymbol{\theta}_0\right)^{\boldsymbol{c}_k}$ from $\boldsymbol{\theta}_0$ to $\tau\left(\boldsymbol{c}_k\right)$ of length $L_k\le E(\boldsymbol{c}_k)^{1/2}+1/k\le\ell_{\boldsymbol{c}_k}+1/k$. A point $\boldsymbol{\theta}$ of γ_k lies in B, so $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le L_k$ and $\|g\left(\boldsymbol{\theta}\right)-\boldsymbol{c}_k\|\le CL_k+\|g\left(\boldsymbol{\theta}_0\right)-\boldsymbol{c}_k\|$, and the defining function of $\mathfrak{B}_{\boldsymbol{c}_k}$ is at most $(1+C^2/c^2)L_k+\ell_{\boldsymbol{c}_k}$ along γ_k. This is below some $R''<R$ for all large k, because $(2+C^2/c^2)\ell_{\boldsymbol{c}}<R$ on $\bar{N}_1$. The paths $(\gamma_k,\boldsymbol{c}_k)$ therefore lie in the set of points of $\mathcal{F}$ with $\boldsymbol{c}\in\bar{N}_1$ and defining function at most R'', which is compact, $\mathcal{F}$ being closed in $\hat{\mathfrak{B}}$. Transporting them by $\Phi_{\boldsymbol{c}_1-\boldsymbol{c}_k}$ and joining endpoints as before gives $E(\boldsymbol{c}_1)^{1/2}\le\liminf_k E(\boldsymbol{c}_k)^{1/2}$; with upper semicontinuity, E is continuous on $\mathcal{N}'$.

(ii) The chord. Let $p\in g^{-1}(\boldsymbol{c})$ with $p-\boldsymbol{\theta}_0=\boldsymbol{v}+\boldsymbol{w}$ and $\|\boldsymbol{w}\|=O(\|\delta\boldsymbol{c}\|^2)$ as $\delta\boldsymbol{c}\to\boldsymbol{0}$. Taylor's theorem for g at $\boldsymbol{\theta}_0$ (A.1) gives $\delta\boldsymbol{c}=J_0(\boldsymbol{v}+\boldsymbol{w})+\frac{1}{2}\left[\boldsymbol{v}^\top\nabla^2 g_i(\boldsymbol{\theta}_0)\boldsymbol{v}\right]_i+O(\|\delta\boldsymbol{c}\|^3)$, and $J_0\boldsymbol{v}=K_0\boldsymbol{a}=\delta\boldsymbol{c}$, so $J_0\boldsymbol{w}=-\frac{1}{2}\left[\boldsymbol{v}^\top\nabla^2 g_i(\boldsymbol{\theta}_0)\boldsymbol{v}\right]_i+O(\|\delta\boldsymbol{c}\|^3)$. With $\|\boldsymbol{v}\|^2=\boldsymbol{a}^\top K_0\boldsymbol{a}$, $\langle\boldsymbol{v},\boldsymbol{w}\rangle=\boldsymbol{a}^\top J_0\boldsymbol{w}$ and $\boldsymbol{a}=O(\|\delta\boldsymbol{c}\|)$,

$$\|p-\boldsymbol{\theta}_0\|^2\;=\;\|\boldsymbol{v}\|^2+2\langle\boldsymbol{v},\boldsymbol{w}\rangle+\|\boldsymbol{w}\|^2\;=\;\delta\boldsymbol{c}^\top K_0^{-1}\delta\boldsymbol{c}-\sum_i a_i\,\boldsymbol{v}^\top\nabla^2 g_i(\boldsymbol{\theta}_0)\boldsymbol{v}+O(\|\delta\boldsymbol{c}\|^4),$$

in which the vertical part of $\boldsymbol{w}$ does not appear. The i-th row of $J\left(\cdot\right)$ is $\nabla g_i^\top$, so $D_{\boldsymbol{v}}J$ has rows $(\nabla^2 g_i(\boldsymbol{\theta}_0)\boldsymbol{v})^\top$, $D_{\boldsymbol{v}}K=(D_{\boldsymbol{v}}J)J_0^\top+J_0(D_{\boldsymbol{v}}J)^\top$, and $\boldsymbol{a}^\top(D_{\boldsymbol{v}}K)\boldsymbol{a}=2\sum_i a_i\,\boldsymbol{v}^\top\nabla^2 g_i(\boldsymbol{\theta}_0)\boldsymbol{v}$: the displayed expansion holds for $\|p-\boldsymbol{\theta}_0\|^2$. Two points qualify. $\tau\left(\boldsymbol{c}\right)-\boldsymbol{\theta}_0=D\tau(g\left(\boldsymbol{\theta}\right))\delta\boldsymbol{c}+O(\|\delta\boldsymbol{c}\|^2)=\boldsymbol{v}+O(\|\delta\boldsymbol{c}\|^2)$ by Theorem F.2(b),(d). A point p of $g^{-1}(\boldsymbol{c})$ nearest to $\boldsymbol{\theta}_0$ exists, $g^{-1}(\boldsymbol{c})\cap\overline{B}(\boldsymbol{\theta}_0,\ell_{\boldsymbol{c}})$ being compact and containing $\tau\left(\boldsymbol{c}\right)$, and $\|p-\boldsymbol{\theta}_0\|\le\|\tau\left(\boldsymbol{c}\right)-\boldsymbol{\theta}_0\|\le\ell_{\boldsymbol{c}}<R$. At $p\in B$ the fibre is a submanifold with tangent space $\ker J\left(p\right)$, so minimality gives $p-\boldsymbol{\theta}_0=J\left(p\right)^\top\boldsymbol{b}$ with $\|\boldsymbol{b}\|\le\|p-\boldsymbol{\theta}_0\|/c$; since $J\left(p\right)=J_0+O(\|p-\boldsymbol{\theta}_0\|)$, $p-\boldsymbol{\theta}_0=J_0^\top\boldsymbol{b}+O(\|\delta\boldsymbol{c}\|^2)$ and $\delta\boldsymbol{c}=J_0(p-\boldsymbol{\theta}_0)+O(\|\delta\boldsymbol{c}\|^2)=K_0\boldsymbol{b}+O(\|\delta\boldsymbol{c}\|^2)$, so $\boldsymbol{b}=\boldsymbol{a}+O(\|\delta\boldsymbol{c}\|^2)$ and $p-\boldsymbol{\theta}_0=\boldsymbol{v}+O(\|\delta\boldsymbol{c}\|^2)$. This gives the expansions of $\|\tau\left(\boldsymbol{c}\right)-\boldsymbol{\theta}_0\|^2$ and $\mathrm{dist}^2(\boldsymbol{\theta}_0,g^{-1}(\boldsymbol{c}))$.

(ii) A uniform chord–arc estimate. Let $E\subseteq\mathbb{R}^m$ be a subspace, h a C^1 map from $\{\boldsymbol{x}\in E:\|\boldsymbol{x}\|<r_0\}$ to $E^\perp$ with $h(\boldsymbol{0})=\boldsymbol{0}$ and Dh L-Lipschitz, and M a C^1 submanifold containing its graph $\{\boldsymbol{\theta}_0+\boldsymbol{x}+h(\boldsymbol{x})\}$. For $q=\boldsymbol{\theta}_0+\boldsymbol{x}+h(\boldsymbol{x})$ the curve $\gamma(t):=\boldsymbol{\theta}_0+t\boldsymbol{x}+h(t\boldsymbol{x})$, $t\in[0,1]$, lies in M; the mean of its velocity is $q-\boldsymbol{\theta}_0$, from which the velocity differs by $\left\|\int_0^1\left(Dh(t\boldsymbol{x})-Dh(t'\boldsymbol{x})\right)\boldsymbol{x}\,dt'\right\|\le L\|\boldsymbol{x}\|^2$. By the Cauchy–Schwarz inequality, Lemma C.6, and $\|\boldsymbol{x}\|\le\|q-\boldsymbol{\theta}_0\|$ (as $h(\boldsymbol{x})\perp\boldsymbol{x}$),

$$\begin{aligned}\|q-\boldsymbol{\theta}_0\|\;\le\;d_M(\boldsymbol{\theta}_0,q)\;&\le\;\Big(\int_0^1\|\dot\gamma\|^2\,dt\Big)^{1/2}\;\le\;\left(\|q-\boldsymbol{\theta}_0\|^2+L^2\|\boldsymbol{x}\|^4\right)^{1/2}\\&\le\;\|q-\boldsymbol{\theta}_0\|\left(1+\tfrac{L^2}{2}\|q-\boldsymbol{\theta}_0\|^2\right),\end{aligned}$$

the estimate of Lemma C.8, with a constant uniform over every family of graphs with a common L.

(ii) The intrinsic energies. Let $E_0:=\mathrm{Im}\,J_0^\top$ and $Q_0:=I-P_{\boldsymbol{\theta}_0}$, the orthogonal projector onto $\ker J_0=E_0^\perp$. *Reachable manifold.* By Theorem F.2(d), $\boldsymbol{c}\mapsto P_{\boldsymbol{\theta}_0}(\tau\left(\boldsymbol{c}\right)-\boldsymbol{\theta}_0)$ has the invertible differential $J_0^+:\mathbb{R}^n\to E_0$ at $g\left(\boldsymbol{\theta}_0\right)$, so near $\boldsymbol{\theta}_0$ the surface $\mathcal{R}\left(\boldsymbol{\theta}_0\right)$ is the graph over a ball of E_0 of the C^∞ map $h_{\mathcal{R}}(\boldsymbol{x}):=Q_0(\tau\left(\boldsymbol{c}(\boldsymbol{x})\right)-\boldsymbol{\theta}_0)$, $\boldsymbol{c}(\cdot)$ the local inverse, with $h_{\mathcal{R}}(\boldsymbol{0})=\boldsymbol{0}$ and $Dh_{\mathcal{R}}$ Lipschitz on a smaller ball. *Flow manifolds.* Apply Lemma D.1 with $r=2$ to the augmented field of the proof of Theorem F.2 about $(\boldsymbol{\theta}_0,g\left(\boldsymbol{\theta}_0\right))\in\Sigma$, obtaining $\Pi(\boldsymbol{\theta},\boldsymbol{c})=(\pi_{\boldsymbol{c}}(\boldsymbol{\theta}),\boldsymbol{c})$ on a neighbourhood $U'\subseteq\hat{\mathfrak{B}}$. The map $\boldsymbol{p}\mapsto(Q_0\boldsymbol{p},g\left(\boldsymbol{p}\right))$ has the invertible differential (Q_0,J_0) at $\boldsymbol{\theta}_0$, so

it is injective on a neighbourhood W_0 of $\boldsymbol{\theta}_0$, and Q_0 is injective on $g^{-1}(\boldsymbol{c}) \cap W_0$ for every $\boldsymbol{c}$. The C^2 map $F(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}) := Q_0\big[\pi_{\boldsymbol{c}}(\boldsymbol{\theta}_0 + \boldsymbol{x} + \boldsymbol{y}) - \pi_{\boldsymbol{c}}(\boldsymbol{\theta}_0)\big]$, $\boldsymbol{x} \in E_0$, $\boldsymbol{y} \in \ker J_0$, vanishes at $(\mathbf{0}, \mathbf{0}, \boldsymbol{c})$, and $D_{\boldsymbol{y}} F(\mathbf{0}, \mathbf{0}, g(\boldsymbol{\theta}_0))$ is the identity of $\ker J_0$, because $D\Pi(\boldsymbol{\theta}_0, g(\boldsymbol{\theta}_0))$ fixes $(\boldsymbol{y}, \mathbf{0}) \in T\Sigma$ ((iv) of the lemma). The implicit function theorem gives a box $\{\|\boldsymbol{x}\| < r_1,\ \|\boldsymbol{y}\| < r_2,\ \|\boldsymbol{c} - g(\boldsymbol{\theta}_0)\| < r_3\}$ and a C^2 map h with values in $\ker J_0$ such that, in the box, $F = \mathbf{0}$ if and only if $\boldsymbol{y} = h(\boldsymbol{x}, \boldsymbol{c})$. Shrinking the box so that it lies in U' and $\pi_{\boldsymbol{c}}$ maps it into W_0, $F = \mathbf{0}$ says that $\boldsymbol{\theta}_0 + \boldsymbol{x} + \boldsymbol{y}$ and $\boldsymbol{\theta}_0$ have the same limit $\tau(\boldsymbol{c})$ under the flow with labels $\boldsymbol{c}$; so the graph of $h(\cdot, \boldsymbol{c})$ lies in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$, and on a smaller box $D_{\boldsymbol{x}} h$ is Lipschitz with a constant L independent of $\boldsymbol{c}$. For $\delta\boldsymbol{c}$ small, $\tau(\boldsymbol{c})$ lies in the box, it tends to $\boldsymbol{\theta}_0$, and it is its own limit, so it lies on that graph. The chord–arc estimate on $\mathcal{R}(\boldsymbol{\theta}_0)$ and on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$, with $q = \tau(\boldsymbol{c})$ and $\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\| \le \ell_{\boldsymbol{c}} = O(\|\delta\boldsymbol{c}\|)$, gives $d^2_{\mathcal{R}(\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0, \tau(\boldsymbol{c}))$ and $E(\boldsymbol{c}) = d_{\boldsymbol{c}}(\boldsymbol{\theta}_0, \tau(\boldsymbol{c}))^2$ equal to $\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\|^2 + O(\|\delta\boldsymbol{c}\|^4)$, whose expansion is the chord's.

(iii) Let $Y \in \mathcal{N}'$ and $\boldsymbol{c} \in C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}'$, so that $q := \tau(\boldsymbol{c}) = g|^{-1}_{\boldsymbol{\theta}_0}(\boldsymbol{c})$ (Theorem F.2(f)). By Theorem D.9(v) for the two problems, $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ and $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$ are the components containing $\boldsymbol{\theta}_0$, in the leaf topology, of $\mathcal{L}(\boldsymbol{\theta}_0) \cap \mathfrak{B}$ and $\mathcal{L}(\boldsymbol{\theta}_0) \cap \mathfrak{B}_{\boldsymbol{c}}$; let $d_{\mathcal{L}}$ be the intrinsic distance of the leaf (Corollary D.5), so that $d_{\mathcal{L}}(\boldsymbol{\theta}_0, q) \le d_{\boldsymbol{c}}(\boldsymbol{\theta}_0, q) \le \ell_{\boldsymbol{c}}$ by (i). A piecewise-C^1 path in the leaf from $\boldsymbol{\theta}_0$ of length L' lies in B, so along it the defining functions of $\mathfrak{B}$ and $\mathfrak{B}_{\boldsymbol{c}}$ are at most $(1 + C^2/c^2)L' + \ell$ and $(1 + C^2/c^2)L' + \ell_{\boldsymbol{c}}$, as in the proof of (i). Since ℓ and $\ell_{\boldsymbol{c}}$ are below $R/(2 + C^2/c^2)$, $(1 + C^2/c^2)\ell_{\boldsymbol{c}} + \max(\ell, \ell_{\boldsymbol{c}}) < R$, so a leaf path from $\boldsymbol{\theta}_0$ to q of length close enough to $d_{\mathcal{L}}(\boldsymbol{\theta}_0, q)$ lies in $\mathfrak{B} \cap \mathfrak{B}_{\boldsymbol{c}}$, and, being connected in the leaf topology and containing $\boldsymbol{\theta}_0$, in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)^{\boldsymbol{c}}$. Hence $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, q) = d_{\boldsymbol{c}}(\boldsymbol{\theta}_0, q) = d_{\mathcal{L}}(\boldsymbol{\theta}_0, q)$, that is, $E_Y(\boldsymbol{c}) = E(\boldsymbol{c})$. The kernel regime is Proposition D.15(c). □

On the model of Remark D.3 with $\boldsymbol{\theta}_0 = (0, 0, 1)$ and $\delta\boldsymbol{c} = \varepsilon(-1, 2)/\sqrt{5}$, a direction that is not an eigenvector of K_0, write q_2 and T_3 for the second- and third-order terms of the expansion. Then, at $\varepsilon = 0.025, 0.0125, 0.00625$, all within $\mathcal{N}'$, $(\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\|^2 - q_2)/\varepsilon^3$ exceeds the predicted $T_3/\varepsilon^3 = 0.179$ by $1.22 \cdot 10^{-3}$, $6.1 \cdot 10^{-4}$ and $3.0 \cdot 10^{-4}$, and $(\|\tau(\boldsymbol{c}) - \boldsymbol{\theta}_0\|^2 - q_2 - T_3)/\varepsilon^4$ lies between 0.0489 and 0.0491.

Read as a prior, $\exp(-\beta E)$ is Gaussian about $g(\boldsymbol{\theta}_0)$ with covariance $(2\beta)^{-1} K_0$ to second order, in every regime: the NTK Gaussian-process marginal is its osculating Gaussian at initialisation, as it is of the contraction of anchored ridge, which shares the expansion (Remark F.2). The first departure is cubic and odd in $\delta\boldsymbol{c}$, set by how fast the NTK moves when the network is trained towards $\boldsymbol{c}$: along $\boldsymbol{v}$ the kernel grows or shrinks, and output vectors in whose direction training enlarges the kernel become cheaper to reach. In the kernel regime the departure vanishes and the Gaussian is exact (Proposition D.15(f)); whether E extends beyond the training inputs to a process is open.

F.4 Regulariser of a problem, prior of a class

The flow manifold anchors the regulariser of one training problem; the reachable manifold is the only anchor that serves every problem of the family at once.

Theorem F.4 (Unbiasedness for every label vector forces the training map)**.** *Let $O \subseteq \mathcal{N}$ be open and $\mathcal{W}' \subseteq g^{-1}(O)$ a neighbourhood of $\tau(O)$. Call $R : \Theta \to [0, \infty]$ an* additive lift *over O if $R = A \circ g + S$ on $\mathcal{W}'$ and $R = +\infty$ off $\mathcal{W}'$, with $A : O \to [0, \infty)$ continuous, $S : \mathcal{W}' \to [0, \infty)$ continuous, and S vanishing on each fibre $g^{-1}(\boldsymbol{c}) \cap \mathcal{W}'$, $\boldsymbol{c} \in O$, at exactly one point $\sigma(\boldsymbol{c})$, with $\sigma : O \to \mathcal{W}'$ continuous. Then for every $Y \in O$ the minimisers of $\mathcal{L}(g(\cdot), Y) + \lambda R$ exist for all small $\lambda > 0$, and R is* unbiased *for every $Y \in O$ – every selection of these minimisers converges to $\tau(Y)$ as $\lambda \downarrow 0$ – if and only if $\sigma = \tau$ on O.*

Proof. Fix $Y \in O$ and $\lambda > 0$. For $\boldsymbol{\theta} \in \mathcal{W}'$ with $\boldsymbol{c} := g(\boldsymbol{\theta})$,

$$\mathcal{L}(g(\boldsymbol{\theta}), Y) + \lambda R(\boldsymbol{\theta}) \;=\; \|\boldsymbol{c} - Y\|^2 + \lambda A(\boldsymbol{c}) + \lambda S(\boldsymbol{\theta}) \;\ge\; \|\boldsymbol{c} - Y\|^2 + \lambda A(\boldsymbol{c}),$$

with equality if and only if $\boldsymbol{\theta} = \sigma(\boldsymbol{c})$; the objective is $+\infty$ off $\mathcal{W}'$, and $\sigma(\boldsymbol{c}) \in \mathcal{W}' \cap g^{-1}(\boldsymbol{c})$ for every $\boldsymbol{c} \in O$. Hence $\inf_{\boldsymbol{\theta}}[\mathcal{L} + \lambda R] = \inf_{\boldsymbol{c} \in O}[\|\boldsymbol{c} - Y\|^2 + \lambda A(\boldsymbol{c})]$, and $\boldsymbol{\theta}$ minimises the left-hand side if and only if $\boldsymbol{\theta} = \sigma(\boldsymbol{c})$ for a minimiser $\boldsymbol{c}$ of the right-hand side, as in Proposition D.17(i). Let $\overline{B}(Y, \rho_1) \subset O$. If $\lambda A(Y) < \rho_1^2$, the right-hand side is at most $\lambda A(Y)$ at $\boldsymbol{c} = Y$ and at least ρ_1^2 off $B(Y, \rho_1)$, so it is minimised over the compact ball $\overline{B}(Y, \rho_1)$, on which A is continuous: minimisers

$\boldsymbol{c}_\lambda$ exist, and every one satisfies $\|\boldsymbol{c}_\lambda - Y\|^2 \le \lambda A(Y)$. So $\boldsymbol{c}_\lambda \to Y$, and, σ being continuous, every selection of minimisers $\sigma(\boldsymbol{c}_\lambda)$ converges to $\sigma(Y)$. R is therefore unbiased for Y if and only if $\sigma(Y) = \tau(Y)$, and for every $Y \in O$ if and only if $\sigma = \tau$ on O. □

Remark F.2 (Regulariser of a problem, prior of a class). Two anchors are canonical for two purposes. Geodesic ridge is anchored on $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ through $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$, which trajectory-level minimality forces (Theorem 4.2(b)); it is the regulariser of the training problem, label-dependent as early stopping is, and it keeps the sandwich and the invariant dynamics. As a function of the parameters it is unbiased for its own labels, and, by the argument of Theorem F.4 with $A = E_Y$ and anchor $g|_{\boldsymbol{\theta}_0}^{-1}(\cdot)$, for another label vector $Y' \in C_0(\boldsymbol{\theta}_0) \cap \mathcal{N}$ only if $g|_{\boldsymbol{\theta}_0}^{-1}(Y') = \tau(Y')$, that is, only if $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ contains the network trained on Y'. Anchoring on $\mathcal{R}(\boldsymbol{\theta}_0)$ through τ gives the additive lift $R^{\mathcal{R}} := E \circ g + d^2_{\text{fibre}}(\tau(g(\cdot)), \cdot)$, over the ball $\mathcal{N}'$ on which E is continuous (Theorem F.3(i)) and on a neighbourhood of $\tau(\mathcal{N}')$ on which the fibre term is continuous, as it is near each point of $\mathcal{R}(\boldsymbol{\theta}_0)$ by the argument of Proposition D.17(ii). It is unbiased for every label vector in $\mathcal{N}'$ (Theorem F.4) and label-free, but not trajectory-minimal: at a trajectory point off $\mathcal{R}(\boldsymbol{\theta}_0)$ its fibre term is positive, so $R^{\mathcal{R}}(\phi_{\text{gf}}(\boldsymbol{\theta}_0, t)) > E(g(\phi_{\text{gf}}(\boldsymbol{\theta}_0, t)))$, and such points occur whenever training is route-dependent (Theorem F.2(e)): on the model of Remark D.3 every intermediate time tested gives one. The two anchors can agree for all label vectors of an open set O, in the sense that every $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)^Y$, $Y \in O$, contains a common open piece of $\mathcal{R}(\boldsymbol{\theta}_0)$, only if that piece is tangent to the gradient field of every $Y \in O$, hence an integral surface of $\mathcal{H}$ (Theorem D.8): outside the integrable class the regulariser of a problem and the regulariser of the class differ. In the integrable class they coincide near $\boldsymbol{\theta}^\star$ whenever $Y \in \mathcal{N}'$ (Theorem F.2(f), Theorem F.3(iii)), and both are anchored ridge in the kernel regime (Corollary D.8). E and E_Y agree at Y; whether they are tangent there in general is not claimed. Nor does the NTK match of Theorem F.3(ii) single out geodesic ridge: every label-independent energy listed there shares the expansion through third order, including $\text{dist}^2(\boldsymbol{\theta}_0, g^{-1}(\boldsymbol{c}))$, the contraction of anchored ridge under g, so near initialisation the outputs priced by the canonical regulariser and by anchored ridge agree up to $O(\|\delta\boldsymbol{c}\|^4)$. The two differ in which point of each output fibre they select, and in their energies at fourth order and globally, which we do not study.

In Bayesian terms, E_Y is the prior that the regulariser of the training problem implies on outputs, exact as a MAP statement for that problem and dependent on its labels as early stopping is, and E is the prior of the class: the cost of reaching each output vector from $\boldsymbol{\theta}_0$, fixed before the labels are seen, Gaussian with the NTK covariance at second order, and equal to the problem's own prior at the labels themselves.

G What Width Adds

The results above hold for any smooth overparametrised model under Assumption 2.1. This appendix records what the wide-network structure adds, separating theorems from conjectures. Propositions G.1–G.4 are theorems; Remark G.1 collects what is conjectural and what width does not give. Nothing in Appendices A–F depends on this appendix.

Proposition G.1 (The optimiser's metric)**.** *Let each parameter tensor k carry a width-dependent scale $s_k > 0$ and a learning-rate multiplier $m_k > 0$, and let $\boldsymbol{\theta}_k := s_k \boldsymbol{\theta}_k^{\text{st}}$ be its effective parameters, $\boldsymbol{\theta}_k^{\text{st}}$ the stored ones. Gradient descent on the stored parameters with per-tensor learning rates ηm_k, with or without heavy-ball momentum, is the same iteration on the effective parameters with per-tensor rates $\eta m_k s_k^2$. In the μP implementation of Appendix H the scale of each tensor is its μP width scaling, for which $m_k s_k^2 = 1$, m_k being the μP learning-rate multiplier of the tensor; training is then Euclidean gradient descent in effective coordinates at every width. Consequently:*

- *(i) the measurement rule of Lemma D.6, calibrated in effective coordinates, is the same at every width, and every object of the paper is stated in effective coordinates;*
- *(ii) if instead $m_k s_k^2 = d_k$ with fixed $d_k > 0$, the per-tensor $O(1)$ multipliers that the parametrisation leaves free, and D is the diagonal matrix carrying d_k on the coordinates of tensor k, gradient flow $\dot{\boldsymbol{\theta}} = -D\nabla_{\boldsymbol{\theta}}\mathcal{L}$ is the gradient flow of $\mathcal{L}$ for the metric $\langle \cdot, D^{-1}\cdot \rangle$, and every result of the paper holds with $\langle \cdot, \cdot \rangle$ replaced by $\langle \cdot, D^{-1}\cdot \rangle$, K by $JDJ^\top$, P by $DJ^\top(JDJ^\top)^{-1}J$, every distance by its D^{-1}-distance, and Assumption 2.1 imposed on the singular values of $JD^{1/2}$.*

Proof. By the chain rule $\partial\mathcal{L}/\partial\boldsymbol{\theta}_k^{\mathrm{st}} = s_k\,\partial\mathcal{L}/\partial\boldsymbol{\theta}_k$, so the stored update $\boldsymbol{\theta}_k^{\mathrm{st}} \leftarrow \boldsymbol{\theta}_k^{\mathrm{st}} - \eta m_k s_k\,\partial\mathcal{L}/\partial\boldsymbol{\theta}_k$ is the effective update $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \eta m_k s_k^2\,\partial\mathcal{L}/\partial\boldsymbol{\theta}_k$; a heavy-ball buffer $\boldsymbol{b}_k^{\mathrm{st}} \leftarrow \beta_{\mathrm{m}}\boldsymbol{b}_k^{\mathrm{st}} + \partial\mathcal{L}/\partial\boldsymbol{\theta}_k^{\mathrm{st}}$ is s_k times the buffer $\boldsymbol{b}_k \leftarrow \beta_{\mathrm{m}}\boldsymbol{b}_k + \partial\mathcal{L}/\partial\boldsymbol{\theta}_k$ of the effective iteration, with the same effect. *(i)* Lemma D.6 uses only that training is gradient flow in the Euclidean metric of the coordinates in which the rule is written. *(ii)* The linear map $\boldsymbol{\xi} \mapsto \boldsymbol{\theta} = D^{1/2}\boldsymbol{\xi}$ carries the Euclidean gradient flow $\dot{\boldsymbol{\xi}} = -\nabla_{\boldsymbol{\xi}}\mathcal{L}$ of the model $\boldsymbol{\xi} \mapsto g\left(D^{1/2}\boldsymbol{\xi}\right)$, whose Jacobian is $JD^{1/2}$ and whose NTK is $JDJ^\top$, to $\dot{\boldsymbol{\theta}} = -D\nabla_{\boldsymbol{\theta}}\mathcal{L}$, and it is an isometry from the Euclidean metric in $\boldsymbol{\xi}$ to $\langle\cdot, D^{-1}\cdot\rangle$ in $\boldsymbol{\theta}$. Every result applies in $\boldsymbol{\xi}$. Transported back, Euclidean distances in $\boldsymbol{\xi}$ become D^{-1}-distances in $\boldsymbol{\theta}$, and the orthogonal projector onto $\mathrm{Im}(JD^{1/2})^\top$ becomes $D^{1/2}\big[D^{1/2}J^\top(JDJ^\top)^{-1}JD^{1/2}\big]D^{-1/2} = DJ^\top(JDJ^\top)^{-1}J$, the D^{-1}-orthogonal projector onto $\mathrm{Im}\,DJ^\top$, the horizontal space of the preconditioned flow. □

Width fixes the metric up to these multipliers: among the parametrisations whose per-tensor scalings are powers of the width, μP is, up to the symmetry that leaves effective coordinates unchanged, the unique one whose infinite-width limit is stable and learns features in every layer (Yang & Hu, 2021). Adaptive optimisers such as Adam (Kingma & Ba, 2015) are not gradient flows of a fixed metric and lie outside the framework.

Proposition G.2 (Width-uniform constants at initialisation)**.** *Under μP for MLPs with smooth non-polynomial activations and fixed data, with biases or pairwise non-proportional inputs, $K(\boldsymbol{\theta}_0)$ converges almost surely, as the width tends to infinity, to a positive-definite limit K_∞ (Yang, 2020a, Cor. 7.3; Carvalho et al., 2025, Thms. 1–2), so that with probability tending to one $\lambda_{\min}(K(\boldsymbol{\theta}_0)) \geq \lambda_{\min}(K_\infty)/2$ and $\lambda_{\max}(K(\boldsymbol{\theta}_0)) \leq 2\lambda_{\max}(K_\infty)$. The singular values of $J(\boldsymbol{\theta}_0)$, and with them the normal spectral gap $2\sigma_{\min}(J(\boldsymbol{\theta}_0))^2 \geq \lambda_{\min}(K_\infty)$ of the equilibrium manifold Σ of the proof of Theorem F.2 at $(\boldsymbol{\theta}_0, g(\boldsymbol{\theta}_0))$, are therefore bounded away from 0 and ∞ uniformly in width, and so, on some ball about $\boldsymbol{\theta}_0$ whose radius may depend on the width, are the constants of Assumption 2.1 (Remark A.4). The radius condition is not supplied by width.*

Proof. Write N for the width and use the abc-parametrisation $W^l = N^{-a_l}w^l$ of Yang & Hu (2021, Table 1), with the same standard Gaussians for μP and the NTK parametrisation. The effective weights of the non-readout layers then have the same law in both, so the forward features coincide; the μP readout is $N^{-1/2}$ times the NTK-parametrisation one, so every backpropagated vector is scaled by $N^{-1/2}$, while the readout gradient, the last hidden feature, is unchanged. The learning-rate-weighted contribution of layer l is $N^{-2a_l}\langle\nabla_{W^l}f(\boldsymbol{x}), \nabla_{W^l}f(\boldsymbol{x}')\rangle$ (both learning-rate exponents vanish), and a_l drops by $\frac{1}{2}$ from the NTK parametrisation to μP at the input and hidden layers and is unchanged at the readout, so the factor N^{+1} cancels the backpropagated N^{-1} layer by layer and the two Gram matrices are equal at every width. The symmetry of Yang & Hu (2021, Eq. (5)) leaves the learning-rate-weighted kernel, and so $K(\boldsymbol{\theta}_0)$ of Proposition G.1, unchanged. Almost-sure convergence and positive definiteness of the limit are the cited results; under Proposition G.1(ii) the limit is the d_k-weighted sum of the per-tensor limits, positive definite because the readout term alone is (Carvalho et al., 2025, Cor. 1). Almost-sure convergence implies convergence in probability, and eigenvalues are continuous functions of a symmetric matrix, which gives the two bounds; a finite-width lower bound for the NTK parametrisation, such as that of Banerjee et al. (2023, Thm. 4.1), transfers verbatim through the identity. $\sigma_i(J(\boldsymbol{\theta}_0))^2 = \lambda_i(K(\boldsymbol{\theta}_0))$, and the normal spectrum of Σ at $(\boldsymbol{\theta}_0, g(\boldsymbol{\theta}_0))$ is $\{-2\lambda_i(K(\boldsymbol{\theta}_0))\}$ (proof of Theorem F.2(b)). Continuity of $\sigma_{\min}(J(\cdot))$ and $\sigma_{\max}(J(\cdot))$ extends the bounds, halved and doubled, to a ball about $\boldsymbol{\theta}_0$ whose radius nothing in the argument controls; in particular nothing bounds it below by ℓ. □

Proposition G.3 (Permutation equivariance)**.** *Let $\mathcal{G}$ be a group of linear isometries of effective-parameter space with $g \circ \sigma = g$ and $\sigma\Theta = \Theta$ for every $\sigma \in \mathcal{G}$; for a network, the permutations of the hidden units of each layer, which permute coordinates within tensors and so preserve the per-tensor scales. For $\sigma \in \mathcal{G}$, Assumption 2.1 holds at $\sigma\boldsymbol{\theta}_0$ with the same radius and constants, and: $X \circ \sigma = \sigma X$; $\phi_t \circ \sigma = \sigma \circ \phi_t$; $\pi(\sigma\boldsymbol{\theta}) = \sigma\pi(\boldsymbol{\theta})$; $\sigma\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) = \mathcal{M}_{\mathrm{flow}}(\sigma\boldsymbol{\theta}_0)$, the flow manifold of $\sigma\boldsymbol{\theta}_0$ for the same problem; $\sigma\mathcal{R}(\boldsymbol{\theta}_0) = \mathcal{R}(\sigma\boldsymbol{\theta}_0)$; $d_{\mathrm{flow}}(\sigma\boldsymbol{\theta}, \sigma\boldsymbol{\theta}') = d_{\mathrm{flow}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, the left-hand side computed on $\mathcal{M}_{\mathrm{flow}}(\sigma\boldsymbol{\theta}_0)$; the energy of the training problem and the training-cost function of $\sigma\boldsymbol{\theta}_0$ are those of $\boldsymbol{\theta}_0$; and arc ridge is invariant, $L(\sigma \circ \gamma) = L(\gamma)$ and $\boldsymbol{F}_{\mathrm{arc}}(\sigma\boldsymbol{\theta}, s) = \sigma\boldsymbol{F}_{\mathrm{arc}}(\boldsymbol{\theta}, s)$.*

Proof. Differentiating $g \circ \sigma = g$ gives $J(\sigma\boldsymbol{\theta})\,\sigma = J(\boldsymbol{\theta})$, so $J(\sigma\boldsymbol{\theta}) = J(\boldsymbol{\theta})\,\sigma^\top$ has the singular values of $J(\boldsymbol{\theta})$; with $\sigma B(\boldsymbol{\theta}_0, R) = B(\sigma\boldsymbol{\theta}_0, R)$ and $g(\sigma\boldsymbol{\theta}_0) = g(\boldsymbol{\theta}_0)$, Assumption 2.1 transfers, $\mathcal{N}$ is unchanged, and $\sigma\mathfrak{B}$ is the trapping region of $\sigma\boldsymbol{\theta}_0$. Then $X(\sigma\boldsymbol{\theta}) = -2\sigma J(\boldsymbol{\theta})^\top (g(\boldsymbol{\theta}) - Y) = \sigma X(\boldsymbol{\theta})$, so $t \mapsto \sigma\phi_t(\boldsymbol{\theta})$ solves the flow equation from $\sigma\boldsymbol{\theta}$ and equals $\phi_t(\sigma\boldsymbol{\theta})$ by uniqueness, and limits commute with the continuous σ. Hence $\sigma\pi^{-1}(\boldsymbol{\theta}^\star)$ is the set of points of $\sigma\mathfrak{B}$ training to $\sigma\boldsymbol{\theta}^\star$, the trained network from $\sigma\boldsymbol{\theta}_0$: $\sigma\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0) = \mathcal{M}_{\text{flow}}(\sigma\boldsymbol{\theta}_0)$. The same argument for each label vector in $\mathcal{N}$ gives $\sigma \circ \tau$ as the training map of $\sigma\boldsymbol{\theta}_0$ and $\sigma\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)^c$ as its flow manifolds. An isometry of $\mathbb{R}^m$ carries a submanifold onto its image isometrically for the induced metrics and preserves the lengths of paths and the outputs g, which gives the statements on distances, on E_Y and E (Proposition F.1, Definition F.1), and on L; finally $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\sigma\boldsymbol{\theta}) = \sigma\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ and σ preserves norms, so $\boldsymbol{F}_{\text{arc}}(\sigma\boldsymbol{\theta}, s) = \sigma\boldsymbol{F}_{\text{arc}}(\boldsymbol{\theta}, s)$ (Definition 4.2). □

For two-layer networks the quotient by permutations is a space of measures, and the optimiser's metric is the Wasserstein metric on it.

Proposition G.4 (Two-layer networks: the optimiser's metric is Wasserstein)**.** *Let $f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=1}^{N} a_j\varphi(\boldsymbol{w}_j \cdot \boldsymbol{x})$ with neurons $z_j := (a_j, \boldsymbol{w}_j) \in \mathbb{R}^{d+1}$ and φ smooth, in μP, which at depth two is the mean-field parametrisation (Yang & Hu, 2021; Chizat & Bach, 2018; Mei et al., 2018): gradient flow moves the neurons by $\dot{z}_j = -N\nabla_{z_j}\mathcal{L}$, and in the effective coordinates $\tilde{z}_j := z_j/\sqrt{N}$ it is Euclidean. Write $\mu_{\boldsymbol{\theta}} := \frac{1}{N}\sum_j \delta_{z_j}$ for the neuron measure, $\mathcal{L}(\mu) := \sum_i \big(\int a\varphi(\boldsymbol{w}\cdot X_i)\,d\mu(z) - Y_i\big)^2$, so that $\mathcal{L}(\mu_{\boldsymbol{\theta}}) = \mathcal{L}(g(\boldsymbol{\theta}), Y)$, and S_N for the permutations of the neurons. Then:*

- *(i)* $\min_{\sigma\in S_N}\left\|\tilde{\boldsymbol{\theta}} - \sigma\tilde{\boldsymbol{\theta}}'\right\| = W_2(\mu_{\boldsymbol{\theta}}, \mu_{\boldsymbol{\theta}'})$*: the Euclidean distance of effective coordinates descends to the quotient by S_N as the Wasserstein distance of the neuron measures;*
- *(ii)* $W_2^2(\mu_{\boldsymbol{\theta}}, \mu_{\boldsymbol{\theta}_0}) \le \left\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_0\right\|^2$*: anchored ridge dominates the squared Wasserstein distance from the initial neuron measure;*
- *(iii) the gradient-flow velocity of each neuron is the value at it of one gradient field, $\dot{z}_j = v_t(z_j)$ with $v_t := -\nabla_z \frac{\delta\mathcal{L}}{\delta\mu}(\cdot\,;\mu_t)$ and $\mu_t := \mu_{\boldsymbol{\theta}(t)}$; if the initial neurons are distinct, $|\dot{\mu}_t|_{W_2} = \|v_t\|_{L^2(\mu_t)} = \left\|\dot{\tilde{\boldsymbol{\theta}}}(t)\right\|$, and $s(T)$ is the W_2-length of $t \mapsto \mu_t$ on $[0,T]$: arc ridge is the squared Wasserstein length of the path of the neuron measure, at every finite width;*
- *(iv) S_N permutes flow manifolds (Proposition G.3), and if the neurons of $\boldsymbol{\theta}^\star$ are distinct, the quotient map restricted to $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ is injective and preserves lengths of paths: $\mathcal{M}_{\text{flow}}(\boldsymbol{\theta}_0)$ with its intrinsic distance is isometric to its image in the space of neuron measures with the length metric of W_2.*

Proof. (i) For $\sigma \in S_N$, $\left\|\tilde{\boldsymbol{\theta}} - \sigma\tilde{\boldsymbol{\theta}'}\right\|^2 = \frac{1}{N}\sum_j \left\|z_j - z'_{\sigma(j)}\right\|^2$ is the cost of the coupling of $\mu_{\boldsymbol{\theta}}$ and $\mu_{\boldsymbol{\theta}'}$ that sends z_j to $z'_{\sigma(j)}$. Every coupling of two uniform measures on N atoms, counted with multiplicity, is $\frac{1}{N}$ times a doubly stochastic matrix, the cost is linear in that matrix, and the extreme points of the doubly stochastic matrices are the permutation matrices (Birkhoff–von Neumann), so the minimum over couplings, W_2^2, is attained at a permutation. *(ii)* The identity is a permutation.

(iii) $\mathcal{L}(\mu_{\boldsymbol{\theta}}) = \sum_i (f(X_i;\boldsymbol{\theta}) - Y_i)^2$ with $\nabla_{z_j} f(X_i;\boldsymbol{\theta}) = \frac{1}{N}\nabla_z[a\varphi(\boldsymbol{w}\cdot X_i)]_{z=z_j}$, so $N\nabla_{z_j}\mathcal{L} = \nabla_z \frac{\delta\mathcal{L}}{\delta\mu}(z_j;\mu_{\boldsymbol{\theta}})$ with $\frac{\delta\mathcal{L}}{\delta\mu}(z;\mu) = 2\sum_i \big(\int a'\varphi(\boldsymbol{w}'\cdot X_i)\,d\mu(z') - Y_i\big)a\varphi(\boldsymbol{w}\cdot X_i)$, and $\dot{z}_j = v_t(z_j)$. The field v_t is locally Lipschitz in z, so the neurons solve one non-autonomous equation $\dot{z} = v_t(z)$ and two of them that coincide at one time coincide at all times: distinct initial neurons stay distinct. Fix t, let $\delta > 0$ be the least distance between neurons at times near t and M a bound on their speeds there. For times t_1, t_2 near t, a coupling of μ_{t_1} and μ_{t_2} by a permutation other than the identity sends some $z_j(t_1)$ to $z_k(t_2)$ with $k \ne j$, at cost at least $(\delta - M|t_2 - t_1|)^2/N$, while the identity costs at most $M^2|t_2 - t_1|^2$; so for $|t_2 - t_1|$ small the identity is optimal, $W_2(\mu_{t_1}, \mu_{t_2}) = \left\|\tilde{\boldsymbol{\theta}}(t_2) - \tilde{\boldsymbol{\theta}}(t_1)\right\|$, and the metric derivative is $\left\|\dot{\tilde{\boldsymbol{\theta}}}(t)\right\| = (\frac{1}{N}\sum_j \|v_t(z_j)\|^2)^{1/2} = \|v_t\|_{L^2(\mu_t)}$, the finite-width case of the characterisation of the metric derivative by the tangent velocity field (Ambrosio et al., 2008, Ch. 8). Computing lengths over fine partitions, the W_2-length of $t \mapsto \mu_t$ on $[0,T]$ equals the Euclidean length of $t \mapsto \tilde{\boldsymbol{\theta}}(t)$, which is $s(T)$.

(iv) The first statement is Proposition G.3. If $\sigma \neq \mathrm{id}$ and $\boldsymbol{\theta} \in \mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0) \cap \sigma\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, the trajectory from $\boldsymbol{\theta}$ converges both to $\boldsymbol{\theta}^\star$ and to $\sigma\boldsymbol{\theta}^\star \neq \boldsymbol{\theta}^\star$, which is impossible; so the flow manifolds $\sigma\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$, $\sigma \in S_N$, are pairwise disjoint. Hence no point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ is fixed by a transposition, that is, every point of it has distinct neurons, and the quotient map q is injective on $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. Near a configuration with distinct neurons the identity is the optimal coupling between nearby configurations, as in (iii), so q preserves distances between nearby points there and therefore the lengths of paths in $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$. Its inverse on $q(\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0))$ is continuous: the training-limit map is continuous and $\sigma\boldsymbol{\theta}^\star \neq \boldsymbol{\theta}^\star$ for $\sigma \neq \mathrm{id}$, so a neighbourhood of a point of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ meets no $\sigma\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ with $\sigma \neq \mathrm{id}$, and a configuration of $\mathcal{M}_{\mathrm{flow}}(\boldsymbol{\theta}_0)$ whose neuron measure is W_2-close to that of $\boldsymbol{\theta}$ is, up to a permutation that must be the identity, Euclidean-close to $\boldsymbol{\theta}$. Paths therefore correspond with equal lengths in both directions, and the two length metrics agree. □

For depth three or more the symmetry group is a product of permutation groups acting layer by layer, and the quotient distance $\min_\sigma \left\|\tilde{\boldsymbol{\theta}} - \sigma\tilde{\boldsymbol{\theta}}'\right\|$ is an assignment problem coupled across layers, not the Wasserstein distance of a single neuron measure; (i)–(iii) and the isometry of (iv) are two-layer statements.

Remark G.1 (What is conjectural, and what width does not give). *Conjecture (Tensor-Programs limits).* For a μP network with smooth polynomially bounded activations, fixed data and a fixed horizon T, the arc length $s(T)$, the chord $\|\boldsymbol{\theta}(T) - \boldsymbol{\theta}_0\|$, the NTK $K(\boldsymbol{\theta}(t))$, the cross-time Gram matrices $J(\boldsymbol{\theta}(t))\,J(\boldsymbol{\theta}(t'))^\top$ for $t, t' \leq T$, the symmetric and antisymmetric parts of the tensor B of Remark A.1 along the trajectory, and the second fundamental form $\mathrm{II}_{\boldsymbol{\theta}^\star}$ of Proposition D.7 have deterministic limits as the width $N \to \infty$: each is a continuous function of coordinate-averaged quantities of the training process of the kind the Tensor-Programs master theorem controls (Yang & Hu, 2021), although the expressibility of the cross-time and second-order quantities as Tensor-Program variables is not written out here. Passing to $T = \infty$ needs the constants of Assumption 2.1 uniform in width along the trajectory, which is the open radius condition; once they are, Corollary D.2 controls the tail uniformly. *Open.* A deterministic limit of $d_{\mathrm{flow}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star)$ or of E: both are infima over paths, boundary-value objects that are not evidently Tensor-Program quantities. *A conditional statement.* By Proposition D.7, with $\|\mathcal{B}(\boldsymbol{v}, \boldsymbol{w})\| \leq \|\boldsymbol{a}\|\, M_2(\boldsymbol{\theta}^\star)\, \|\boldsymbol{w}\| \leq M_2(\boldsymbol{\theta}^\star)\, \|\boldsymbol{v}\|\, \|\boldsymbol{w}\|\,/c$ for $\boldsymbol{v} = J(\boldsymbol{\theta}^\star)^\top \boldsymbol{a}$ and $M_2(\boldsymbol{\theta}^\star) := (\sum_k \left\|\nabla^2 g_k(\boldsymbol{\theta}^\star)\right\|_{\mathrm{op}}^2)^{1/2}$, every entry of $\mathrm{II}_{\boldsymbol{\theta}^\star}$ in the K-eigenframe of that proposition has norm at most $2M_2(\boldsymbol{\theta}^\star)/c$; under the NTK parametrisation, conditional on the Hessian scaling $M_2 = O(N^{-1/2})$ of that parametrisation (Liu et al., 2020; 2022) and on c bounded below, $\mathrm{II}_{\boldsymbol{\theta}^\star} = O(N^{-1/2})$ and the flow manifold flattens to the plane of the kernel regime. *What width does not give.* Integrability of $\mathcal{H}$. *Conjecture (the twist is a law of large numbers):* for two-layer μP networks the antisymmetric part of B at initialisation, which in a K-orthonormal frame is a continuous function of averages over the independent neurons, converges to a deterministic limit that is non-zero for generic data, so that width makes non-integrability deterministic, not small, as the finite examples of Remark A.3(iii)–(iv) suggest. Nor does width give the radius condition of Assumption 2.1 (Proposition G.2), embeddedness or completeness of the flow manifold beyond Theorem D.3 (the chart of Theorem D.4 is local), a Riemannian metric on output space whose energy is E (the metrics G and $G_{\mathcal{R}}$ are those of particular charts, and G depends on the labels), or anything for adaptive optimisers.

H Experimental Details

We discuss here the main details surrounding the two large-scale experiments in Section 5. Our codebase, which is predominantly based on the PyTorch (Paszke et al., 2019) and `mup` (Yang et al., 2021) libraries, is shared at `https://anonymous.4open.science/r/canonical-regularisation-7349/`. All regularisers act on effective parameters, the stored parameters multiplied by their μP width scalings. The implementation weights $\frac{1}{2}R$ rather than R for every regulariser, so that the objective minimised at a reported strength λ is $\mathcal{L} + \frac{\lambda}{2}R$: reported values of λ are twice those of the convention $\mathcal{L} + \lambda R$ of this paper.

H.1 UTKFACE

For this experiment, we adapt the ResNet18 architecture (He et al., 2016) by replacing the classification head with a μP output layer (`MuReadout`) as regression head, using the default PyTorch `Linear` initialisation (Kaiming-uniform). We fine-tune all weights from the pretrained checkpoint, with per-layer learning-rate scaling for μP. Anchored ridge anchors every parameter, the freshly initialised head included, at its initial value, and so differs from L2-SP, which applies standard ridge to the head. We train each regulariser setup over three seeds for 50 epochs with a batch size of 128, base learning rate of 0.003 (decreased to 0.0003 for the largest regularisation strength due to instability; early stopping prevents non-regulariser-induced undertraining here nonetheless), and momentum of 0.9 (to account for stochasticity induced by minibatching). UTKFace (Zhang et al., 2017) defines no evaluation split; we report test performance on the held-out `valid` split of the `deedax/UTK-Face-Revised` Hugging Face release (deedax, 2023), and use an 80–20 validation split of its training portion for early stopping. We further corrupt targets with a per-datapoint fixed (for a given training run) additive isotropic zero-mean Gaussian noise with a standard deviation of 0.3 years.

H.2 YELP REVIEW

For this experiment, we use the DistilBERT (Sanh et al., 2019) base encoder architecture with the same μP regression head and initialisation, and again fine-tune all weights from the pretrained checkpoint with per-layer learning-rate scaling. We train each regulariser setup over two seeds (due to the runtime) for a number of epochs determined by a patience mechanism: we train for a base 20 epochs, then, whenever the final epoch exhibits the best validation loss, for a further three epochs, the runtime preventing substantial overtraining. We train with a base learning rate of 0.003 (decreased to 0.0001 for the largest regularisation strength) and momentum of 0.9. We process the training dataset by first transforming it to a valid regression target, normalising the scores (1 to 5 stars) to the interval $[0, 1]$ by an affine transformation and transforming to the reals by a logit transform (clipping scores at 0.01 and 0.99 to prevent infinite targets), then corrupting with a per-datapoint fixed (for a given training run) additive isotropic zero-mean Gaussian noise with a standard deviation of 0.1 (unitless).

H.3 COMPUTE

Both experiments were carried out remotely on NVIDIA A10 GPUs via Modal. The UTKFace experiments run relatively quickly at roughly 10 minutes per regulariser per seed, while the Yelp Review experiment is considerably slower at roughly 5 hours per regulariser per seed. Total project compute spend was roughly $315 up to the point of submission.