ON FEATURE LEARNING IN FLOW MATCHING UNDER DIFFERENT COUPLINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching learns velocity fields from paired noise and data samples without simulating trajectories during training. We discuss feature learning of flow matching, focusing on the difference between the independent and optimal-transport (OT) couplings for the same source and target distributions. To understand the difference, we employ a theoretical model, where the data lie on an unknown low-dimensional subspace, with a non-Gaussian distribution along one unknown direction (feature) and Gaussian variation in the orthogonal directions. We show that, near the target endpoint, the nonlinear component remains nonzero under OT but vanishes under independent coupling, while both couplings have the same velocity component normal to the subspace. To examine how this difference changes feature learning, we analyze joint learning of the support and feature direction at a fixed flow time using a model with one cubic neuron. Although nonlinear feature learning becomes increasingly slow near the target endpoint under independent coupling, OT avoids this slowdown, presenting more eminent feature learning. We also empirically confirm this advantage of OT in joint training across flow times, using narrow windows near the endpoint or separate features for each time bin. These results suggest that OT coupling may benefit the learning of nonlinear structure in the target density near the endpoint.

1 INTRODUCTION

Conditional flow matching (CFM) learns velocities along paths joining noise and data samples, when applied to generative tasks. For linear interpolation, the model takes their weighted average as input and predicts the velocity from noise to data (Lipman et al., 2023; Tong et al., 2024). The coupling determines sample pairings and hence the distribution of training inputs and velocity labels. *Independent coupling* refers to the independent pairing of noise and data. *Optimal-transport (OT) coupling* minimizes their average cost, typically the squared distance. Even with fixed source and target distributions, these choices can change the learned velocity field, label variance, and training speed (Tong et al., 2024; Pooladian et al., 2023).

When high-dimensional data concentrate near a set of much lower intrinsic dimension (Fefferman et al., 2016; Loaiza-Ganem et al., 2024), coupling may change how quickly the model learns where data lie and how they are distributed there (Li et al., 2026; Liu et al., 2025). Recent theory shows that these two tasks can behave differently. Near the target endpoint, OT-CFM reduces deviations normal (perpendicular) to the target support (the set on which the data lie) (Fukumizu et al., 2026), while diffusion training can separate recovery of the manifold projection from learning the density on the manifold (Huang et al., 2026).

This difference suggests that support recovery and learning nonlinear features within the support may proceed at different rates. Then, which of these two tasks is affected by the choice of coupling?

To study this question, we adapt the single-index framework from feature learning theory to flow matching. In regression, $y = \sigma_*(\langle \boldsymbol{b}_*, \boldsymbol{x} \rangle)$ with nonlinear scalar link σ_*, and feature learning aims to recovers the unknown direction $\boldsymbol{b}_*$ (Damian et al., 2022; 2023). In our setting, data lie on an unknown linear subspace, with a non-Gaussian component along one unknown direction within it. Feature learning recovers this direction, while support recovery identifies the subspace. For example, if data lie on a plane with non-Gaussian distribution in one direction and Gaussian variation in the

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.

orthogonal direction, the feature is the non-Gaussian direction and the support is the plane. We compare how quickly the support and feature direction are learned under each coupling.

To compare these learning speeds, we analyze training at a fixed flow time $t \in (0, 1)$, with noise at $t = 0$ and data at $t = 1$. The ratio of noise to data weights, $\varepsilon_t := (1 - t)/t$, is small near the data endpoint. When the nonlinear velocity component weakens as $t \to 1$, fitting the nonlinear feature would not reduce loss beyond a linear predictor. In our model, the weaker signal also reduces updates to the feature direction, slowing its learning even when support recovery is fast. Our main contributions are summarized as follows:

- For polynomial nonlinearity, we derive the velocity field to first order in the nonlinearity strength (Proposition 1). Both couplings have the same component normal to the support. Within it, the nonlinear component remains nonzero under OT but decreases in proportion to ε_t under independent coupling near the endpoint.
- With the target distribution fixed, we quantify the nonlinear effects by removing the best linear fit from the velocity field (Theorem 1). Under some assumptions, the remaining component has a positive limiting mean squared norm under OT but scales as ε_t^2 under independent coupling. Given the interpolated input, velocity labels have zero conditional variance under OT but have nonzero variance along the feature direction under independent coupling.
- We analyze the time interval that achieves a fixed fraction of the available nonlinear loss reduction. We use a model with one cubic neuron that trains the subspace jointly under continuous-time gradient descent (Theorem 2). With high probability over random initialization, this time interval remains bounded near the endpoint for OT but scales as ε_t^{-2} for independent coupling. Although our theory fixes t, experiments show that independent coupling also learns endpoint features more slowly when training uses a narrow window near the endpoint or separate features for different time bins (Figure 2).

2 Related Work

Flow matching and coupling. Different vector fields can connect the same source and target distributions. In CFM, the coupling selects the interpolation paths and regression problem (Lipman et al., 2023; Tong et al., 2024). Compared with independent coupling, OT coupling can reduce label variance and produce straighter paths. In practice, OT-CFM typically constructs noise–data pairings by solving OT problems within minibatches (Tong et al., 2024; Pooladian et al., 2023). Fukumizu et al. (2026) formulate population OT-CFM using a convex transport potential and an associated minimization problem. They prove subsequential convergence of minibatch OT vector fields to the population field as batch size grows. Gao & Li (2026) study how flow matching memorizes and generalizes when data lie in sample subspaces. We compare OT and independent coupling as regression problems, studying how they change the population velocity, the nonlinear feature signal, and the time needed to learn that feature.

Generative models on low-dimensional supports. The manifold hypothesis states that many high-dimensional data distributions concentrate on or near sets of much smaller intrinsic dimension (Fefferman et al., 2016; Loaiza-Ganem et al., 2024). Theory for diffusion models explains several consequences of this structure (Oko et al., 2023; Chandramoorthy & de Clercq, 2025). Chen et al. (2023) and Chen et al. (2024) study low-dimensional subspaces and tangent–normal structure, while Li et al. (2026) and Liu et al. (2025) study the geometry and singular behavior of the score (the gradient of the log density) near the data support. Score-based models can detect manifold support (Pidstrigach, 2022), and gradient descent on a diffusion objective can recover low-dimensional structure through subspace clustering (Wang et al., 2025). Separating support and density learning, Huang et al. (2026) identify a collapse-and-refine mechanism in which small-noise training learns projection onto the manifold and moderate-noise training learns its intrinsic density. For manifold-supported targets, Fukumizu et al. (2026) show that the terminal dynamics of OT-CFM contract normal perturbations while remaining neutral in tangent directions after time rescaling. This distinction motivates our comparison of how coupling changes support and feature learning.

Feature learning. Feature learning adapts a model's representation to the target and can form target-aligned nonlinear features beyond fixed random features (Ba et al., 2022; Moniri et al., 2024; Yang & Hu, 2021). In contrast, a model in the lazy regime remains close to its linearization at initialization and behaves like a kernel method (Jacot et al., 2018; Chizat et al., 2019). Related work studies recovery of unknown directions using Hermite polynomials, an orthogonal basis for functions of Gaussian inputs (Damian et al., 2022; 2023). The lowest degree carrying information about the direction helps determine learning difficulty. We use this decomposition to compare velocities under the two couplings. Feature learning and representation dynamics have also been studied in diffusion models through low-dimensional data models (Han et al., 2025; Li et al., 2025).

3 Feature Learning under OT and Independent Couplings

Both couplings move samples toward the same support, but they provide different signals for learning the nonlinear feature within it. We first compare these signals, then determine how they govern training dynamics.

3.1 Problem setup

Conditional flow matching. The symbols $\|\cdot\|$ and $\|\cdot\|_F$ denote the Euclidean and Frobenius norms, respectively. Let P_0 and P_1 be the source and target distributions on $\mathbb{R}^D$, where D is the ambient dimension. A coupling π is a joint distribution of $(\boldsymbol{X}_0, \boldsymbol{X}_1)$ with these marginals. It specifies how source and target samples are paired without changing either distribution.

For a flow time $t \in [0,1]$, an endpoint pair $(\boldsymbol{X}_0, \boldsymbol{X}_1) \sim \pi$, and differentiable scalar schedules α_t, β_t, define the conditional path and velocity by $\boldsymbol{X}_t = \alpha_t \boldsymbol{X}_0 + \beta_t \boldsymbol{X}_1$ and $\boldsymbol{U}_t = \dot{\alpha}_t \boldsymbol{X}_0 + \dot{\beta}_t \boldsymbol{X}_1$. We assume $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$ (Lipman et al., 2023; Tong et al., 2024). A dot denotes differentiation with respect to t. Let μ denote the distribution used to sample flow times on $[0,1]$, independently of the endpoint pair. A training example is generated as follows:

$$(\boldsymbol{X}_0, \boldsymbol{X}_1) \sim \pi, \qquad t \sim \mu, \qquad \text{input } (\boldsymbol{X}_t, t), \qquad \text{label } \boldsymbol{U}_t. \tag{1}$$

Our main results use the affine schedule $\alpha_t = 1 - t$ and $\beta_t = t$, so $\boldsymbol{X}_t = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1$ and $\boldsymbol{U}_t = \boldsymbol{X}_1 - \boldsymbol{X}_0$. At a fixed time t, the squared-loss population optimum is the conditional mean (Lipman et al., 2023):

$$\boldsymbol{v}_{t,\pi}(\boldsymbol{x}) := \mathbb{E}_\pi[\boldsymbol{U}_t \mid \boldsymbol{X}_t = \boldsymbol{x}]. \tag{2}$$

We call $\boldsymbol{v}_{t,\pi}$ the population velocity. This conditional mean depends on how π pairs source and target samples, so the same marginals can yield different velocity fields.

To study this dependence, we compare two couplings. The independent coupling is $\pi_{\mathrm{ind}} = P_0 \otimes P_1$, where $\otimes$ denotes the product of probability measures. The OT coupling minimizes the transport cost over the set $\Pi(P_0, P_1)$ of joint distributions with marginals P_0, P_1, using the average squared distance between paired points (Peyré & Cuturi, 2019):

$$\pi_{\mathrm{OT}} \in \operatorname*{arg\,min}_{\pi \in \Pi(P_0,P_1)} \mathbb{E}_\pi\left[\tfrac{1}{2}\|\boldsymbol{X}_1 - \boldsymbol{X}_0\|^2\right], \tag{3}$$

Here the expectation is over pairs $(\boldsymbol{X}_0, \boldsymbol{X}_1) \sim \pi$. We use this exact population OT coupling. For $C \in \{\mathrm{OT}, \mathrm{ind}\}$, write $\boldsymbol{v}_{t,C} := \boldsymbol{v}_{t,\pi_C}$.

Target model. We nonlinearly transform one coordinate of a Gaussian distribution on a fixed subspace to distinguish support recovery from feature learning.

Let $r \le D$ be the support dimension, and let $\boldsymbol{A} \in \mathbb{R}^{D \times r}$ have orthonormal columns spanning the target subspace $\mathcal{M}$. We use *support learning* to mean recovery of this linear subspace. The topological support of the target can be a proper subset of $\mathcal{M}$ under the assumptions below. Write $\boldsymbol{I}_k$ for the $k \times k$ identity matrix and let $\boldsymbol{A}_\perp$ be an orthonormal basis of the orthogonal complement. For independent $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_r)$ and $\boldsymbol{\Xi} \sim \mathcal{N}(0, \boldsymbol{I}_{D-r})$, the source is $\boldsymbol{X}_0 = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{A}_\perp \boldsymbol{\Xi}$, with law $P_0 = \mathcal{N}(0, \boldsymbol{I}_D)$.

Let $\boldsymbol{u} \in \mathbb{R}^r$ be a unit vector specifying the feature direction, so its ambient direction is $\boldsymbol{A}\boldsymbol{u}$. A continuously differentiable link function $H : \mathbb{R} \to \mathbb{R}$ specifies the transformation and $\rho > 0$ its

strength. The target distribution P_1 is the law of

$$\boldsymbol{X}_1 = \boldsymbol{A}\boldsymbol{Z} + \rho H(\boldsymbol{u}^\top \boldsymbol{Z})\boldsymbol{A}\boldsymbol{u}. \tag{4}$$

equation 4 specifies the target marginal, not a coupling. Under OT, its $\boldsymbol{Z}$ is shared with the source representation, whereas under independent coupling, it is an independent copy of the source Gaussian coordinate. Thus $\boldsymbol{A}$ determines the subspace containing the samples, while $\boldsymbol{u}$, H, and ρ determine their distribution within that subspace. Only the coordinate along $\boldsymbol{A}\boldsymbol{u}$ is changed. Unlike single-index regression, where the link specifies a label (Damian et al., 2022; 2023), here it specifies a data distribution. The regression target is then determined by the coupling.

The following conditions remove constant and linear contributions from H and make the transformation monotone. They give an explicit OT map and separate the nonlinear feature from a first-order covariance change.

Assumption 3.1 (Conditions on the nonlinear feature)**.** For $S \sim \mathcal{N}(0,1)$, the link H satisfies

$$\mathbb{E}[H(S)] = 0, \qquad \mathbb{E}[SH(S)] = 0, \qquad 0 < \mathbb{E}[H(S)^2] < \infty. \tag{5}$$

We also assume that $s \mapsto s + \rho H(s)$ is nondecreasing.

These conditions need not give a target density. Section 3.3 states the additional regularity needed for the independent endpoint rate.

Since the source is Gaussian and the target has finite second moment, the optimal coupling is induced by a transport map $\boldsymbol{X}_1 = T(\boldsymbol{X}_0)$, even when the target is supported on a lower-dimensional subspace (Brenier, 1991). The following lemma gives this map explicitly and shows that the transformation changes the covariance only at second order in ρ.

Lemma 3.2 (Support, covariance, and OT map)**.** *Under Assumption 3.1, P_1 is supported on $\mathcal{M}$ and*

$$\mathrm{Cov}(\boldsymbol{A}^\top \boldsymbol{X}_1) = \boldsymbol{I}_r + \rho^2 \mathbb{E}[H(S)^2]\boldsymbol{u}\boldsymbol{u}^\top. \tag{6}$$

The population quadratic-cost OT map from P_0 to P_1 is

$$T_\rho(\boldsymbol{g}) = \boldsymbol{A}\left[\boldsymbol{A}^\top \boldsymbol{g} + \rho H(\boldsymbol{u}^\top \boldsymbol{A}^\top \boldsymbol{g})\boldsymbol{u}\right]. \tag{7}$$

The proof is in Appendix A.1. The leading eigenvector of the target covariance identifies $\boldsymbol{A}\boldsymbol{u}$ up to sign. Information sufficient for direction recovery therefore exists in the data under both couplings. Our question is how quickly velocity regression uses this information. Identifiability alone does not imply fast recovery by the specified learning dynamics. Here $T = T_\rho$ induces π_{OT}. We write $\boldsymbol{X}_{t,C}$ for the path variable and $\boldsymbol{U}_{t,C}$ for its velocity label under coupling $C \in \{\mathrm{OT}, \mathrm{ind}\}$.

3.2 How coupling changes the population velocity

At a fixed flow time t, we expand the population velocity to first order in the transformation strength ρ. We then examine the endpoint limit $\varepsilon_t = (1-t)/t \to 0$. The first limit changes the target, while the second moves the input toward that target.

We expand H in Hermite polynomials because conditional averaging multiplies each degree by its own coefficient. Let He_j denote the probabilists' Hermite polynomial, normalized so that $\mathbb{E}[\mathrm{He}_j(S)\,\mathrm{He}_k(S)] = j!\,\mathbf{1}_{\{j=k\}}$ for $S \sim \mathcal{N}(0,1)$, and let J be the polynomial degree and h_j the expansion coefficients. In this subsection, we assume an exact finite expansion:

$$H(s) = \sum_{j=2}^{J} h_j\,\mathrm{He}_j(s). \tag{8}$$

Hermite polynomials form an orthogonal basis under a Gaussian input. The sum starts at degree two because the moment conditions in Assumption 3.1 remove the constant and linear components. For $c \in [0,1]$, define two linear operators on this basis:

$$\mathcal{N}\,\mathrm{He}_j = j\,\mathrm{He}_j, \qquad \mathcal{P}_c\,\mathrm{He}_j = c^j\,\mathrm{He}_j\,. \tag{9}$$

The operator $\mathcal{N}$ records the degree of each component, while $\mathcal{P}_c$ multiplies a degree-j component by c^j. These multipliers determine the first-order nonlinear velocity. We use $\|\boldsymbol{Y}\|_{L^2} = (\mathbb{E}\|\boldsymbol{Y}\|^2)^{1/2}$

for a random vector $\boldsymbol{Y}$. The notation $O_{t,L^2}(\rho^2)$ denotes a remainder with L^2 norm bounded by a constant times ρ^2, with the constant allowed to depend on t.

Let $\boldsymbol{P} := \boldsymbol{A}\boldsymbol{A}^\top$ project onto the support and $\boldsymbol{P}^\perp := \boldsymbol{I}_D - \boldsymbol{P}$ onto its orthogonal complement. Define $\sigma_t := (\alpha_t^2 + \beta_t^2)^{1/2}$. The last quantity is the standard deviation within the subspace for the independent path when $\rho = 0$. Write $\mathcal{T}_t = (\alpha_t/\sigma_t^2)(I + \mathcal{N})\mathcal{P}_{\beta_t/\sigma_t}$ for the operator describing conditional averaging, where I is the identity operator. The expansions below are in L^2 under the corresponding path law.

Proposition 1 (First-order population velocities)**.** Let H be as in equation 8 and satisfy Assumption 3.1. Under the affine schedule $\alpha_t = 1 - t$ and $\beta_t = t$, for every fixed $t \in (0, 1)$, as $\rho \downarrow 0$,

$$\boldsymbol{v}_{t,\mathrm{OT}}(\boldsymbol{x}) = -\frac{1}{\alpha_t}\boldsymbol{P}^\perp\boldsymbol{x} + \rho H(\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{x})\boldsymbol{A}\boldsymbol{u} + O_{t,L^2}(\rho^2). \tag{10}$$

Under independent coupling, the population velocity satisfies

$$\boldsymbol{v}_{t,\mathrm{ind}}(\boldsymbol{x}) = -\frac{1}{\alpha_t}\boldsymbol{P}^\perp\boldsymbol{x} + \frac{\beta_t - \alpha_t}{\sigma_t^2}\boldsymbol{P}\boldsymbol{x} + \rho(\mathcal{T}_t H)\left(\frac{\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{x}}{\sigma_t}\right)\boldsymbol{A}\boldsymbol{u} + O_{t,L^2}(\rho^2). \tag{11}$$

Sketch of the derivation. The normal component of the input is $\alpha_t\boldsymbol{A}_\perp\boldsymbol{\Xi}$, while the target has no normal component. Both couplings therefore give the normal velocity $-\boldsymbol{P}^\perp\boldsymbol{x}/\alpha_t$. Along the feature direction, set $S = \boldsymbol{u}^\top\boldsymbol{Z}$. OT pairs S with $S + \rho H(S)$. The observed coordinate $S + \beta_t\rho H(S)$ determines S, so its velocity is $\rho H(S)$. Replacing S by the observed coordinate changes this velocity only at order ρ^2.

Under independent coupling, conditional averaging acts separately on each Hermite degree. The Gaussian conditional expectation contributes $(\beta_t/\sigma_t)^j$, and expanding the velocity contributes $(j + 1)\alpha_t/\sigma_t^2$. Hence

$$(\mathcal{T}_t H)(s) = \sum_{j=2}^{J}(j+1)h_j\frac{\alpha_t\beta_t^j}{\sigma_t^{j+2}}\,\mathrm{He}_j(s).$$

As $\varepsilon_t = (1 - t)/t \to 0$, the multiplier of degree j is $(j + 1)\varepsilon_t + O(\varepsilon_t^2)$. The proof and uniform remainder estimate are in Appendix A.4. □

Thus independent coupling weakens the first-order nonlinear velocity and changes the relative strengths of its components. This expansion is valid for small transformation strength ρ. To compare the couplings without changing the target, the next theorem fixes ρ and takes the endpoint limit.

3.3 Nonlinear signal and label noise at fixed target

We now fix ρ and allow general links satisfying Assumption 3.1, without the polynomial restriction, to measure the nonlinear signal as the input approaches the target. The total velocity norm does not isolate this signal because it also includes motion toward the support and linear motion within it. We therefore remove the best linear fit. Inputs and labels have zero mean, so no intercept is needed. For coupling $C \in \{\mathrm{OT}, \mathrm{ind}\}$, minimize over linear maps $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ to define the *nonlinear signal energy*:

$$\mathcal{E}_C(t) := \min_{\boldsymbol{M}\in\mathbb{R}^{D\times D}} \mathbb{E}\left\|\boldsymbol{v}_{t,C}(\boldsymbol{X}_{t,C}) - \boldsymbol{M}\boldsymbol{X}_{t,C}\right\|^2. \tag{12}$$

Write $\boldsymbol{M}^*_{C,t}$ for the minimizing matrix. This removes everything a linear predictor can fit, including the covariance change in Lemma 3.2. The residual is the target for nonlinear prediction.

A training label also contains variation around its conditional mean. This part is unpredictable from the input, even when the population velocity is known. Along the feature direction $\boldsymbol{A}\boldsymbol{u}$, we measure this *label noise* by

$$\mathcal{V}_C(t) := \mathbb{E}\left[\left((\boldsymbol{A}\boldsymbol{u})^\top\{\boldsymbol{U}_{t,C} - \boldsymbol{v}_{t,C}(\boldsymbol{X}_{t,C})\}\right)^2\right]. \tag{13}$$

Expectations use the corresponding coupling and measure variance only along $\boldsymbol{A}\boldsymbol{u}$.

Let $S \sim \mathcal{N}(0, 1)$ and write

$$W = S + \rho H(S), \qquad m_H = \mathbb{E}[H(S)^2], \qquad q_\rho = \mathbb{E}[W^2] = 1 + \rho^2 m_H. \tag{14}$$

Here W is the target coordinate along $\boldsymbol{Au}$. For OT, Assumption 3.1 is sufficient. For the independent endpoint rate, we also require W to have a positive C^1 density p_ρ on $\mathbb{R}$ with finite Fisher information, $I_\rho := \mathbb{E}[(p'_\rho(W)/p_\rho(W))^2] = \int_{\mathbb{R}} p'_\rho(w)^2/p_\rho(w)\,dw < \infty$. This bounds the mean squared score of the target density.

Thm 1 (Nonlinear signal and label noise at fixed ρ)**.** Fix $\rho > 0$ and H satisfying Assumption 3.1. Use $\boldsymbol{X}_t = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1$ and $\varepsilon_t = (1-t)/t$. For every $0 < t < 1$, OT coupling satisfies

$$\mathcal{E}_{\mathrm{OT}}(t) = \frac{\rho^2 m_H}{1 + t^2\rho^2 m_H}, \qquad \mathcal{V}_{\mathrm{OT}}(t) = 0. \tag{15}$$

If W additionally has a positive C^1 density with $I_\rho < \infty$, then

$$\lim_{t\uparrow 1} \frac{\mathcal{E}_{\mathrm{ind}}(t)}{\varepsilon_t^2} = I_\rho - \frac{1}{q_\rho} > 0, \qquad \lim_{t\uparrow 1} \mathcal{V}_{\mathrm{ind}}(t) = 1. \tag{16}$$

Sketch of the derivation. Only the coordinate along $\boldsymbol{Au}$ contains a nonlinear signal. Under OT, its input is $S + t\rho H(S)$ and its velocity label is $\rho H(S)$. The input determines S, so the conditional label variance is zero. Removing the best linear fit from this scalar velocity gives $\rho^2 m_H/(1 + t^2\rho^2 m_H)$.

For independent coupling, write the input coordinate as $Y = \alpha_t N + \beta_t W$, where $N \sim \mathcal{N}(0,1)$ is independent of W. Let ℓ_t be the derivative of the log density of Y and $d_t = \mathbb{E}[Y^2]$. Gaussian conditioning shows that the nonlinear velocity remaining after the best linear fit is $\varepsilon_t(\ell_t(Y) + Y/d_t)$. Its squared norm divided by ε_t^2 converges to $I_\rho - 1/q_\rho$. Meanwhile, the input approaches W without revealing N, so the label variance tends to one. Finally, under either coupling, the normal component of the input is $(1-t)\boldsymbol{A}_\perp\boldsymbol{\Xi}$ and the normal label is $-\boldsymbol{A}_\perp\boldsymbol{\Xi}$. The input therefore determines the normal label. Since $\boldsymbol{A}_\perp$ has orthonormal columns and $\mathbb{E}\|\boldsymbol{\Xi}\|^2 = D - r$,

$$\boldsymbol{P}^\perp \boldsymbol{v}_{t,C}(\boldsymbol{X}_{t,C}) = -\boldsymbol{A}_\perp\boldsymbol{\Xi}, \qquad \mathbb{E}\|\boldsymbol{P}^\perp \boldsymbol{v}_{t,C}(\boldsymbol{X}_{t,C})\|^2 = D - r. \tag{17}$$

Appendix A.2 gives the full proof. □

The independent endpoint gives a direct explanation for the vanishing nonlinear signal. Along the feature direction, the input at $t = 1$ is W and the label is $W - N$, so independence gives $\mathbb{E}[W - N \mid W] = W$. Thus a non-Gaussian target can have a linear endpoint regression function. For OT with an invertible transformation $g(s) = s + \rho H(s)$, the endpoint input instead determines the paired source $g^{-1}(W)$, leaving the nonlinear velocity $W - g^{-1}(W)$. The theorem quantifies how the nonlinear signal approaches these endpoint values.

Under independent coupling, the nonlinear signal vanishes while label noise remains. OT has a nonzero signal without conditional label noise. Its sample gradients can still vary across inputs. Both couplings have normal signal energy $D - r$ for $r < D$. Label noise adds a parameter-independent term to the population loss, so it does not itself slow population gradient flow. The next result isolates the slowdown caused by the weaker nonlinear signal.

3.4 Joint support and feature learning

To translate signal strength into training time, we jointly train the support, linear velocity, and feature direction. We construct a predictor with the three components of the velocity decomposition above. Let $\boldsymbol{B} \in \mathbb{R}^{D\times r}$ have orthonormal columns spanning a candidate support. Substituting $\boldsymbol{BB}^\top$ for $\boldsymbol{AA}^\top$ in the exact normal velocity makes this component contract toward the learned subspace, with the known coefficient $1/(1-t)$. For the Gaussian target ($\rho = 0$), the within-subspace linear velocity has coefficient $\ell_{\mathrm{OT},t} = 0$ or $\ell_{\mathrm{ind},t} = (2t-1)/\sigma_t^2$. We use this baseline and learn a correction $\boldsymbol{K} \in \mathbb{R}^{r\times r}$ to fit the linear velocity when $\rho > 0$.

To capture nonlinear variation beyond this linear fit, we add one cubic neuron with activation $\mathrm{He}_3(z) = z^3 - 3z$, unit input direction $\boldsymbol{w} \in \mathbb{S}^{r-1} := \{\boldsymbol{w} \in \mathbb{R}^r : \|\boldsymbol{w}\| = 1\}$, and output weights $\boldsymbol{a} \in \mathbb{R}^r$. We normalize its input by the within-subspace Gaussian standard deviation, $d_{\mathrm{OT},t} = 1$ or $d_{\mathrm{ind},t} = \sigma_t$. The cubic predictor makes the dependence of learning dynamics on feature alignment explicit. We study direction recovery without assuming that this predictor represents the full nonlinear velocity for a general link H. Combining these three components gives

$$\widehat{\boldsymbol{v}}_{\boldsymbol{B},\boldsymbol{K},\boldsymbol{a},\boldsymbol{w}}(\boldsymbol{x}) = -\frac{(\boldsymbol{I}_D - \boldsymbol{BB}^\top)\boldsymbol{x}}{1-t} + \boldsymbol{B}(\ell_{C,t}\boldsymbol{I}_r + \boldsymbol{K})\boldsymbol{B}^\top\boldsymbol{x} + \boldsymbol{Ba}\,\mathrm{He}_3(\boldsymbol{w}^\top\boldsymbol{B}^\top\boldsymbol{x}/d_{C,t}). \tag{18}$$

We train all four parameter blocks jointly, without knowing $\boldsymbol{A}$. At fixed flow time t, they follow gradient flow in training time s on the population loss $\mathcal{J}_{C,t} = \frac{1}{2}\mathbb{E}\|\widehat{\boldsymbol{v}}(\boldsymbol{X}_{t,C}) - \boldsymbol{U}_{t,C}\|^2$, with unit learning rates and Euclidean gradients projected onto the tangent spaces of $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}_r$ and $\|\boldsymbol{w}\| = 1$. Subscripts s and 0 denote the current and initial values. We initialize $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$ and choose $\boldsymbol{B}_0, \boldsymbol{w}_0$ at random. Linear weights follow gradient flow without instantaneous refitting. Appendix B.1 gives the exact updates.

Fitting the support and linear velocity can reduce the loss without recovering the nonlinear feature. We therefore measure improvement beyond the best linear predictor, relative to what the cubic model can achieve. Define $J_{\mathrm{lin},C,t} := \frac{1}{2}\mathbb{E}\|\boldsymbol{U}_{t,C} - \boldsymbol{M}^*_{C,t}\boldsymbol{X}_{t,C}\|^2$. The same matrix minimizes label and population-velocity regression losses by conditional expectation. Let $J_{\mathrm{cub},C,t} := \min_{\boldsymbol{K},\boldsymbol{a},\|\boldsymbol{w}\|=1} \mathcal{J}_{C,t}(\boldsymbol{A}, \boldsymbol{K}, \boldsymbol{a}, \boldsymbol{w})$. For $\theta \in (0,1)$, define the *hitting time* $T^{\mathrm{full}}_{C,\theta}(t)$ as the first gradient-flow time at which a fraction θ of this gap has been closed:

$$T^{\mathrm{full}}_{C,\theta}(t) = \inf\left\{s \geq 0 : J_{\mathrm{lin},C,t} - \mathcal{J}_{C,t}(s) \geq \theta\big(J_{\mathrm{lin},C,t} - J_{\mathrm{cub},C,t}\big)\right\}. \tag{19}$$

For $\theta = 1/2$, half the available nonlinear improvement is required, even when the signal is small. When the cubic correlation is nonzero, optimizing the linear and cubic weights on the true support gives a loss reduction strictly increasing with absolute feature alignment (Appendix A.5). The threshold thus measures direction learning without requiring exact velocity recovery. The reference losses are not used in training.

The energy $\mathcal{E}_C(t)$ measures the entire nonlinear residual, whereas a cubic neuron can use only the part correlated with its feature. Define the *residual cubic correlation*

$$\kappa_C(t) := \mathbb{E}\left[(\boldsymbol{A}\boldsymbol{u})^\top(\boldsymbol{U}_{t,C} - \boldsymbol{M}^*_{C,t}\boldsymbol{X}_{t,C})\ \mathrm{He}_3\left(\frac{(\boldsymbol{A}\boldsymbol{u})^\top\boldsymbol{X}_{t,C}}{d_{C,t}}\right)\right]. \tag{20}$$

For this training analysis, assume $\mathbb{E}|H(S)|^6 < \infty$. Recall $W = S + \rho H(S)$ and $q_\rho = \mathbb{E}[W^2]$, and define

$$K_{\mathrm{OT}} := \frac{\mathbb{E}[W^4]}{q_\rho} - \mathbb{E}[SW^3], \quad K_{\mathrm{ind}} := \mathbb{E}[W^4] - 3q_\rho^2. \tag{21}$$

Here K_{OT} and K_{ind} are scalar, while $\boldsymbol{K}$ is a trainable matrix. The mixed moment in K_{OT} uses the pairing $W = S + \rho H(S)$. Near the endpoint, $\kappa_{\mathrm{OT}}(t) \to K_{\mathrm{OT}}$ and $\kappa_{\mathrm{ind}}(t)/\varepsilon_t \to K_{\mathrm{ind}}/q_\rho$ (Appendix B.1). If $\kappa_C(t) = 0$, no feature direction on the true support improves on the best linear loss, even when $\mathcal{E}_C(t) > 0$ (Appendix A.5). Thus the condition $K_{\mathrm{OT}}K_{\mathrm{ind}} \neq 0$ ensures that the cubic model can learn from the residual near the endpoint. It holds, for example, for $H = \mathrm{He}_3$ with $0 < \rho < 1/3$. The theorem also allows noncubic links.

Thm 2 (Joint support and feature learning)**.** Suppose Assumption 3.1 holds, $\mathbb{E}|H(S)|^6 < \infty$, and $1 + \rho H'(s) > 0$ for all s. Assume $K_{\mathrm{OT}}K_{\mathrm{ind}} \neq 0$. Fix the improvement fraction $\theta \in (0,1)$ and failure probability $\delta \in (0,1)$, and write $\varepsilon_t = (1-t)/t$. There are $r_\delta \geq 2$ and $c_\delta, C_\delta > 0$, depending only on H, ρ, θ, δ, such that for $D \geq r \geq r_\delta$ the following holds. Initialize $\boldsymbol{B}_0$ uniformly among $D \times r$ matrices with orthonormal columns, independently of uniform $\boldsymbol{w}_0 \in \mathbb{S}^{r-1}$. There is a deterministic threshold $t_0 = t_0(D, r, H, \rho, \theta, \delta) < 1$ such that, with probability at least $1 - \delta$, for both couplings and all $t_0 \leq t < 1$,

$$c_\delta r^2 \leq T^{\mathrm{full}}_{\mathrm{OT},\theta}(t) \leq C_\delta r^2, \qquad c_\delta r^2 \varepsilon_t^{-2} \leq T^{\mathrm{full}}_{\mathrm{ind},\theta}(t) \leq C_\delta r^2 \varepsilon_t^{-2}. \tag{22}$$

As $s \to \infty$, the learned support converges to $\mathrm{span}(\boldsymbol{A})$, the feature direction $\boldsymbol{B}_s\boldsymbol{w}_s$ converges to $\boldsymbol{A}\boldsymbol{u}$ up to sign, and the loss converges to $J_{\mathrm{cub},C,t}$. The endpoint threshold t_0 may depend on dimension.

Sketch of the derivation. Both couplings have the same support term $\mathbb{E}\|(\boldsymbol{I}_D - \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{X}_1\|^2/(2\alpha_t^2)$ in the loss. Its coefficient grows near the endpoint, driving rapid recovery of $\mathrm{span}(\boldsymbol{A})$. The subsequent difference in training time comes from feature alignment within this subspace.

After the best linear fit is removed, the cubic correlation converges to K_{OT} under OT, whereas under independent coupling it is $\varepsilon_t K_{\mathrm{ind}}/q_\rho$ to leading order. Write this correlation as κ and the absolute feature alignment as $m = |(\boldsymbol{A}\boldsymbol{u})^\top\boldsymbol{B}\boldsymbol{w}|$. On the true support, the residual correlation with the cubic feature is proportional to κm^3. Optimizing the linear and cubic output weights therefore gives a loss improvement of order $\kappa^2 m^6$ for small m (Appendix A.5). At fixed alignment, the optimal output

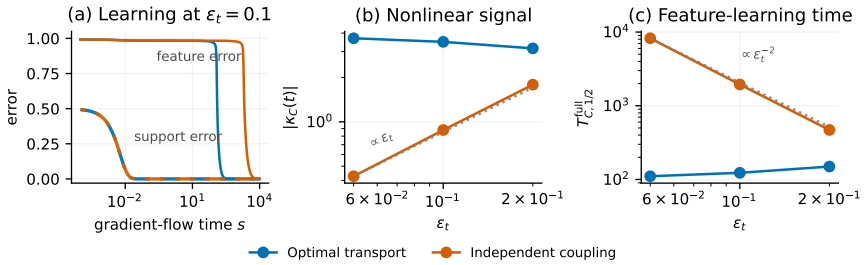

Figure 1: **Fixed-time joint learning with an unknown support.** (a) Support error (dashed) and feature-direction error (solid) at $\varepsilon_t = 0.1$. (b) Residual cubic correlation after the best linear fit. (c) Time to close half the loss gap between the best linear predictor and the optimal cubic model on the true support (equation 19). Gray dotted lines indicate ε_t scaling in (b) and ε_t^{-2} scaling in (c).

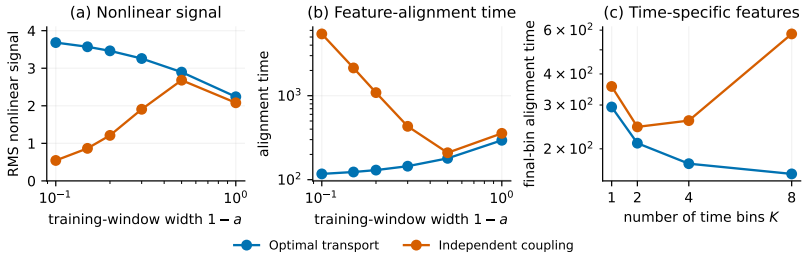

Figure 2: **Feature sharing across flow times.** (a) Nonlinear signal strength and (b) feature-alignment time as the training window narrows toward the endpoint, with features shared across times. (c) Endpoint-bin alignment time with separate features in K time bins. Alignment times use a feature-direction error threshold of $1/2$.

weights are proportional to κ. Combining these weights with the residual signal gives the κ^2 scale of the direction update. In the reduced direction dynamics, dm/ds is of order $\kappa^2 m^5$. Integrating from a small initial alignment m_0 to a fixed positive alignment gives a time of order $\kappa^{-2} m_0^{-4}$. A weaker signal thus slows learning both by reducing the output weights and by reducing the signal driving the direction update. After initial support recovery, m_0 is of order $r^{-1/2}$ with high probability (Appendix B.1), giving the common r^2 factor. The independent correlation contributes the additional ε_t^{-2} factor. Appendix B proves these statements for simultaneous updates of all parameter blocks. □

Both couplings eventually recover the support and feature direction, but independent coupling becomes arbitrarily slow near the endpoint at fixed dimension.

3.5 Numerical Experiments

We compare the training speeds between normal-field and density-direction, test whether sharing features across times reduces the slowdown, and examine the distinction on the real image manifolds.

Fixed-time dynamics. Both couplings recover the support quickly, but independent coupling learns the feature direction more slowly (Figure 1(a)). We jointly train the cubic-neuron model on a two-dimensional target subspace in three dimensions, using the same initialization for both couplings. In panel (b), the residual cubic correlation $|\kappa_C(t)|$ remains nonzero under OT and decreases in proportion to ε_t under independent coupling. Panel (c) measures the time to close half the nonlinear loss gap in equation 19. Halving ε_t leaves the OT time nearly unchanged and increases the independent time by about fourfold, consistent with Theorem 2. Appendix C.1 gives the setup and error definitions.

Sharing parameters across flow times. We compare a shared feature direction with separate directions for different time bins, keeping the support basis shared in both cases. Training minimizes the population loss under a discrete distribution on Gauss–Legendre flow-time nodes strictly inside each window or bin, with the corresponding quadrature weights. With a shared direction, both couplings learn on similar time scales over the full time window. Narrowing the window $(a, 1)$

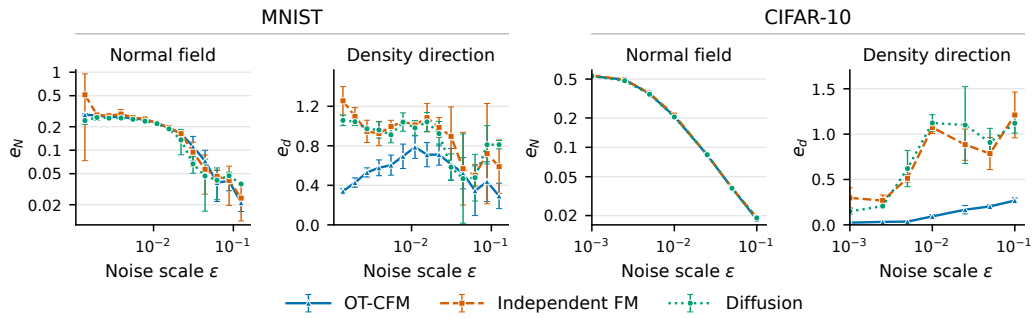

Figure 3: **Normal and density-direction errors on decoder manifolds.** The left pair shows MNIST and the right pair CIFAR-10. Each pair reports normal-field error e_N followed by density-direction error e_d. Lower values are better. Points and error bars show means and standard errors over selected runs. Both axes are logarithmic, with separate vertical ranges. See Appendix C.3 for details.

toward the endpoint weakens the independent signal and slows learning, while OT has a nonzero signal and remains fast (Figure 2(a,b)). With separate directions, endpoint alignment takes longer under independent coupling than under OT for eight bins (panel (c)). Separate feature directions receive direct updates from their own bins, but remain coupled through the shared support basis. Thus, the stronger signal at smaller flow times can help learn the feature used near the endpoint when that feature is shared across times. Appendix C.2 gives the setup.

MNIST and CIFAR-10 decoder manifolds. Frozen class-conditional variational autoencoder decoders generate images from Gaussian latents with a nonlinear transformation of one latent coordinate. At each noise scale, convolutional neural networks learn vector fields from noisy images using OT-CFM, independent FM, or diffusion targets. The density direction is the local tangent gradient of the transformed latent coordinate. Normal-field and density-direction errors are measured against each method's separately trained reference field and normalized by its energy in the corresponding component (equation 165–equation 166). Figure 3 shows MNIST in the left pair and CIFAR-10 in the right pair. Each pair reports normal-field error e_N followed by density-direction error e_d. Normal-field errors are nearly identical across OT-CFM, independent FM, and diffusion, whereas only OT-CFM achieves low density-direction errors at small noise scales. These results suggest that OT-CFM differs from the other methods primarily in its ability to learn the density direction. Appendix C.3 describes the detailed settings.

4 Discussion and Limitations

Conclusions. Coupling changes nonlinear feature learning with a fixed target. In our model, both couplings recover the support quickly near the endpoint under population gradient flow. Independent coupling weakens the nonlinear signal and slows feature learning, whereas OT has a nonzero signal and bounded feature-learning time at fixed dimension. Support recovery alone therefore does not imply learning the nonlinear structure within it.

Limitations. Our training-time bounds use a subspace model with one cubic neuron and population gradient flow. Extending these bounds to multiple interacting features or deeper networks requires additional analysis. In this setting, the feature direction is identifiable from the target covariance. Whether the same slowdown holds when direction recovery requires higher-order statistics also remains open. The analysis also fixes the flow time, restricting the learner to the signal at that time. With shared features, however, training at smaller flow times can help learn the direction used near $t = 1$ (Figure 2). Extending the bounds to time-conditioned networks therefore requires understanding when shared weights produce shared feature directions across times.

Our results clarify nonlinear feature learning, but this does not necessarily imply improved generation quality. Even exact feature-direction recovery can leave velocity approximation error, and the optimal cubic predictor can blow up in finite time during sampling (Appendix A.6). Generation guarantees require controlling approximation errors and sampling stability across flow times.

AI USE STATEMENT

We used large language models to improve the clarity and style of the writing and to seek suggestions on possible approaches to proving mathematical claims. We reviewed all AI-assisted work. For the mathematical results, the authors developed the final rigorous proofs and wrote their exposition to ensure clarity. The authors take full responsibility for the final content of the paper, including all text, mathematical claims, and proofs.

REFERENCES

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *Advances in Neural Information Processing Systems*, 38:141266–141286, 2025.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: https://doi.org/10.1002/cpa.3160440402.

Jack Carr. *Applications of Centre Manifold Theory*, volume 35 of *Applied Mathematical Sciences*. Springer New York, 1981.

Nisha Chandramoorthy and Adriaan de Clercq. When and how can inexact generative models still sample from the data manifold? *Advances in Neural Information Processing Systems*, 38: 170326–170357, 2025.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in Neural Information Processing Systems*, pp. 27340–27371, 2024.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36:752–784, 2023.

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

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Kenji Fukumizu, Wei Huang, Han Bao, Shuntuo Xu, and Nisha Chandramoothy. Flow matching from viewpoint of proximal operators. *arXiv preprint arXiv:2602.12683*, 2026.

Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces. *npj Artificial Intelligence*, 2026.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *International Conference on Learning Representations*, volume 2025, 2025.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*, volume 2024, 2024.

Wei Huang, Andi Han, Mingyuan Bai, Huanjian Zhou, Qixin Zhang, Taiji Suzuki, and Kenji Fukumizu. Provably learning diffusion models under the manifold hypothesis: Collapse and refine. *arXiv preprint arXiv:2605.20235*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Xiang Li, Zebang Shen, Ya-Ping Hsieh, and Niao He. When scores learn geometry: Rate separations under the manifold hypothesis. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiao Li, Zekai Zhang, Xiang Li, Siyi Chen, Zhihui Zhu, Peng Wang, and Qing Qu. Understanding representation dynamics of diffusion models via low-dimensional modeling. *Advances in Neural Information Processing Systems*, 38:107365–107404, 2025.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Zichen Liu, Wei Zhang, and Tiejun Li. Improving the euclidean diffusion generation of manifold data by mitigating score function singularity. *Advances in Neural Information Processing Systems*, 38:110885–110917, 2025.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 21–27 Jul 2024.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019.

Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T Q Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28100–28127. PMLR, July 2023.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 211–215. IEEE, 2025.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737, 2021.

A PROOFS

A.1 PROOF OF LEMMA 3.2

Proof. Write $S = \boldsymbol{u}^\top \boldsymbol{Z}$ and $\boldsymbol{Z} = S\boldsymbol{u} + \boldsymbol{Z}_\perp$. Since S and $\boldsymbol{Z}_\perp$ are independent centered Gaussians, the moment conditions in equation 5 imply

$$\mathbb{E}[\boldsymbol{Z}H(S)] = \boldsymbol{u}\mathbb{E}[SH(S)] + \mathbb{E}[\boldsymbol{Z}_\perp]\mathbb{E}[H(S)] = 0. \tag{23}$$

The representation $\boldsymbol{X}_1 = \boldsymbol{A}[\boldsymbol{Z} + \rho H(S)\boldsymbol{u}]$ gives $\boldsymbol{X}_1 \in \mathcal{M}$ and $\mathbb{E}[\boldsymbol{X}_1] = 0$. Expanding the covariance therefore yields

$$\mathrm{Cov}(\boldsymbol{A}^\top \boldsymbol{X}_1) = \mathbb{E}[(\boldsymbol{Z} + \rho H(S)\boldsymbol{u})(\boldsymbol{Z} + \rho H(S)\boldsymbol{u})^\top] = \boldsymbol{I}_r + \rho^2 \mathbb{E}[H(S)^2]\boldsymbol{u}\boldsymbol{u}^\top. \tag{24}$$

To identify the OT map, let $V' = H$ and define

$$\Psi_\rho(\boldsymbol{g}) = \frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{g}\|^2 + \rho V(\boldsymbol{u}^\top \boldsymbol{A}^\top \boldsymbol{g}). \tag{25}$$

For $\boldsymbol{A}^\top \boldsymbol{g} = s\boldsymbol{u} + \boldsymbol{y}_\perp$, this potential is $\frac{1}{2}\|\boldsymbol{y}_\perp\|^2 + \frac{1}{2}s^2 + \rho V(s)$. Its derivative in s is $s + \rho H(s)$, which is nondecreasing by Assumption 3.1. Thus Ψ_ρ is convex, and $\nabla\Psi_\rho = T_\rho$ by direct differentiation. Moreover, $T_\rho(\boldsymbol{X}_0) = \boldsymbol{A}[\boldsymbol{Z} + \rho H(S)\boldsymbol{u}]$ has law P_1. Both endpoint laws have finite second moments and P_0 is absolutely continuous, so Brenier's theorem identifies T_ρ as the unique quadratic-cost OT map, up to P_0-null sets. □

A.2 PROOF OF THEOREM 1

Proof. Fix $t \in (0, 1)$ and abbreviate $\alpha = 1 - t$, $\beta = t$, and $\varepsilon = \alpha/\beta$. For independent standard Gaussians S, N, set

$$h_\rho(s) = s + \rho H(s), \qquad W = h_\rho(S), \qquad q_\rho = \mathbb{E}[W^2] = 1 + \rho^2 m_H, \tag{26}$$

where $m_H = \mathbb{E}[H(S)^2]$. The moment conditions also give $\mathbb{E}W = 0$ and $\mathbb{E}[SW] = 1$.

Under either coupling, the normal components of the input and label are $\alpha\boldsymbol{A}_\perp\boldsymbol{\Xi}$ and $-\boldsymbol{A}_\perp\boldsymbol{\Xi}$, so the normal velocity is $-\boldsymbol{P}^\perp\boldsymbol{x}/\alpha$. In the tangent directions orthogonal to $\boldsymbol{A}\boldsymbol{u}$, the velocity is zero under OT and $(\beta - \alpha)\boldsymbol{x}/(\alpha^2 + \beta^2)$ under independent coupling. These components are linear in the input. They are also independent of the input–label pair along $\boldsymbol{A}\boldsymbol{u}$, and all coordinates are centered. Thus the best linear fit has no cross-coordinate terms, and

$$\mathcal{E}_C(t) = \min_{b\in\mathbb{R}} \mathbb{E}[(v_C(Y_C) - bY_C)^2], \qquad \mathcal{V}_C(t) = \mathbb{E}[\mathrm{Var}(V_C \mid Y_C)], \tag{27}$$

where $Y_C = (\boldsymbol{A}\boldsymbol{u})^\top \boldsymbol{X}_{t,C}$, $V_C = (\boldsymbol{A}\boldsymbol{u})^\top \boldsymbol{U}_{t,C}$, and $v_C(y) = \mathbb{E}[V_C \mid Y_C = y]$. The normal component also gives $\mathbb{E}\|\boldsymbol{P}^\perp \boldsymbol{v}_{t,C}(\boldsymbol{X}_{t,C})\|^2 = \mathbb{E}\|\boldsymbol{\Xi}\|^2 = D - r$, proving equation 17.

OT coupling. Here

$$Y_{\mathrm{OT}} = S + \beta\rho H(S), \qquad V_{\mathrm{OT}} = \rho H(S). \tag{28}$$

The map $s \mapsto \alpha s + \beta h_\rho(s)$ is strictly increasing and onto: h_ρ is continuous and nondecreasing, and $\alpha > 0$. Consequently Y_{OT} determines S, so $v_{\mathrm{OT}}(Y_{\mathrm{OT}}) = V_{\mathrm{OT}}$ and $\mathcal{V}_{\mathrm{OT}}(t) = 0$. The moment conditions give

$$\mathbb{E}[Y_{\mathrm{OT}}^2] = 1 + \beta^2\rho^2 m_H, \qquad \mathbb{E}[V_{\mathrm{OT}}Y_{\mathrm{OT}}] = \beta\rho^2 m_H, \qquad \mathbb{E}[V_{\mathrm{OT}}^2] = \rho^2 m_H. \tag{29}$$

Substituting these moments into the scalar least-squares formula yields

$$\mathcal{E}_{\mathrm{OT}}(t) = \rho^2 m_H - \frac{\beta^2\rho^4 m_H^2}{1 + \beta^2\rho^2 m_H} = \frac{\rho^2 m_H}{1 + \beta^2\rho^2 m_H}, \tag{30}$$

which proves equation 15.

Independent coupling. Now $Y = \alpha N + \beta W$ and $V = W - N = (Y - N)/\beta$. Let p_t be the density of Y and define

$$d = \mathbb{E}[Y^2] = \alpha^2 + \beta^2 q_\rho, \qquad \ell_t = \frac{p_t'}{p_t}, \qquad J_t = \mathbb{E}[\ell_t(Y)^2]. \tag{31}$$

Differentiating the Gaussian convolution gives

$$\mathbb{E}[N \mid Y = y] = -\alpha \ell_t(y), \qquad v(y) = \frac{y}{\beta} + \frac{\alpha}{\beta}\ell_t(y). \tag{32}$$

In particular, $J_t \le \alpha^{-2}$ by conditional Jensen's inequality, and $\mathbb{E}[Y\ell_t(Y)] = -\alpha^{-1}\mathbb{E}[YN] = -1$. Hence the best linear coefficient and the corresponding residual are

$$b^{\mathrm{lin}} = \frac{\mathbb{E}[Yv(Y)]}{d} = \frac{1}{\beta} - \frac{\alpha}{\beta d}, \qquad v(Y) - b^{\mathrm{lin}} Y = \varepsilon \left(\ell_t(Y) + \frac{Y}{d}\right). \tag{33}$$

For the label variance, the law of total variance gives

$$\mathbb{E}[\mathrm{Var}(V \mid Y)] = \frac{1}{\beta^2}\mathbb{E}[\mathrm{Var}(N \mid Y)] = \frac{1}{\beta^2}\left(1 - \alpha^2 J_t\right). \tag{34}$$

Together with equation 27, these calculations establish

$$\mathcal{E}_{\mathrm{ind}}(t) = \varepsilon^2 \left(J_t - \frac{1}{d}\right), \qquad \mathcal{V}_{\mathrm{ind}}(t) = \frac{1 - \alpha^2 J_t}{\beta^2}. \tag{35}$$

Endpoint limits. Assume now that W has a positive C^1 density p_ρ with finite Fisher information I_ρ, and write $\ell_\rho = p_\rho'/p_\rho$. Since $Y = \beta Z_\varepsilon$ for $Z_\varepsilon = W + \varepsilon N$, we have $J_t = \beta^{-2} I(Z_\varepsilon)$. Since $\int |p_\rho'| \le \sqrt{I_\rho} < \infty$, differentiation of the convolution gives

$$\ell_{Z_\varepsilon}(Z_\varepsilon) = \mathbb{E}[\ell_\rho(W) \mid Z_\varepsilon], \qquad I(Z_\varepsilon) \le I_\rho. \tag{36}$$

For the reverse inequality in the limit, use the variational formula

$$I(Z) = \sup_{f \in C_c^\infty(\mathbb{R})} \mathbb{E}[-2f'(Z) - f(Z)^2]. \tag{37}$$

Indeed, integration by parts identifies the expectation with $I(Z) - \mathbb{E}[(f(Z) - \ell_Z(Z))^2]$, and $C_c^\infty(\mathbb{R})$ is dense in L^2 of the law of Z. For every such f, weak convergence $Z_\varepsilon \Rightarrow W$ gives

$$\mathbb{E}[-2f'(W) - f(W)^2] = \lim_{\varepsilon \downarrow 0} \mathbb{E}[-2f'(Z_\varepsilon) - f(Z_\varepsilon)^2] \le \liminf_{\varepsilon \downarrow 0} I(Z_\varepsilon). \tag{38}$$

Taking the supremum and combining with equation 36 proves

$$I(Z_\varepsilon) \longrightarrow I_\rho, \qquad J_t \longrightarrow I_\rho. \tag{39}$$

Since $d \to q_\rho$, $\alpha \to 0$, and $\beta \to 1$, equation 35 now gives both limits in equation 16.

Strict positivity of the limiting signal. It remains to show that $I_\rho - 1/q_\rho$ is strictly positive. Applying integration by parts to $w\chi(w/R)$ and sending $R \to \infty$ gives $\mathbb{E}[W\ell_\rho(W)] = -1$, where χ is a smooth cutoff equal to one near the origin. The limit is justified by $\mathbb{E}|W\ell_\rho(W)| \le \sqrt{q_\rho I_\rho} < \infty$. Therefore

$$I_\rho - \frac{1}{q_\rho} = \mathbb{E}\left[\left(\ell_\rho(W) + \frac{W}{q_\rho}\right)^2\right] \ge 0. \tag{40}$$

Equality would imply $\ell_\rho(w) = -w/q_\rho$ everywhere, by continuity and positivity of p_ρ, and hence $W \sim \mathcal{N}(0, q_\rho)$. Because h_ρ is nondecreasing, it would then coincide almost everywhere with the Gaussian quantile map $h_\rho(s) = \sqrt{q_\rho}s$. But $\mathbb{E}[SW] = 1$ would force $q_\rho = 1$, contradicting $q_\rho = 1 + \rho^2 m_H > 1$. □

A.3 Links covered by the fixed-strength comparison

Write $h_\rho(s) = s + \rho H(s)$ for the scalar target map.

Single Hermite examples. For $H = \mathrm{He}_q$ with odd $q \geq 3$, define

$$M_q := -q \min_{s \in \mathbb{R}} \mathrm{He}_{q-1}(s) > 0. \tag{41}$$

Then $0 < \rho < M_q^{-1}$ makes $s + \rho H(s)$ strictly increasing. A single even degree is excluded because its derivative is negative in one tail. For the cubic feature, equation 63 gives the independent first-order coefficient relative to OT, after standardizing the scalar input:

$$a_{3,t} = 4\frac{\alpha_t \beta_t^3}{\sigma_t^5}, \qquad |a_{3,t} - 4\varepsilon_t| \leq 14\varepsilon_t^2 \quad (0 < \varepsilon_t \leq 1). \tag{42}$$

The target density. If h_ρ is a C^2 bijection with $h'_\rho > 0$, its density satisfies $p_\rho(h_\rho(s)) = \phi(s)/h'_\rho(s)$, where ϕ is the standard Gaussian density. Differentiating gives

$$\ell_\rho(h_\rho(s)) = -\frac{s}{h'_\rho(s)} - \frac{h''_\rho(s)}{h'_\rho(s)^2}. \tag{43}$$

In particular, the target has finite Fisher information whenever

$$\mathbb{E}\left[\left(\frac{S}{h'_\rho(S)} + \frac{h''_\rho(S)}{h'_\rho(S)^2}\right)^2\right] < \infty. \tag{44}$$

The polynomial and nonpolynomial examples below satisfy this condition.

Polynomial links. For $H = \mathrm{He}_j$ with odd $j \geq 3$, the bound $0 < \rho < M_j^{-1}$ in equation 41 makes h'_ρ bounded below by a positive constant. The map h_ρ is onto, and the score in equation 43 is bounded by a polynomial in s. Thus Theorem 1 applies with $m_H = j!$. Mixtures may also contain even degrees. For example, $H(s) = \mathrm{He}_3(s) + c\,\mathrm{He}_2(s)$ has $m_H = 6 + 2c^2$ and

$$\inf_s h'_\rho(s) = 1 - \rho(3 + c^2/3). \tag{45}$$

The same argument applies whenever $0 < \rho < (3 + c^2/3)^{-1}$.

A nonpolynomial link. Let $H(s) = \sin s - e^{-1/2}s$. Gaussian integration gives $\mathbb{E}H(S) = \mathbb{E}[SH(S)] = 0$ and $m_H = (1 - e^{-2})/2 - e^{-1} > 0$. If $0 < \rho < (1 + e^{-1/2})^{-1}$, then h'_ρ is bounded below by a positive constant and h''_ρ is bounded. Hence h_ρ is onto and equation 43 has finite Gaussian second moment, as required.

A link without endpoint regularity. Assumption 3.1 alone does not imply the additional density condition. Fix $\rho > 0$ and set

$$h_\rho(s) = 2\sqrt{\pi}\{\Phi(s) - 1/2\}, \qquad H(s) = \frac{h_\rho(s) - s}{\rho}, \tag{46}$$

where Φ is the standard Gaussian distribution function. This map is smooth and strictly increasing. Symmetry gives $\mathbb{E}H(S) = 0$, and Gaussian integration by parts gives $\mathbb{E}[Sh_\rho(S)] = \mathbb{E}[h'_\rho(S)] = 1$. Moreover, $W = h_\rho(S)$ is uniform on $[-\sqrt{\pi}, \sqrt{\pi}]$, so $m_H = (\pi/3 - 1)/\rho^2 > 0$. Thus the moment and monotonicity assumptions hold, but the target density has jumps at the boundary of its support. The independent endpoint rate in Theorem 1 does not cover this example.

A.4 Proof of Proposition 1

Lemma A.1 (Stability of polynomial regression)**.** *Let H, K be fixed polynomials and let Z_0, S be independent standard Gaussians. For $t \in (0,1)$ and $0 \leq \rho \leq \rho_0$, put $\alpha = 1-t$, $\beta = t$, $\sigma^2 = \alpha^2 + \beta^2$, and $T_\rho(s) = s + \rho H(s)$. Define*

$$Y_\rho = \alpha Z_0 + \beta T_\rho(S), \qquad Y_0 = \alpha Z_0 + \beta S, \qquad Q_t(y) = (\mathcal{P}_{\beta/\sigma} K)(y/\sigma). \tag{47}$$

There is a constant C, depending only on H, K, ρ_0, such that

$$\|\mathbb{E}[K(T_\rho(S)) \mid Y_\rho] - Q_t(Y_\rho)\|_{L^2} \leq C\rho. \tag{48}$$

Proof. Gaussian regression gives $\mathbb{E}[K(S) \mid Y_0] = Q_t(Y_0)$. Introduce independent standard Gaussians

$$W = Y_0/\sigma, \qquad V = (\beta Z_0 - \alpha S)/\sigma. \tag{49}$$

The polynomial $K(S) - Q_t(Y_0)$ has zero conditional mean given W. Its Hermite expansion in V therefore has the form $\sum_{k\geq 1} b_{k,t}(W)\,\mathrm{He}_k(V)$, with finitely many terms. For $B_t(W,V) = -\sum_{k\geq 1} b_{k,t}(W)\,\mathrm{He}_{k-1}(V)$, we have

$$(\partial_V - V)B_t = K(S) - Q_t(Y_0). \tag{50}$$

For a smooth compactly supported test function f, let $D_t = K(S) - Q_t(Y_0)$. Since $\partial_V Y_\rho = -\rho\alpha\beta H'(S)/\sigma$ and $\partial_{Z_0} Y_\rho = \alpha$, integration by parts gives

$$\begin{aligned}
\mathbb{E}[D_t f(Y_\rho)] &= -\mathbb{E}[B_t f'(Y_\rho)\partial_V Y_\rho] \\
&= \frac{\rho\alpha\beta}{\sigma}\mathbb{E}[B_t H'(S) f'(Y_\rho)] \\
&= \frac{\rho\beta}{\sigma}\mathbb{E}[H'(S)(Z_0 B_t - \partial_{Z_0} B_t) f(Y_\rho)].
\end{aligned} \tag{51}$$

In the last step, S is held fixed when integrating in Z_0. Since the identity holds for every such f, it yields

$$\mathbb{E}[K(S) - Q_t(Y_0) \mid Y_\rho] = \frac{\rho\beta}{\sigma}\mathbb{E}[H'(S)(Z_0 B_t - \partial_{Z_0} B_t) \mid Y_\rho]\,. \tag{52}$$

The coefficients of B_t are uniformly bounded because $\alpha/\sigma, \beta/\sigma \in [0,1]$ and $\sigma \geq 1/\sqrt{2}$. Conditional Jensen's inequality and Gaussian polynomial moments bound the right-hand side in L^2 by $C\rho$. The same polynomial-moment bound applies to $K(T_\rho(S)) - K(S)$ and $Q_t(Y_\rho) - Q_t(Y_0)$. The error in equation 48 decomposes as

$$\begin{aligned}
&\mathbb{E}[K(T_\rho(S)) \mid Y_\rho] - Q_t(Y_\rho) \\
&\quad = \mathbb{E}[K(T_\rho(S)) - K(S) \mid Y_\rho] \\
&\qquad + \mathbb{E}[K(S) - Q_t(Y_0) \mid Y_\rho] \\
&\qquad + \mathbb{E}[Q_t(Y_0) - Q_t(Y_\rho) \mid Y_\rho].
\end{aligned} \tag{53}$$

Each term has L^2 norm at most $C\rho$, by conditional Jensen's inequality and the preceding bounds, proving equation 48.

□

Lemma A.2 (Scalar velocity under independent coupling)**.** *Let H be a fixed polynomial whose derivative H' is bounded below on $\mathbb{R}$. Choose $\rho_0 > 0$ small enough that $1 + \rho H'(s) \geq 1/2$ for every $s \in \mathbb{R}$ and $0 \leq \rho \leq \rho_0$. For independent standard Gaussians Z_0, S, $t \in (0,1)$, and $\alpha = 1-t$, $\beta = t$, define*

$$\begin{aligned}
T_\rho(s) = s + \rho H(s), \qquad Y_\rho = \alpha Z_0 + \beta T_\rho(S), \qquad \sigma^2 = \alpha^2 + \beta^2, \\
K = (I + \mathcal{N})H = H + sH' - H'', \qquad Q_t(y) = (\mathcal{P}_{\beta/\sigma} K)(y/\sigma).
\end{aligned}$$

The score $\ell_{t,\rho}$ of Y_ρ satisfies

$$\ell_{t,\rho}(Y_\rho) = -\frac{Y_\rho}{\sigma^2} + \rho\frac{\beta}{\sigma^2} Q_t(Y_\rho) + O_{L^2}(\beta\rho^2). \tag{54}$$

The scalar conditional velocity has the expansion

$$\mathbb{E}[T_\rho(S) - Z_0 \mid Y_\rho] = \frac{Y_\rho}{\beta} + \frac{\alpha}{\beta}\ell_{t,\rho}(Y_\rho) = \frac{\beta - \alpha}{\sigma^2} Y_\rho + \rho\frac{\alpha}{\sigma^2} Q_t(Y_\rho) + O_{L^2}(\alpha\rho^2). \tag{55}$$

Both L^2 remainders are evaluated under the law of Y_ρ, with constants depending only on H, ρ_0, uniformly in $t \in (0,1)$.

Proof. Let ℓ_ρ be the score of $T_\rho(S)$. The change-of-variables formula gives

$$\ell_\rho(T_\rho(S)) = -\frac{S}{1+\rho H'(S)} - \frac{\rho H''(S)}{(1+\rho H'(S))^2} = -T_\rho(S) + \rho K(T_\rho(S)) + O_{L^2}(\rho^2). \tag{56}$$

To bound the remainder, expand the two reciprocal factors to first order and use $1+\rho H' \geq 1/2$. The resulting error is bounded by $C\rho^2$ times a polynomial in $|S|$. Replacing $K(S)$ by $K(T_\rho(S))$ has the same bound after multiplication by ρ.

Lemma A.1 gives equation 48 for this K, with $Q_t(y) = (\mathcal{P}_{\beta/\sigma}K)(y/\sigma)$.

Integration by parts in the two independent summands gives

$$\ell_{t,\rho}(Y_\rho) = \frac{1}{\beta}\mathbb{E}[\ell_\rho(T_\rho(S)) \mid Y_\rho], \qquad \mathbb{E}[Z_0 \mid Y_\rho] = -\alpha\ell_{t,\rho}(Y_\rho). \tag{57}$$

Substitute equation 56 into the first identity and use $\beta\mathbb{E}[T_\rho(S) \mid Y_\rho] = Y_\rho + \alpha^2\ell_{t,\rho}(Y_\rho)$. Together with equation 48, this gives equation 54. Finally, $\alpha+\beta=1$ gives

$$\mathbb{E}[T_\rho(S) - Z_0 \mid Y_\rho] = \frac{Y_\rho}{\beta} - \frac{1}{\beta}\mathbb{E}[Z_0 \mid Y_\rho] = \frac{Y_\rho}{\beta} + \frac{\alpha}{\beta}\ell_{t,\rho}(Y_\rho). \tag{58}$$

Substituting equation 54 proves equation 55. The factors β and α in the remainders follow from these identities and $\sigma^2 \geq 1/2$. □

Proof of Proposition 1. Write $\alpha = \alpha_t$, $\beta = \beta_t$, and $\sigma^2 = \alpha^2 + \beta^2$. All L^2 remainders below are evaluated under the corresponding path law. Since H is a polynomial and $s+\rho H(s)$ is nondecreasing for some $\rho > 0$, H' is bounded below. The limit $\rho \downarrow 0$ is taken with H fixed. We may therefore restrict to $0 \leq \rho \leq \rho_0$ such that $1+\rho H'(s) \geq 1/2$ for every s.

OT coupling. For independent standard Gaussians $\boldsymbol{Z}$ and $\boldsymbol{\Xi}$, set $S = \boldsymbol{u}^\top \boldsymbol{Z}$. The path and its velocity are

$$\begin{aligned}\boldsymbol{X}_{t,\mathrm{OT}} &= \boldsymbol{A}\{\boldsymbol{Z} + \beta\rho H(S)\boldsymbol{u}\} + \alpha\boldsymbol{A}_\perp\boldsymbol{\Xi},\\ \boldsymbol{U}_{t,\mathrm{OT}} &= \rho H(S)\boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}_\perp\boldsymbol{\Xi}.\end{aligned}$$

The map $s \mapsto s + \beta\rho H(s)$ is strictly increasing and onto, and the remaining coordinates are linear. Thus $\boldsymbol{X}_{t,\mathrm{OT}}$ determines both $\boldsymbol{Z}$ and $\boldsymbol{\Xi}$, so the marginal velocity equals $\boldsymbol{U}_{t,\mathrm{OT}}$. Define the remainder field by

$$\boldsymbol{R}_{t,\mathrm{OT}}(\boldsymbol{x}) := \boldsymbol{v}_{t,\mathrm{OT}}(\boldsymbol{x}) + \frac{1}{\alpha}\boldsymbol{P}^\perp\boldsymbol{x} - \rho H(\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{x})\boldsymbol{A}\boldsymbol{u}. \tag{59}$$

Then

$$\boldsymbol{R}_{t,\mathrm{OT}}(\boldsymbol{X}_{t,\mathrm{OT}}) = \rho\{H(S) - H(S+\beta\rho H(S))\}\boldsymbol{A}\boldsymbol{u}. \tag{60}$$

The expression in braces is ρ times a polynomial in S with coefficients bounded for $\beta \in [0,1]$ and $\rho \in [0,\rho_0]$. Gaussian polynomial moments give $\|\boldsymbol{R}_{t,\mathrm{OT}}(\boldsymbol{X}_{t,\mathrm{OT}})\|_{L^2} = O(\rho^2)$, which proves equation 10.

Independent coupling. Only the coordinate along $\boldsymbol{u}$ is nonlinear. Let Z_0, S be independent standard Gaussian scalars and define

$$T_\rho(s) = s + \rho H(s), \qquad Y_\rho = \alpha Z_0 + \beta T_\rho(S), \qquad K = (I+\mathcal{N})H = H + sH' - H''. \tag{61}$$

Lemma A.2 gives equation 55, with $Q_t(y) = (\mathcal{P}_{\beta/\sigma}K)(y/\sigma)$. The coordinates in $\boldsymbol{u}^\perp$ have the Gaussian conditional velocity $(\beta-\alpha)y/\sigma^2$, while the normal velocity is $-\boldsymbol{P}^\perp\boldsymbol{x}/\alpha$. Define $\boldsymbol{R}_{t,\mathrm{ind}}$ as the difference between the exact field and the first three terms below. Then

$$\begin{aligned}\boldsymbol{v}_{t,\mathrm{ind}}(\boldsymbol{x}) = &-\frac{1}{\alpha}\boldsymbol{P}^\perp\boldsymbol{x} + \frac{\beta-\alpha}{\sigma^2}\boldsymbol{P}\boldsymbol{x} + \rho(\mathcal{T}_t H)\left(\frac{\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{x}}{\sigma}\right)\boldsymbol{A}\boldsymbol{u} + \boldsymbol{R}_{t,\mathrm{ind}}(\boldsymbol{x}),\\ &\|\boldsymbol{R}_{t,\mathrm{ind}}(\boldsymbol{X}_{t,\mathrm{ind}})\|_{L^2} \leq C\alpha\rho^2,\end{aligned} \tag{62}$$

where

$$\mathcal{T}_t H = \frac{\alpha}{\sigma^2}\mathcal{P}_{\beta/\sigma}K = \sum_{j=2}^{J}(j+1)h_j\frac{\alpha\beta^j}{\sigma^{j+2}}\,\mathrm{He}_j\,. \tag{63}$$

This proves equation 11.

Endpoint limit. Write $\varepsilon = \alpha/\beta$. Since $\alpha + \beta = 1$,

$$\frac{\alpha}{\sigma^2} = \frac{\varepsilon(1+\varepsilon)}{1+\varepsilon^2}, \qquad \frac{\beta}{\sigma} = (1+\varepsilon^2)^{-1/2}. \tag{64}$$

Set $a_{j,t} = (j+1)\alpha\beta^j/\sigma^{j+2}$. For $p = (j+2)/2$ and $0 < \varepsilon \le 1$, the inequality $0 \le 1-(1+\varepsilon^2)^{-p} \le p\varepsilon^2$ yields

$$|a_{j,t} - (j+1)\varepsilon| \le (j+1)\varepsilon(\varepsilon + p\varepsilon^2) \le \tfrac{1}{2}(j+1)(j+4)\varepsilon^2. \tag{65}$$

Applying these identities to each of the finitely many Hermite coefficients proves

$$\mathcal{T}_t H = \varepsilon K + O_{L^2(\gamma)}(\varepsilon^2). \tag{66}$$

Moreover, $\sigma = 1 + O(\varepsilon)$ and the path variables have uniformly bounded polynomial moments for $0 \le \rho \le \rho_0$. Thus replacing $\boldsymbol{u}^\top \boldsymbol{A}^\top \boldsymbol{x}/\sigma$ by $\boldsymbol{u}^\top \boldsymbol{A}^\top \boldsymbol{x}$ in the first-order feature contributes $O_{L^2}(\rho\varepsilon^2)$. Since $\alpha \le \varepsilon$, equation 62 gives

$$\Delta \boldsymbol{v}_{t,\mathrm{ind}}(\boldsymbol{x}) = \rho\varepsilon K(\boldsymbol{u}^\top \boldsymbol{A}^\top \boldsymbol{x})\boldsymbol{A}\boldsymbol{u} + O_{L^2}(\varepsilon\rho^2 + \varepsilon^2\rho), \tag{67}$$

with a constant uniform near $t = 1$. In particular, for $H = \mathrm{He}_q$ the multiplier is

$$a_{q,t} = (q+1)\frac{\alpha\beta^q}{\sigma^{q+2}} = (q+1)\frac{\varepsilon(1+\varepsilon)}{(1+\varepsilon^2)^{(q+2)/2}}. \tag{68}$$

For $q = 3$, this reduces to

$$a_{3,t} = 4\frac{\alpha\beta^3}{\sigma^5}. \tag{69}$$

□

A.5 Cubic correlations and the loss on the true support

Fix C, t and use the scales from Appendix B.1. In coordinates on the true support, write

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{A}^\top \boldsymbol{X}_{t,C}, \qquad \boldsymbol{\Delta} = \boldsymbol{A}^\top \boldsymbol{U}_{t,C}, \\ \boldsymbol{L}_{C,t} &= \mathbb{E}[\boldsymbol{\Delta}\boldsymbol{y}^\top](\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top])^{-1}, \\ \boldsymbol{\eta} &= \boldsymbol{\Delta} - \boldsymbol{L}_{C,t}\boldsymbol{y}, \qquad \boldsymbol{z} = \boldsymbol{y}/d_{C,t}, \qquad V = \boldsymbol{u}^\top \boldsymbol{z}. \end{aligned} \tag{70}$$

Write $\kappa = \kappa_C(t)$, $v = \mathbb{E}[V^2]$, and $\mu_j = \mathbb{E}[V^j]$. All moments used below are finite under the sixth-moment assumption on $H(S)$.

Scalar residuals and cubic correlations. Set $\alpha = 1 - t$, $\sigma^2 = \alpha^2 + t^2$, and $g_t(S) = S + t\rho H(S)$. Under either coupling the normalized input has the form $\boldsymbol{z} = \boldsymbol{u}V + \boldsymbol{G}_\perp$, with $\boldsymbol{G}_\perp$ a standard Gaussian on $\boldsymbol{u}^\perp$ independent of V. The conditional residual is $\mathbb{E}[\boldsymbol{\eta} \mid \boldsymbol{z}] = \boldsymbol{u}R(V)$. For OT,

$$V = g_t(S), \qquad D_t = 1 + t^2\rho^2 m_H, \qquad \boldsymbol{u}^\top \boldsymbol{\eta} = \rho H(S) - \frac{t\rho^2 m_H}{D_t} g_t(S) = \frac{1}{t}\left(\frac{g_t(S)}{D_t} - S\right). \tag{71}$$

For independent coupling, take $N \sim \mathcal{N}(0,1)$ independent of W. Then

$$Y = \alpha N + tW, \qquad V = Y/\sigma, \qquad \boldsymbol{u}^\top \boldsymbol{\eta} = \frac{\alpha W - tq_\rho N}{\alpha^2 + t^2 q_\rho}. \tag{72}$$

In the other support directions, the independent residual is jointly Gaussian with, and uncorrelated with, the input, so its conditional mean is zero. In both cases $\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{\eta})V] = 0$. Expanding the third powers therefore gives

$$\begin{aligned} \kappa_{\mathrm{OT}}(t) &= \frac{1}{t}\left(\frac{\mathbb{E}[g_t(S)^4]}{D_t} - \mathbb{E}[S g_t(S)^3]\right), \\ \kappa_{\mathrm{ind}}(t) &= \frac{\alpha t^3 K_{\mathrm{ind}}}{(\alpha^2 + t^2 q_\rho)(\alpha^2 + t^2)^{3/2}}. \end{aligned} \tag{73}$$

For example, $\mathbb{E}[WY^3] = 3\alpha^2 t q_\rho + t^3\mathbb{E}[W^4]$ and $\mathbb{E}[NY^3] = 3\alpha^3 + 3\alpha t^2 q_\rho$ prove the second identity. Moment convergence as $t \uparrow 1$ yields $\kappa_{\mathrm{OT}}(t) \to K_{\mathrm{OT}}$ and $\kappa_{\mathrm{ind}}(t)/\varepsilon_t \to K_{\mathrm{ind}}/q_\rho$. Consequently, after choosing $t_0 < 1$ sufficiently close to one,

$$\frac{|K_{\mathrm{OT}}|}{2} \le |\kappa_{\mathrm{OT}}(t)| \le 2|K_{\mathrm{OT}}|, \qquad \frac{|K_{\mathrm{ind}}|}{2q_\rho}\varepsilon_t \le |\kappa_{\mathrm{ind}}(t)| \le \frac{2|K_{\mathrm{ind}}|}{q_\rho}\varepsilon_t. \tag{74}$$

The linear coefficient. Write $\boldsymbol{L}_{C,t} = \ell_{C,t}\boldsymbol{I}_r + \lambda_C(t)\boldsymbol{u}\boldsymbol{u}^\top$ and set $\zeta = \rho^2 m_H$. Direct regression gives

$$\lambda_{\mathrm{OT}}(t) = \frac{t\zeta}{1+t^2\zeta}, \qquad \lambda_{\mathrm{ind}}(t) = \frac{(1-t)t\zeta}{((1-t)^2+t^2 q_\rho)\sigma_t^2}. \tag{75}$$

In particular, $|\lambda_{\mathrm{ind}}(t)| = O(|\kappa_{\mathrm{ind}}(t)|)$ near the endpoint under $K_{\mathrm{ind}} \neq 0$.

The loss as a function of alignment. Suppress C, t, put $m = \boldsymbol{u}^\top \boldsymbol{w}$, and let $M(m) = \mathbb{E}[\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})^2]$. Conditionally on V, $\boldsymbol{w}^\top \boldsymbol{z}$ has law $mV + \sqrt{1-m^2}G$, with G an independent standard Gaussian. The Hermite addition identity gives

$$\begin{aligned}\mathbb{E}[\boldsymbol{\eta}\,\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})] &= \boldsymbol{u}\kappa m^3,\\ M(m) &= 6(1-m^2)^3 + 18\mathbb{E}[V^2]m^2(1-m^2)^2\\ &\quad + 9\mathbb{E}[(V^2-1)^2]m^4(1-m^2) + \mathbb{E}[\mathrm{He}_3(V)^2]m^6.\end{aligned} \tag{76}$$

Optimizing the linear and cubic weights. On the true support, the predictor is $\boldsymbol{g}_{\boldsymbol{K},\boldsymbol{a},\boldsymbol{w}}(\boldsymbol{y}) = (\ell_{C,t}\boldsymbol{I}_r + \boldsymbol{K})\boldsymbol{y} + \boldsymbol{a}\,\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})$, and its loss is $J = \frac{1}{2}\mathbb{E}\|\boldsymbol{g}_{\boldsymbol{K},\boldsymbol{a},\boldsymbol{w}}(\boldsymbol{y}) - \boldsymbol{\Delta}\|^2$. Let $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{I}_r + (v-1)\boldsymbol{u}\boldsymbol{u}^\top$ and write $\boldsymbol{w} = m\boldsymbol{u} + \sqrt{1-m^2}\boldsymbol{e}$, with $\boldsymbol{e} \perp \boldsymbol{u}$ a unit vector. The linear projection of the feature is determined by

$$\boldsymbol{c}(m) = \mathbb{E}[\boldsymbol{z}\,\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})] = (\mu_4 - 3v)m^3\boldsymbol{u} + 3(v-1)m^2\sqrt{1-m^2}\boldsymbol{e}. \tag{77}$$

Removing this projection gives the squared feature norm

$$\begin{aligned}\widetilde{M}(m) &= M(m) - \boldsymbol{c}(m)^\top \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}\boldsymbol{c}(m)\\ &= 6(1-m^2)^3 + 18vm^2(1-m^2)^2\\ &\quad + 9(\mu_4 - v^2)m^4(1-m^2) + D_3 m^6, \qquad D_3 = \mu_6 - \mu_4^2/v.\end{aligned} \tag{78}$$

All four coefficients are positive for the continuous nondegenerate law of V. In particular, $D_3 > 0$ by strict Cauchy–Schwarz. Define the feature after removing its linear projection by $\psi_{\boldsymbol{w}}(\boldsymbol{z}) = \mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z}) - \boldsymbol{c}(m)^\top \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}\boldsymbol{z}$. Then $\mathbb{E}[\boldsymbol{z}\psi_{\boldsymbol{w}}(\boldsymbol{z})] = 0$, $\mathbb{E}[\psi_{\boldsymbol{w}}(\boldsymbol{z})^2] = \widetilde{M}(m)$, and $\mathbb{E}[\boldsymbol{\eta}\psi_{\boldsymbol{w}}(\boldsymbol{z})] = \kappa m^3\boldsymbol{u}$. After optimizing the linear weights, the remaining loss is therefore

$$J - J_{\mathrm{lin}} = \tfrac{1}{2}\widetilde{M}(m)\|\boldsymbol{a}\|^2 - \kappa m^3 \boldsymbol{u}^\top \boldsymbol{a}. \tag{79}$$

Its minimizer is $\boldsymbol{a}^* = \kappa m^3 \boldsymbol{u}/\widetilde{M}(m)$, so

$$\min_{\boldsymbol{K},\boldsymbol{a}} J - J_{\mathrm{lin}} = -\frac{\kappa^2 m^6}{2\widetilde{M}(m)}. \tag{80}$$

To verify strict monotonicity of $m^6/\widetilde{M}(m)$, write $\widetilde{M}(m) = P(m^2)$ with $P(x) = \sum_{j=0}^3 A_j x^j (1-x)^{3-j}$ and $A_j > 0$. Then

$$3P(x) - xP'(x) = 3A_0(1-x)^2 + 2A_1 x(1-x) + A_2 x^2 > 0. \tag{81}$$

For $\kappa \neq 0$, strict convexity in the weights forces every stationary point to use these minimizing weights. Thus stationary points have either $m = 0$ and $J = J_{\mathrm{lin}}$, or $|m| = 1$ and $J = J_{\mathrm{lin}} - \kappa^2/(2D_3)$. The linear-plus-cubic feature Gram matrix is positive definite uniformly in $\boldsymbol{w}$, since its Schur complement is $\widetilde{M}(m) > 0$. The aligned optimal predictor adds $\kappa D_3^{-1}\boldsymbol{u}(V^3 - \mu_4 V/v)$ to $\boldsymbol{L}_{C,t}\boldsymbol{y}$. Consequently,

$$2(J_{\mathrm{lin},C,t} - J_{\mathrm{cub},C,t}) = \frac{\kappa_C(t)^2}{D_{3,C,t}} \leq \mathcal{E}_C(t). \tag{82}$$

Verification for cubic data. For $H = \mathrm{He}_3$ and $0 < \rho < 1/3$, Gaussian moments give

$$\begin{aligned}q_\rho &= 1 + 6\rho^2,\\ K_{\mathrm{ind}} &= 24\rho + 216\rho^2 + 1296\rho^3 + 3240\rho^4 > 0,\\ K_{\mathrm{OT}} &= \frac{6\rho(1 + 18\rho + 144\rho^2 + 432\rho^3 - 324\rho^4)}{1+6\rho^2} > 0.\end{aligned}$$

The final numerator is positive since $432\rho^3 - 324\rho^4 = 108\rho^3(4 - 3\rho) > 0$.

A.6 Approximation error and existence of the sampling flow

Approximation at a fixed time. For the jointly learned linear and cubic weights, equation 80 gives the best aligned predictor. Its remaining mean squared OT velocity error is

$$\mathcal{A}_{\mathrm{OT}}^{\mathrm{joint}}(t) = \frac{\rho^2 m_H}{1 + t^2\rho^2 m_H} - \frac{\kappa_{\mathrm{OT}}(t)^2}{D_{3,\mathrm{OT},t}} \geq 0. \tag{83}$$

This follows by subtracting twice the optimal loss reduction from $\mathcal{E}_{\mathrm{OT}}(t)$. For cubic data with $0 < \rho < 1/3$, the error is strictly positive for $t > 0$, even after exact direction recovery. Otherwise the exact scalar OT velocity would be a linear-plus-cubic polynomial in $y = g_t(S)$. A nonzero cubic term would give degree nine in S, whereas the true label $\rho(S^3 - 3S)$ has degree three. A linear term alone cannot match its linear and cubic coefficients simultaneously. Since S has full support, equality almost surely would imply a polynomial identity, which is impossible.

Existence of the sampling flow. For the same cubic target, the aligned OT predictor, with jointly optimized linear and cubic weights, has scalar form $\widehat{v}_t(y) = c(t)y + \beta(t)y^3$. Its moment formulas extend continuously to $t = 0$, where the predictor is exact and $\beta(0) = \rho > 0$. Thus there are $t_1 > 0$ and $C_0 \geq 0$ such that $\beta(t) \geq \rho/2$ and $c(t) \geq -C_0$ on $[0, t_1]$. For $y > 0$ with $y^2 \geq 4C_0/\rho$,

$$\dot{y} \geq \frac{\rho}{4}y^3, \qquad \frac{d}{dt}y^{-2} \leq -\frac{\rho}{2}. \tag{84}$$

Every positive initial value satisfying both this threshold and $2/(\rho y(0)^2) < t_1$ therefore leads to explosion no later than $2/(\rho y(0)^2)$. A Gaussian source exceeds this threshold with positive probability. This example explains why fixed-time feature-learning bounds alone do not guarantee a globally defined sampling flow.

B Joint support and feature learning

We prove Theorem 2 for the joint gradient flow equation 86. The support first approaches the true subspace in time $O(\alpha^2 \log(1/\alpha))$. The subsequent alignment dynamics require separate arguments for independent coupling, where $\kappa \to 0$, and OT, where κ has a nonzero limit.

We write $\alpha = 1 - t$, $\nu = q_\rho$, $\zeta = \nu - 1$, $\boldsymbol{q}_* = \boldsymbol{A}\boldsymbol{u}$, $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{Q}_{\boldsymbol{B}} = \boldsymbol{I}_D - \boldsymbol{B}\boldsymbol{B}^\top$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{P} + \zeta \boldsymbol{q}_* \boldsymbol{q}_*^\top$. Constants below are finite and positive. Unless uniformity is explicitly asserted, they may depend on the fixed target, dimensions, and a compact set of initializations with nonsingular support overlap. They do not depend on t sufficiently close to one.

B.1 Model specifications and deterministic initialization

For the predictor equation 18, set

$$(d_{\mathrm{OT},t}, \ell_{\mathrm{OT},t}) = (1, 0), \qquad d_{\mathrm{ind},t} = \sigma_t, \quad \ell_{\mathrm{ind},t} = \frac{2t-1}{\sigma_t^2}, \quad \sigma_t^2 = (1-t)^2 + t^2. \tag{85}$$

Writing $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{B}^\top$, the gradient flow for the induced Euclidean metric on the Stiefel manifold is

$$\dot{\boldsymbol{B}} = -(\boldsymbol{I}_D - \boldsymbol{P}_{\boldsymbol{B}})\nabla_{\boldsymbol{B}}\mathcal{J}_{C,t} - \boldsymbol{B}\,\mathrm{skew}(\boldsymbol{B}^\top \nabla_{\boldsymbol{B}}\mathcal{J}_{C,t}), \qquad \mathrm{skew}(\boldsymbol{M}) = \tfrac{1}{2}(\boldsymbol{M} - \boldsymbol{M}^\top), \tag{86}$$

with $\dot{\boldsymbol{K}} = -\nabla_{\boldsymbol{K}}\mathcal{J}_{C,t}$, $\dot{\boldsymbol{a}} = -\nabla_{\boldsymbol{a}}\mathcal{J}_{C,t}$, $\dot{\boldsymbol{w}} = -(\boldsymbol{I}_r - \boldsymbol{w}\boldsymbol{w}^\top)\nabla_{\boldsymbol{w}}\mathcal{J}_{C,t}$, and $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$. The dot denotes gradient-flow time s, while flow time t stays fixed. Every parameter block has unit learning rate.

Optimal losses and cubic correlations. In coordinates on the true support, let $\boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{X}_{t,C}$, $\boldsymbol{\Delta} = \boldsymbol{A}^\top \boldsymbol{U}_{t,C}$, and $\boldsymbol{L}_{C,t} = \mathbb{E}[\boldsymbol{\Delta}\boldsymbol{y}^\top](\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top])^{-1}$, $\boldsymbol{\eta} = \boldsymbol{\Delta} - \boldsymbol{L}_{C,t}\boldsymbol{y}$, $\boldsymbol{z} = \boldsymbol{y}/d_{C,t}$, and $V = \boldsymbol{u}^\top \boldsymbol{z}$. The best linear loss is $J_{\mathrm{lin},C,t} = \frac{1}{2}\mathbb{E}\|\boldsymbol{\eta}\|^2$. The cubic correlation after removing this linear fit is $\kappa_C(t) = \mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{\eta})\,\mathrm{He}_3(V)]$. For the aligned cubic feature, the squared norm after removing its linear projection is

$$D_{3,C,t} = \mathbb{E}[V^6] - \frac{\mathbb{E}[V^4]^2}{\mathbb{E}[V^2]}. \tag{87}$$

At the correct support, regression over the linear and cubic weights gives

$$J_{\mathrm{cub},C,t} = J_{\mathrm{lin},C,t} - \frac{\kappa_C(t)^2}{2D_{3,C,t}}. \tag{88}$$

Thus the threshold in equation 19 is exactly $\theta\kappa_C(t)^2/(2D_{3,C,t})$. Define

$$K_{\mathrm{ind}} = \mathbb{E}[W^4] - 3q_\rho^2, \qquad K_{\mathrm{OT}} = \frac{\mathbb{E}[W^4]}{q_\rho} - \mathbb{E}[SW^3]. \tag{89}$$

The endpoint correlations satisfy

$$\lim_{t\uparrow 1} \kappa_{\mathrm{OT}}(t) = K_{\mathrm{OT}}, \qquad \lim_{t\uparrow 1} \frac{\kappa_{\mathrm{ind}}(t)}{\varepsilon_t} = \frac{K_{\mathrm{ind}}}{q_\rho}. \tag{90}$$

Thus, if both constants are nonzero, $|\kappa_{\mathrm{OT}}(t)|$ is bounded above and below by positive constants near the endpoint, whereas $|\kappa_{\mathrm{ind}}(t)|$ is bounded above and below by positive multiples of ε_t. The explicit bounds are equation 74.

$K_{\mathrm{ind}} = 0$ makes the independent cubic correlation zero at every t, even if the nonlinear velocity is nonzero.

Effective initial alignment. Let $\boldsymbol{B}_\infty^{(0)}$ be the limiting basis of the auxiliary flow

$$\frac{d\bar{\boldsymbol{B}}}{d\xi} = (\boldsymbol{I}_D - \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top)\boldsymbol{\Sigma}_1\bar{\boldsymbol{B}}, \qquad \bar{\boldsymbol{B}}(0) = \boldsymbol{B}_0, \qquad m_0 = |\boldsymbol{q}_*^\top \boldsymbol{B}_\infty^{(0)}\boldsymbol{w}_0|. \tag{91}$$

This limit exists when $\boldsymbol{A}^\top\boldsymbol{B}_0$ is invertible. Lemma B.2 shows that the joint flow approaches this basis while $\boldsymbol{w}$ changes by a vanishing amount. Thus m_0 is the initial alignment for the limiting dynamics on the true support.

The transformation $(\boldsymbol{a}, \boldsymbol{w}) \mapsto (-\boldsymbol{a}, -\boldsymbol{w})$ preserves the predictor, the loss, and the gradient flow. Since $\boldsymbol{a}_0 = 0$, we may choose the sign of $\boldsymbol{w}_0$ so that $\boldsymbol{q}_*^\top \boldsymbol{B}_\infty^{(0)}\boldsymbol{w}_0 = m_0 > 0$. We use this orientation in the signed-alignment equations below. The final conclusions concern absolute alignment.

Proposition 2 (Deterministic initialization)**.** Under the assumptions of Theorem 2, initialize $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$. Fix $\theta \in (0,1)$ and $D \geq r \geq 2$. There exist $m_* > 0$ and $c_0, C_0 > 0$, depending only on H, ρ, θ, such that the following holds. If $\boldsymbol{A}^\top\boldsymbol{B}_0$ is invertible and $0 < m_0 \leq m_*$, there is $t_0 < 1$, which may also depend on $D, r, \boldsymbol{B}_0, \boldsymbol{w}_0$, such that for $t_0 \leq t < 1$ the full joint flow satisfies

$$\frac{c_0}{\kappa_C(t)^2 m_0^4} \leq T_{C,\theta}^{\mathrm{full}}(t) \leq \frac{C_0}{\kappa_C(t)^2 m_0^4}, \qquad C \in \{\mathrm{OT}, \mathrm{ind}\}. \tag{92}$$

Moreover, as $s \to \infty$,

$$\boldsymbol{B}_s\boldsymbol{B}_s^\top \to \boldsymbol{P}, \qquad |\boldsymbol{q}_*^\top \boldsymbol{B}_s\boldsymbol{w}_s| \to 1, \qquad \mathcal{J}_{C,t}(s) \to J_{\mathrm{lin},C,t} - \frac{\kappa_C(t)^2}{2D_{3,C,t}}. \tag{93}$$

The endpoint threshold can be chosen uniformly on compact sets where $\boldsymbol{A}^\top\boldsymbol{B}_0$ is nonsingular and m_0 is bounded away from zero. Applying these bounds to random initialization will give equation 22.

The proof is given in Section B.7.

B.2 The loss decomposition and a sufficient basin condition

With $\boldsymbol{y} = \boldsymbol{B}^\top\boldsymbol{X}_t$, put $\boldsymbol{g}(\boldsymbol{y}) = (\ell\boldsymbol{I}_r + \boldsymbol{K})\boldsymbol{y} + \boldsymbol{a}\,\mathrm{He}_3(\boldsymbol{w}^\top\boldsymbol{y}/d)$. The interpolation identity $\boldsymbol{U}_t + \boldsymbol{X}_t/\alpha = \boldsymbol{X}_1/\alpha$ gives the exact orthogonal decomposition

$$\mathcal{J} = \frac{\mathcal{G}(\boldsymbol{B})}{\alpha^2} + \mathcal{F}_t(\boldsymbol{B}, \boldsymbol{K}, \boldsymbol{a}, \boldsymbol{w}), \qquad \mathcal{G}(\boldsymbol{B}) = \tfrac{1}{2}\,\mathrm{tr}(\boldsymbol{Q}_{\boldsymbol{B}}\boldsymbol{\Sigma}_1), \qquad \mathcal{F}_t = \tfrac{1}{2}\mathbb{E}\|\boldsymbol{g}(\boldsymbol{B}^\top\boldsymbol{X}_t) - \boldsymbol{B}^\top\boldsymbol{U}_t\|^2. \tag{94}$$

The first term is the squared normal residual, and the second is the squared tangential residual. Both are nonnegative. Sixth moments make $\mathcal{F}_t$ and its parameter derivatives bounded on bounded weight

sets, uniformly near $t = 1$: the predictor is polynomial of degree three in the input, and the loss involves moments of degree at most six.

At $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{P}$, the normal Gaussian coordinate of $\boldsymbol{X}_t$ is independent of the entire tangential endpoint pair and has mean zero. Differentiating $\mathcal{F}_t$ in a horizontal direction and averaging this coordinate therefore gives

$$\boldsymbol{Q}_{\boldsymbol{B}} \nabla_{\boldsymbol{B}} \mathcal{F}_t = 0 \quad \text{if } \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{P}, \qquad \|\boldsymbol{Q}_{\boldsymbol{B}} \nabla_{\boldsymbol{B}} \mathcal{F}_t\|_F \le C_R \|\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{P}\|_F \tag{95}$$

near the true support when $\|\boldsymbol{K}\|_F + \|\boldsymbol{a}\| \le R$. For example, write the predictor as $-\boldsymbol{x}/\alpha + \boldsymbol{B}\boldsymbol{f}(\boldsymbol{B}^\top \boldsymbol{x})$. At $\boldsymbol{B} = \boldsymbol{A}$, its derivative in $\boldsymbol{A}_\perp \boldsymbol{T}$ is $\boldsymbol{A}_\perp \boldsymbol{T} \boldsymbol{f}(\boldsymbol{y}) + \alpha \boldsymbol{A} D\boldsymbol{f}(\boldsymbol{y}) \boldsymbol{T}^\top \boldsymbol{N}$. The residual is tangential, so its inner product with the first term is zero. The inner product with the second term averages to zero. Thus the horizontal derivative of $\mathcal{J}$ is zero. The geometry term satisfies $\boldsymbol{Q}_{\boldsymbol{A}} \nabla_{\boldsymbol{B}} \mathcal{G}(\boldsymbol{A}) = -\boldsymbol{Q}_{\boldsymbol{A}} \boldsymbol{\Sigma}_1 \boldsymbol{A} = 0$ because $\mathrm{range}(\boldsymbol{\Sigma}_1) \subseteq \mathrm{span}(\boldsymbol{A})$. Subtracting its derivative in equation 94 gives the first identity for $\mathcal{F}_t$. The second follows from bounded derivatives.

Let $E_{\boldsymbol{P}} = r - \mathrm{tr}(\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{P})$ and $\boldsymbol{C}_{\boldsymbol{B}} = \boldsymbol{B}^\top \boldsymbol{P} \boldsymbol{B}$. The geometry part of the support update gives

$$\dot{E}_{\boldsymbol{P}} \le -2\alpha^{-2} \,\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{B}} - \boldsymbol{C}_{\boldsymbol{B}}^2). \tag{96}$$

The covariance-spike contribution has the same sign: its nonnegative factor is $\boldsymbol{q}_*^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{q}_* - \boldsymbol{q}_*^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{P} \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{q}_*$. If the eigenvalues of $\boldsymbol{C}_{\boldsymbol{B}}$ exceed $1/2$, then $\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{B}} - \boldsymbol{C}_{\boldsymbol{B}}^2) \ge E_{\boldsymbol{P}}/2$. Using equation 95, the remaining contribution is bounded in absolute value by $C_R E_{\boldsymbol{P}}$. Basis rotations do not change $E_{\boldsymbol{P}}$. Hence

$$\dot{E}_{\boldsymbol{P}} \le -(\alpha^{-2} - C_R) E_{\boldsymbol{P}} \tag{97}$$

in this neighborhood.

Lemma B.1 (A basin defined by the linear loss)**.** *For t sufficiently close to one, any full joint trajectory that reaches $\mathcal{J} < J_{\mathrm{lin},C,t}$ satisfies equation 93.*

Proof. At the correct support the exact best linear losses are

$$J_{\mathrm{lin,OT},t} = \frac{\zeta}{2(1 + t^2\zeta)}, \qquad J_{\mathrm{lin,ind},t} = \frac{1}{2}\left[\frac{r-1}{\sigma_t^2} + \frac{\nu}{\alpha^2 + t^2\nu}\right]. \tag{98}$$

They are bounded near one for fixed r. Loss decreases along the product gradient flow, so equation 94 gives $E_{\boldsymbol{P}} \le 2\alpha^2 J_{\mathrm{lin},C,t}$ throughout the strict sublevel set. This places the support in the neighborhood used above. The Gram matrix of the linear and cubic features has a positive uniform lower bound there: at the true support its Schur complement is equation 78, and continuity extends the bound to a neighborhood, uniformly over $\boldsymbol{w}$. The loss bound and bounded label second moments then bound $\boldsymbol{K}, \boldsymbol{a}$. For small enough α, equation 97 makes $\boldsymbol{B}\boldsymbol{B}^\top \to \boldsymbol{P}$ exponentially. The trajectory lies in a compact parameter set, and gradient-flow dissipation implies that every limit point is stationary. At the correct support, minimizing over the weights leaves $-\kappa^2 m^6/(2\widetilde{M}(m))$, as in equation 80. Its strict monotonicity excludes stationary alignments other than $m = 0$ and $|m| = 1$. The former have loss $J_{\mathrm{lin},C,t}$, excluded by the strict sublevel condition. Every limit point therefore represents the aligned optimum, proving all three limits in equation 93. The basis itself need not be unique. □

B.3 Rotations of the support basis

Factor $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{O}$ along the trajectory, where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{C}^\top \dot{\boldsymbol{C}} = 0$. In this frame the internal parameters are $\bar{\boldsymbol{K}} = \boldsymbol{O}\boldsymbol{K}\boldsymbol{O}^\top$, $\bar{\boldsymbol{a}} = \boldsymbol{O}\boldsymbol{a}$, and $\bar{\boldsymbol{w}} = \boldsymbol{O}\boldsymbol{w}$. The condition $\boldsymbol{C}^\top \dot{\boldsymbol{C}} = 0$ removes rotations within the span of $\boldsymbol{C}$. We henceforth omit bars on the transformed parameters. Define $\boldsymbol{Q}_{\boldsymbol{C}} = \boldsymbol{I}_D - \boldsymbol{C}\boldsymbol{C}^\top$ and $[\boldsymbol{\Omega}, \boldsymbol{K}] = \boldsymbol{\Omega}\boldsymbol{K} - \boldsymbol{K}\boldsymbol{\Omega}$. For skew-symmetric $\boldsymbol{\Omega}$, set

$$\begin{aligned} R_\Theta \boldsymbol{\Omega} &= ([\boldsymbol{\Omega}, \boldsymbol{K}], \boldsymbol{\Omega}\boldsymbol{a}, \boldsymbol{\Omega}\boldsymbol{w}), \qquad \Theta = (\boldsymbol{K}, \boldsymbol{a}, \boldsymbol{w}), \\ R_\Theta^* \boldsymbol{G} &= \mathrm{skew}(\boldsymbol{G}_{\boldsymbol{K}} \boldsymbol{K}^\top - \boldsymbol{K}^\top \boldsymbol{G}_{\boldsymbol{K}} + \boldsymbol{g}_{\boldsymbol{a}} \boldsymbol{a}^\top + \boldsymbol{g}_{\boldsymbol{w}} \boldsymbol{w}^\top). \end{aligned} \tag{99}$$

The map R_Θ describes how an infinitesimal basis rotation changes the internal parameters. Its adjoint uses Frobenius inner products for matrices. Invariance of the predictor under $(\boldsymbol{B}, \boldsymbol{K}, \boldsymbol{a}, \boldsymbol{w}) \mapsto (\boldsymbol{B}\boldsymbol{O}, \boldsymbol{O}^\top \boldsymbol{K}\boldsymbol{O}, \boldsymbol{O}^\top \boldsymbol{a}, \boldsymbol{O}^\top \boldsymbol{w})$ yields $\mathrm{skew}(\boldsymbol{C}^\top \nabla_{\boldsymbol{C}} \mathcal{J}) = R_\Theta^* \boldsymbol{G}$, where $\boldsymbol{G} = \mathrm{grad}_\Theta \mathcal{J}$. The gradient flow in these coordinates is

$$\dot{\boldsymbol{C}} = \alpha^{-2} \boldsymbol{Q}_{\boldsymbol{C}} \boldsymbol{\Sigma}_1 \boldsymbol{C} - \boldsymbol{Q}_{\boldsymbol{C}} \nabla_{\boldsymbol{C}} \mathcal{F}_t, \qquad \dot{\Theta} = -\mathsf{A}_\Theta \boldsymbol{G}, \qquad \mathsf{A}_\Theta = \boldsymbol{I} + R_\Theta R_\Theta^*. \tag{100}$$

On bounded weight sets,

$$\boldsymbol{I} \preceq \mathsf{A}_\Theta \preceq (2+4\|\boldsymbol{K}\|_{\mathrm{op}}^2+\|\boldsymbol{a}\|^2)\boldsymbol{I}, \qquad -\dot{\mathcal{J}} = \|\boldsymbol{Q}_{\boldsymbol{C}}\nabla_{\boldsymbol{C}}\mathcal{J}\|_F^2 + \|\boldsymbol{G}\|^2 + \|R_\Theta^*\boldsymbol{G}\|_F^2. \tag{101}$$

Since A_Θ is positive definite, the stationary points on the true support are exactly the zeros of $\boldsymbol{G}$.

B.4 Support contraction and the limiting basis

Lemma B.2 (Initial support contraction)**.** *Fix a compact set of $(\boldsymbol{B}_0, \boldsymbol{w}_0)$ on which the least singular value of $\boldsymbol{A}^\top\boldsymbol{B}_0$ is positive, and set $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$. For any fixed $p > 2$ there are $C_p, c > 0$ and, for small enough α, a time $s_g \le C_p\alpha^2\log(1/\alpha)$ such that*

$$\begin{aligned} \|\boldsymbol{C}_{s_g}\boldsymbol{C}_{s_g}^\top - \boldsymbol{P}\|_F &\le C_p\alpha^p, \\ \|\boldsymbol{K}_{s_g}\|_F + \|\boldsymbol{a}_{s_g}\| + \|\boldsymbol{w}_{s_g} - \boldsymbol{w}_0\| &\le C_p\alpha^2\log(1/\alpha), \\ \|\boldsymbol{C}_{s_g} - \boldsymbol{B}_\infty^{(0)}\|_F &\le C_p\alpha^2. \end{aligned} \tag{102}$$

If the weights remain in a fixed bounded set afterward, then for $\boldsymbol{C}_* = \Pi(\boldsymbol{C}_{s_g})$, a basis of the true support,

$$\|\boldsymbol{C}_s - \boldsymbol{C}_*\|_F \le C_p\left[\alpha^p e^{-c(s-s_g)/\alpha^2} + \alpha^{p+2}\right] \quad (s \ge s_g). \tag{103}$$

Here $\Pi(\boldsymbol{C})$ is the limiting basis of the pure geometry flow starting at $\boldsymbol{C}$.

Proof. Use fast time $\xi = s/\alpha^2$. The support equation is its pure geometry flow plus $-\alpha^2\boldsymbol{Q}_{\boldsymbol{C}}\nabla_{\boldsymbol{C}}\mathcal{F}_t$; the internal equation is $-\alpha^2\mathsf{A}_\Theta\,\mathrm{grad}_\Theta\,\mathcal{F}_t$. Both perturbations are bounded on bounded weight sets. For the pure geometry flow, graph coordinates

$$\boldsymbol{Z} = \boldsymbol{A}_\perp^\top\boldsymbol{C}(\boldsymbol{A}^\top\boldsymbol{C})^{-1}, \qquad \mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z}) = (\boldsymbol{A}+\boldsymbol{A}_\perp\boldsymbol{Z})(\boldsymbol{I}_r + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1/2}, \qquad \boldsymbol{C} = \mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z})\boldsymbol{O} \tag{104}$$

satisfy $\boldsymbol{Z}' = -\boldsymbol{Z}\boldsymbol{\Lambda}$, with $\boldsymbol{\Lambda} = \boldsymbol{I}_r + \zeta\boldsymbol{u}\boldsymbol{u}^\top$, and $\boldsymbol{O}' = -\mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z})^\top\mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z})'\boldsymbol{O}$. Explicitly, $\boldsymbol{Z}(\xi) = \boldsymbol{Z}(0)e^{-\boldsymbol{\Lambda}\xi}$. The rotation coefficient $\mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z})^\top D\mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z})[-\boldsymbol{Z}\boldsymbol{\Lambda}]$ is smooth and vanishes to second order at $\boldsymbol{Z} = 0$. On compact sets of initial graph coordinates it and each of its derivatives with respect to those coordinates are bounded by $C_j e^{-2\xi}$, where j is the derivative order. The integral equation for $\boldsymbol{O}$, differentiated with respect to the initial coordinates, therefore gives a smooth limit $\boldsymbol{O}_\infty$. Hence $\Pi(\boldsymbol{C}) = \boldsymbol{A}\boldsymbol{O}_\infty$ is smooth. Near $\boldsymbol{Z}(0) = 0$, the same integral estimate gives $\|\boldsymbol{O}_\infty - \boldsymbol{O}(0)\|_F \le C\|\boldsymbol{Z}(0)\|_F^2$. Together with $\|\mathsf{F}_{\mathrm{gr}}(\boldsymbol{Z}) - \boldsymbol{A}\|_F \le C\|\boldsymbol{Z}\|_F$ and the local equivalence of $\|\boldsymbol{Z}\|_F$ and $\|\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{P}\|_F$, this proves the second bound below. The first follows because the pure geometry flow does not change its limiting basis:

$$D\Pi(\boldsymbol{C})[\boldsymbol{Q}_{\boldsymbol{C}}\boldsymbol{\Sigma}_1\boldsymbol{C}] = 0, \qquad \|\boldsymbol{C} - \Pi(\boldsymbol{C})\|_F \le C\|\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{P}\|_F \tag{105}$$

near the true support.

Continuous dependence on a fixed fast-time interval first brings the perturbed flow into this neighborhood. The bound equation 97 then gives exponential contraction in ξ, uniformly for small α. Continuing for at most $C_p\log(1/\alpha)$ fast time makes the support error at most $C_p\alpha^p$. The weight and direction displacements on that interval are bounded by $C_p\alpha^2\log(1/\alpha)$. We therefore choose a fixed weight ball and reduce α until the displacement bounds keep the trajectory inside it, completing the continuation argument. Along the actual flow,

$$\frac{d}{d\xi}\Pi(\boldsymbol{C}) = -\alpha^2 D\Pi(\boldsymbol{C})[\boldsymbol{Q}_{\boldsymbol{C}}\nabla_{\boldsymbol{C}}\mathcal{F}_t]. \tag{106}$$

The integrand without α^2 is bounded before local entrance and bounded by $C\|\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{P}\|_F$ afterward. Its integral is bounded. This proves the third bound in equation 102, using equation 105. After s_g, the support error is at most $C_p\alpha^p e^{-c(s-s_g)/\alpha^2}$. Its integral with respect to s is at most $C_p\alpha^{p+2}$, which bounds the remaining drift of $\Pi(\boldsymbol{C}_s)$ and proves equation 103. Every constant can be chosen uniformly on the specified compact set. □

B.5 WEIGHT DYNAMICS UNDER INDEPENDENT COUPLING

Let $\kappa = \kappa_{\mathrm{ind}}(t)$. Take $p = 6$ in Lemma B.2 and use $\boldsymbol{C}_*$ as the reference support basis. Write $\boldsymbol{u}_* := \boldsymbol{C}_*^\top \boldsymbol{q}_*$. In this basis, use $\boldsymbol{y} = \boldsymbol{C}_*^\top \boldsymbol{X}_{t,C}$, $\boldsymbol{z} = \boldsymbol{y}/d$, and $V = \boldsymbol{u}_*^\top \boldsymbol{z}$ for the correct-support problem. Let $\boldsymbol{L} = \ell \boldsymbol{I}_r + \lambda \boldsymbol{u}_* \boldsymbol{u}_*^\top$, and set

$$\boldsymbol{Z}_h = \frac{d(\boldsymbol{K} - \lambda \boldsymbol{u}_* \boldsymbol{u}_*^\top)}{\kappa}, \qquad \boldsymbol{b} = \frac{\boldsymbol{a}}{\kappa}, \qquad \boldsymbol{Q} = (\boldsymbol{Z}_h, \boldsymbol{b}), \qquad \tau = \kappa^2 s. \tag{107}$$

Here $\boldsymbol{Z}_h \in \mathbb{R}^{r \times r}$ and $\boldsymbol{b} \in \mathbb{R}^r$. Under independent coupling, equation 74 and the exact formula for λ in Appendix A.5 give constants $c_0, C_0 > 0$ such that

$$c_0 \alpha \le |\kappa| \le C_0 \alpha, \qquad |\lambda| \le C_0 |\kappa|, \qquad d \ge c_0 \tag{108}$$

near one. Thus $\boldsymbol{Q}(\tau_g)$ is bounded, where $\tau_g = \kappa^2 s_g$.

On the true support, the normalized loss is

$$\begin{gathered} E_t(\boldsymbol{Q}, \boldsymbol{w}) = \tfrac{1}{2}\mathbb{E}\|\boldsymbol{Z}_h \boldsymbol{z} + \boldsymbol{b}\phi_{\boldsymbol{w}}(\boldsymbol{z})\|^2 - m^3 \boldsymbol{u}_*^\top \boldsymbol{b}, \qquad J - J_{\mathrm{lin}} = \kappa^2 E_t, \\ \phi_{\boldsymbol{w}}(\boldsymbol{z}) = \mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z}), \qquad m = \boldsymbol{u}_*^\top \boldsymbol{w}. \end{gathered} \tag{109}$$

Its Hessian with respect to $\boldsymbol{Q}$ consists of copies of

$$\boldsymbol{\Gamma}_t(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{z}} & \boldsymbol{c}(\boldsymbol{w}) \\ \boldsymbol{c}(\boldsymbol{w})^\top & M_t(m) \end{pmatrix} \succeq g_0 \boldsymbol{I}, \quad \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{I}_r + (\mathbb{E}[V^2] - 1)\boldsymbol{u}_* \boldsymbol{u}_*^\top, \quad \boldsymbol{c}(\boldsymbol{w}) = \mathbb{E}[\boldsymbol{z}\phi_{\boldsymbol{w}}(\boldsymbol{z})]. \tag{110}$$

The lower bound follows from the positive Schur complement equation 78, uniformly over $\boldsymbol{w}$ and near-endpoint moments. The unique minimizing weights are

$$\boldsymbol{b}^*(\boldsymbol{w}) = \frac{m^3}{\widetilde{M}_t(m)} \boldsymbol{u}_*, \qquad \boldsymbol{Z}_h^*(\boldsymbol{w}) = -\boldsymbol{b}^*(\boldsymbol{w}) \boldsymbol{c}(\boldsymbol{w})^\top \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}, \qquad \min_{\boldsymbol{Q}} E_t = -\frac{m^6}{2\widetilde{M}_t(m)}. \tag{111}$$

Their derivatives on the sphere are bounded uniformly.

Let $\mu_j^* = \mathbb{E}[W^j]$, $D_* = \mu_6^* - (\mu_4^*)^2/\nu$, and define

$$\begin{aligned} \mathcal{M}_*(m) &= 6(1-m^2)^3 + 18\nu m^2 (1-m^2)^2 + 9(\mu_4^* - \nu^2) m^4 (1-m^2) + D_* m^6, \\ F_*(m) &= \frac{(1-m^2) m^5 [6\mathcal{M}_*(m) - m\mathcal{M}_*'(m)]}{2\mathcal{M}_*(m)^2}. \end{aligned} \tag{112}$$

Lemma B.3 (Independent-coupling limit)**.** *Under the assumptions of Proposition 2, use $s_g, \boldsymbol{C}_*$ from Lemma B.2 with $p = 6$. Set $\tau_g = \kappa^2 s_g$, $\boldsymbol{Q}^*(\boldsymbol{w}) = (\boldsymbol{Z}_h^*(\boldsymbol{w}), \boldsymbol{b}^*(\boldsymbol{w}))$, and $\boldsymbol{e}(\tau) = \boldsymbol{Q}(\tau) - \boldsymbol{Q}^*(\boldsymbol{w}(\tau))$. For every fixed $\tau_{\max} > 0$ and t sufficiently close to one, $\boldsymbol{Q}$ is bounded on $[\tau_g, \tau_{\max}]$, and*

$$\begin{aligned} \|\boldsymbol{e}(\tau)\| &\le C e^{-c(\tau - \tau_g)/\kappa^2} \|\boldsymbol{e}(\tau_g)\| + C\kappa^2 + C\alpha^5, \\ \int_{\tau_g}^{\tau_{\max}} \|\boldsymbol{e}(\tau)\| \, d\tau &\le C\{\kappa^2(1 + \tau_{\max}) + \alpha^5 \tau_{\max}\}. \end{aligned} \tag{113}$$

Put $\boldsymbol{u}_\infty = (\boldsymbol{B}_\infty^{(0)})^\top \boldsymbol{q}_*$ and $g_*(\boldsymbol{w}) = (\boldsymbol{u}_\infty^\top \boldsymbol{w})^6 / [2\mathcal{M}_*(\boldsymbol{u}_\infty^\top \boldsymbol{w})]$. Let $\boldsymbol{w}_*$ solve $\boldsymbol{w}_*' = \frac{3}{2} \operatorname{grad} g_*(\boldsymbol{w}_*)$ with $\boldsymbol{w}_*(0) = \boldsymbol{w}_0$. Its alignment $m = \boldsymbol{u}_\infty^\top \boldsymbol{w}_*$ satisfies

$$m'(\tau) = \tfrac{3}{2} F_*(m(\tau)), \qquad m(0) = m_0. \tag{114}$$

The transported direction obeys

$$\sup_{\tau_g \le \tau \le \tau_{\max}} \|\boldsymbol{w}(\tau) - \boldsymbol{w}_*(\tau)\| \le C \left[\alpha^2 \log(1/\alpha) + \kappa^2(1 + \tau_{\max}) + \varepsilon_t^2 \tau_{\max} + \alpha^5 \tau_{\max}\right] e^{L\tau_{\max}}. \tag{115}$$

The constants $C, c, L > 0$ may depend on $\tau_{\max}$, the target, the dimensions, and the compact initialization set, but not on t near one.

Proof. Split the rotation action into its weight and direction parts. For bounded $\boldsymbol{Q}$, the weights $\boldsymbol{K}, \boldsymbol{a}$ have norm at most $C|\kappa|$, so $\|R_h\| \le C|\kappa|$, whereas $R_{\boldsymbol{w}} R_{\boldsymbol{w}}^* \boldsymbol{g} = \frac{1}{2}\boldsymbol{g}$ for $\boldsymbol{g} \perp \boldsymbol{w}$. Let $\mathsf{S}_h = \mathrm{diag}(d\boldsymbol{I}, \boldsymbol{I})$ on the two weight blocks and $\widehat{R}_h = R_h/\kappa$. On the correct support, write $\boldsymbol{e}_{\boldsymbol{Q}} = \nabla_{\boldsymbol{Q}} E_t$ and $\boldsymbol{g}_{\boldsymbol{w}} = \mathrm{grad}_{\boldsymbol{w}} E_t$. The gradients with respect to $(\boldsymbol{K}, \boldsymbol{a})$ and $\boldsymbol{w}$ are $\kappa \mathsf{S}_h \boldsymbol{e}_{\boldsymbol{Q}}$ and $\kappa^2 \boldsymbol{g}_{\boldsymbol{w}}$, respectively. Substitution into equation 100 gives the exact equations

$$\begin{aligned} \boldsymbol{Q}' &= -\kappa^{-2} \mathsf{S}_h^2 \boldsymbol{e}_{\boldsymbol{Q}} - \mathsf{S}_h \widehat{R}_h (\widehat{R}_h^* \mathsf{S}_h \boldsymbol{e}_{\boldsymbol{Q}} + R_{\boldsymbol{w}}^* \boldsymbol{g}_{\boldsymbol{w}}), \\ \boldsymbol{w}' &= -\tfrac{3}{2} \boldsymbol{g}_{\boldsymbol{w}} - R_{\boldsymbol{w}} \widehat{R}_h^* \mathsf{S}_h \boldsymbol{e}_{\boldsymbol{Q}}. \end{aligned} \tag{116}$$

On bounded $\boldsymbol{Q}$ sets, $\widehat{R}_h$, $\boldsymbol{e}_{\boldsymbol{Q}}$, and $\boldsymbol{g}_{\boldsymbol{w}}$ are bounded uniformly near the endpoint, since λ/κ and d^{-1} are bounded. Adding the support perturbation therefore gives

$$\begin{aligned} &\boldsymbol{Q}' = -\kappa^{-2} \mathsf{D}_h \nabla_{\boldsymbol{Q}} E_t + \boldsymbol{f}_t, \qquad \mathsf{D}_h = \mathrm{diag}(d^2 \boldsymbol{I}, \boldsymbol{I}), \\ &\boldsymbol{w}' = -\tfrac{3}{2} \,\mathrm{grad}_{\boldsymbol{w}} E_t + \boldsymbol{h}_t, \\ &\|\boldsymbol{f}_t\| \le C_R \left(1 + \frac{\Delta_{\boldsymbol{B}}}{|\kappa|^3}\right), \qquad \|\boldsymbol{h}_t\| \le C_R \left(\|\nabla_{\boldsymbol{Q}} E_t\| + \frac{\Delta_{\boldsymbol{B}}}{|\kappa|}\right), \quad \Delta_{\boldsymbol{B}} = \|\boldsymbol{C}_s - \boldsymbol{C}_*\|_F. \end{aligned} \tag{117}$$

Primes denote τ derivatives. A support displacement $\Delta_{\boldsymbol{B}}$ changes the gradient with respect to $(\boldsymbol{K}, \boldsymbol{a})$ by at most $C_R \Delta_{\boldsymbol{B}}$. Dividing by κ for the weights and by κ^2 for time gives the first denominator. A direction gradient contains $\boldsymbol{a}$, and the rotation term contains $\boldsymbol{K}$ or $\boldsymbol{a}$, so its support error before rescaling time is at most $C_R|\kappa|\Delta_{\boldsymbol{B}}$, giving the second.

Fix a slow-time interval $[\tau_g, \tau_{\max}]$. Positive definiteness of $\boldsymbol{\Gamma}_t$ gives

$$\frac{d}{d\tau} \frac{1}{2} \|\boldsymbol{Q}\|^2_{\mathsf{D}_h^{-1}} \le \kappa^{-2}(-c\|\boldsymbol{Q}\|^2 + C\|\boldsymbol{Q}\|) + C_R \|\boldsymbol{Q}\| \left(1 + \Delta_{\boldsymbol{B}}/|\kappa|^3\right). \tag{118}$$

Choose a fixed large ball containing $\boldsymbol{Q}(\tau_g)$. On its boundary the first term points inward. The last terms are dominated for small α, since equation 103 gives $\Delta_{\boldsymbol{B}}/|\kappa|^3 \le C\alpha^3$. Thus $\boldsymbol{Q}$ stays in the ball, and the support estimate extends to $[\tau_g, \tau_{\max}]$.

Tracking the optimal weights. Write $\boldsymbol{Q}^*(\boldsymbol{w}) = (\boldsymbol{Z}_h^*(\boldsymbol{w}), \boldsymbol{b}^*(\boldsymbol{w}))$. For $\boldsymbol{e} = \boldsymbol{Q} - \boldsymbol{Q}^*(\boldsymbol{w})$, the same weighted norm and bounded derivative of $\boldsymbol{Q}^*$ give the upper Dini derivative inequality

$$\frac{d^+}{d\tau} \|\boldsymbol{e}\|_{\mathsf{D}_h^{-1}} \le -\frac{c}{\kappa^2} \|\boldsymbol{e}\|_{\mathsf{D}_h^{-1}} + C \left(1 + \Delta_{\boldsymbol{B}}/|\kappa|^3\right). \tag{119}$$

Variation of constants gives equation 113. The powers of α use $p = 6$ and equation 108.

The limiting direction equation. The function F_* is the spherical gradient-ascent speed of the gain:

$$F_*(m) = (1 - m^2) \frac{d}{dm} \left(\frac{m^6}{2\mathcal{M}_*(m)}\right). \tag{120}$$

The bracket in equation 112 is positive on $[0, 1)$, by the positive-coefficient identity used in equation 78. The even moments of $V = (W + \varepsilon_t N)/\sqrt{1 + \varepsilon_t^2}$ differ from those of W by at most $C\varepsilon_t^2$ through degree six. Using equation 113 in the angular equation and applying Gronwall on $[\tau_g, \tau_{\max}]$ gives equation 115. Lemma B.2 bounds the initial direction error by $C\alpha^2 \log(1/\alpha)$ and gives $\|\boldsymbol{u}_* - \boldsymbol{u}_\infty\| \le C\alpha^2$. The latter contributes $C\alpha^2 \tau_{\max}$, which is absorbed into $C\kappa^2 \tau_{\max}$ by equation 108. Taking the inner product of the limiting sphere equation with $\boldsymbol{u}_\infty$ gives equation 114. □

B.6 The OT endpoint dynamics near zero alignment

Use the limiting support coordinates $\boldsymbol{u}_\infty = (\boldsymbol{B}_\infty^{(0)})^\top \boldsymbol{q}_*$. At the OT endpoint, put $\kappa_* = K_{\mathrm{OT}} \ne 0$ and $\lambda_* = (\nu - 1)/\nu$. The tangential regression problem has $d = 1$, $V = W$, and best linear predictor $\lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top$. The equation for $(\boldsymbol{K}, \boldsymbol{a}, \boldsymbol{w})$ in equation 100 therefore has a regular limit on the true support as $t \uparrow 1$. Write $T_{\mathrm{OT},\theta}^{*,\mathrm{full}}(m_0)$ for its hitting time from zero weights and alignment m_0.

Lemma B.4 (An invariant two-dimensional subsystem)**.** *Let $\mathcal{V} = \mathrm{span}\{\boldsymbol{u}_\infty, \boldsymbol{w}_0\}$ and let $\boldsymbol{P}_\mathcal{V}$ be its orthogonal projector. The endpoint internal flow preserves*

$$\boldsymbol{K} = \boldsymbol{P}_\mathcal{V}\boldsymbol{K}\boldsymbol{P}_\mathcal{V}, \qquad \boldsymbol{a} = \boldsymbol{P}_\mathcal{V}\boldsymbol{a}, \qquad \boldsymbol{w} = \boldsymbol{P}_\mathcal{V}\boldsymbol{w}. \tag{121}$$

For $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$, all coordinates outside this plane remain zero.

Proof. Set $\boldsymbol{E}_K = \boldsymbol{K} - \lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_r + \zeta \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top$. The exact endpoint loss is

$$J - J_{\mathrm{lin}} = \tfrac{1}{2}\,\mathrm{tr}(\boldsymbol{E}_K \boldsymbol{\Sigma} \boldsymbol{E}_K^\top) + \boldsymbol{a}^\top \boldsymbol{E}_K \boldsymbol{c}(\boldsymbol{w}) + \tfrac{1}{2} M(m)\|\boldsymbol{a}\|^2 - \kappa_* m^3 \boldsymbol{u}_\infty^\top \boldsymbol{a}, \quad m = \boldsymbol{u}_\infty^\top \boldsymbol{w}, \tag{122}$$

where $M(m) = \mathbb{E}[\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})^2]$ and

$$\boldsymbol{c}(\boldsymbol{w}) = \mathbb{E}[\boldsymbol{z}\,\mathrm{He}_3(\boldsymbol{w}^\top \boldsymbol{z})] = (\mu_4^* - 3\nu) m^3 \boldsymbol{u}_\infty + 3\zeta m^2 (\boldsymbol{w} - m\boldsymbol{u}_\infty). \tag{123}$$

Thus $\boldsymbol{G}_{\boldsymbol{K}} = \boldsymbol{E}_K \boldsymbol{\Sigma} + \boldsymbol{a}\boldsymbol{c}(\boldsymbol{w})^\top$ and $\boldsymbol{g}_{\boldsymbol{a}} = \boldsymbol{E}_K \boldsymbol{c}(\boldsymbol{w}) + M(m)\boldsymbol{a} - \kappa_* m^3 \boldsymbol{u}_\infty$. On the stated set, these gradients have only the indicated plane blocks. For the angular gradient, $\boldsymbol{E}_K^\top \boldsymbol{a} \in \mathcal{V}$ and the displayed formula for $\boldsymbol{c}(\boldsymbol{w})$ show that the spherical gradient also lies in $\mathcal{V}$. The adjoint in equation 99 consequently satisfies $R_\Theta^* \boldsymbol{G} = \boldsymbol{P}_\mathcal{V}(R_\Theta^* \boldsymbol{G})\boldsymbol{P}_\mathcal{V}$. Both $\boldsymbol{G}$ and $R_\Theta R_\Theta^* \boldsymbol{G}$ are tangent to the stated set. Uniqueness of the ODE solution proves invariance. □

Slow motion near the orthogonal equilibrium. Choose $\boldsymbol{e} \perp \boldsymbol{u}_\infty$ in $\mathcal{V}$ and write $\boldsymbol{w} = \boldsymbol{u}_\infty \sin\varphi + \boldsymbol{e}\cos\varphi$, with $m = \sin\varphi$. The subsystem in Lemma B.4 has six weight coordinates, denoted by $\boldsymbol{h}$, and one angular coordinate.

Lemma B.5 (Angular speed on the center manifold)**.** *The OT endpoint flow has a local C^7 center manifold $\boldsymbol{h} = \boldsymbol{h}_c(\varphi)$ at the equilibrium*

$$\boldsymbol{K} = \lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top, \qquad \boldsymbol{a} = 0, \qquad \varphi = 0, \tag{124}$$

tangent to the angular axis. On this manifold, $m = \sin\varphi$ satisfies

$$\left| \dot{m} - \frac{c_* \kappa_*^2}{2} m^5 \right| \le C|m|^6, \tag{125}$$

where $c_* = 1 + 1/[2(1+\lambda_*^2)]$ and C is independent of m.

Proof. At this equilibrium, the weight Hessian is positive definite and the angular Hessian is zero. Indeed, the weight Gram matrix consists of the positive linear covariance and the independent feature $\mathrm{He}_3(G)$ of norm squared six. The reduced linearization therefore has six strictly stable eigenvalues and one simple zero eigenvalue. The center-manifold theorem gives a local invariant graph $\boldsymbol{h} = \boldsymbol{h}_c(\varphi)$ tangent to the angular axis (Carr, 1981, Chapters 1–2).

Let $\boldsymbol{h}^*(\varphi)$ minimize the loss over the weights at fixed angle, and write $J_{\mathrm{prof}}(\varphi) = J(\boldsymbol{h}^*(\varphi), \varphi)$. We also express this minimized loss as a function of $m = \sin\varphi$. The exact regression formulas equation 111 show that for $|m|$ small,

$$\begin{aligned} &\|\boldsymbol{a}^*(m) - \kappa_* m^3 \boldsymbol{u}_\infty/6\| \le C|m|^5, \qquad \|\boldsymbol{K}^*(m) - \lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top\|_F \le C|m|^5, \\ &\left| J_{\mathrm{prof}}(m) - J_{\mathrm{lin}} + \frac{\kappa_*^2 m^6}{12} \right| \le C|m|^8, \qquad \|\partial_\varphi \boldsymbol{h}^*(\varphi)\| \le C|\varphi|^2. \end{aligned} \tag{126}$$

Here and below the OT endpoint is fixed, so constants may depend on its nonzero κ_*. Let $\boldsymbol{H}_{\boldsymbol{h}}(\varphi)$ be the positive definite Hessian with respect to $\boldsymbol{h}$. Since the loss is quadratic in $\boldsymbol{h}$, setting $\boldsymbol{z} = \boldsymbol{h} - \boldsymbol{h}^*(\varphi)$ gives

$$\nabla_{\boldsymbol{h}} J = \boldsymbol{H}_{\boldsymbol{h}} \boldsymbol{z}, \qquad \partial_\varphi J = J'_{\mathrm{prof}} - (\boldsymbol{h}^*)'^\top \boldsymbol{H}_{\boldsymbol{h}} \boldsymbol{z} + \tfrac{1}{2} \boldsymbol{z}^\top \boldsymbol{H}'_{\boldsymbol{h}} \boldsymbol{z}. \tag{127}$$

Primes in this calculation denote derivatives with respect to φ. Invariance of the graph $\boldsymbol{h} = \boldsymbol{h}_c(\varphi)$ gives

$$\boldsymbol{h}_c'(\varphi)\dot{\varphi} = -\mathsf{A}_{\boldsymbol{h}\boldsymbol{h}} \nabla_{\boldsymbol{h}} J - \mathsf{A}_{\boldsymbol{h}\varphi} \partial_\varphi J, \qquad \boldsymbol{h} = \boldsymbol{h}_c(\varphi). \tag{128}$$

The mobility is smooth and its weight block is invertible. Let A_0 be its value at the equilibrium and put $\boldsymbol{\delta} = \boldsymbol{h}_c - \boldsymbol{h}^*$. Both graphs have the same value and zero derivative there. If the first nonzero

Taylor coefficient of $\boldsymbol{\delta}$ had degree $j \in \{2,3,4\}$, then $\nabla_{\boldsymbol{h}} J$ would have a nonzero term of degree j, while the displayed formula for $\partial_\varphi J$ would give

$$\partial_\varphi J = O(\varphi^5) + O(\varphi^{j+2}) + O(\varphi^{2j}).$$

The angular equation would then give $\dot\varphi = O(\varphi^j)$. Since $\boldsymbol{h}_c' = O(\varphi^{\min\{2,j-1\}})$, the left-hand side of the invariance equation has degree strictly greater than j. Its right-hand side has the nonzero degree-j term $-(\mathsf{A}_0)_{\boldsymbol{hh}}\boldsymbol{H}_{\boldsymbol{h}}(0)\boldsymbol{\delta}$, a contradiction. Thus $\boldsymbol{\delta} = O(\varphi^5)$, $\dot\varphi = O(\varphi^5)$, and $\boldsymbol{h}_c'\dot\varphi = O(\varphi^7)$. At degree five, the invariance equation therefore reduces to the linear weight response to J'_{prof}. Using the blocks of A_0 in the coordinates $(\boldsymbol{h}, \varphi)$, this comparison gives

$$\left\|\boldsymbol{h}_c - \boldsymbol{h}^* + \boldsymbol{H}_{\boldsymbol{h}}(0)^{-1}(\mathsf{A}_0)_{\boldsymbol{hh}}^{-1}(\mathsf{A}_0)_{\boldsymbol{h}\varphi} J'_{\mathrm{prof}}\right\| \le C|\varphi|^6. \tag{129}$$

Taylor's theorem gives this inequality after choosing a C^7 center manifold. The population objective is smooth because it is polynomial in parameters whose coefficients require only moments through degree six.

Only the coordinates $(\boldsymbol{K}_{ue}, \boldsymbol{K}_{eu}, \varphi)$ contribute to the leading angular mobility. For $\boldsymbol{\Omega e} = \boldsymbol{u}_\infty$ and $\boldsymbol{\Omega u}_\infty = -\boldsymbol{e}$, the Frobenius norm squared of $\boldsymbol{\Omega}$ is two and $[\boldsymbol{\Omega}, \lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top]_{ue} = [\boldsymbol{\Omega}, \lambda_* \boldsymbol{u}_\infty \boldsymbol{u}_\infty^\top]_{eu} = -\lambda_*$. Eliminating the leading weight response gives the effective angular coefficient $c_* = (\mathsf{A}_0)_{\varphi\varphi} - (\mathsf{A}_0)_{\varphi \boldsymbol{h}}(\mathsf{A}_0)_{\boldsymbol{hh}}^{-1}(\mathsf{A}_0)_{\boldsymbol{h}\varphi}$. Denote its principal block on $(\boldsymbol{K}_{ue}, \boldsymbol{K}_{eu}, \varphi)$ by $\mathsf{A}_{0,\mathrm{rot}}$. This block and the coefficient are

$$\mathsf{A}_{0,\mathrm{rot}} = \boldsymbol{I}_3 + \tfrac{1}{2}\boldsymbol{v}\boldsymbol{v}^\top, \quad \boldsymbol{v} = (-\lambda_*, -\lambda_*, 1)^\top, \qquad c_* = 1 + \frac{1}{2(1+\lambda_*^2)}. \tag{130}$$

The bounds equation 126 and equation 129 give equation 125 on the center manifold. □

Exact equations during the initial linear fit. Write, in the ordered basis $(\boldsymbol{u}_\infty, \boldsymbol{e})$,

$$\boldsymbol{K} = \begin{pmatrix} k & x \\ y & z \end{pmatrix}, \qquad \boldsymbol{a} = (b, b_\perp)^\top, \qquad n = \cos\varphi, \qquad \chi_4 = \mu_4^* - 3\nu. \tag{131}$$

The second output coordinate is denoted by $b_\perp$. Define $c_u = \chi_4 m^3$, $c_e = 3\zeta m^2 n$ and let a prime on c_u, c_e, M denote differentiation with respect to φ. Differentiating equation 122 gives all seven gradient coordinates:

$$\begin{aligned} g_k &= \nu(k-\lambda_*) + b c_u, & g_x &= x + b c_e, \\ g_y &= \nu y + b_\perp c_u, & g_z &= z + b_\perp c_e, \\ g_b &= (k-\lambda_*)c_u + x c_e + M b - \kappa_* m^3, & g_{b_\perp} &= y c_u + z c_e + M b_\perp, \\ g_\varphi &= b[(k-\lambda_*)c_u' + x c_e'] + b_\perp[y c_u' + z c_e'] \\ &\quad + \tfrac{1}{2}M'(b^2 + b_\perp^2) - 3\kappa_* m^2 n b. \end{aligned} \tag{132}$$

For $\boldsymbol{h} = (k, x, y, z, b, b_\perp)$, set

$$\boldsymbol{v}_{\boldsymbol{h}} = (x+y, z-k, z-k, -x-y, b_\perp, -b)^\top, \qquad \omega = -\tfrac{1}{2}(\boldsymbol{v}_{\boldsymbol{h}}^\top \boldsymbol{g}_{\boldsymbol{h}} + g_\varphi). \tag{133}$$

The coordinate equations are

$$\dot{\boldsymbol{h}} = -\boldsymbol{g}_{\boldsymbol{h}} + \omega \boldsymbol{v}_{\boldsymbol{h}}, \qquad \dot\varphi = -g_\varphi + \omega. \tag{134}$$

Lemma B.6 (Entrance from zero weights)**.** *There are $L, C, c > 0$ such that, for all sufficiently small $m_0 > 0$, the endpoint trajectory with $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$ satisfies, for $0 \le s \le s_b = L\log(1/m_0)$,*

$$\begin{aligned} &|b| \le C m_0^3, && |x| + |y| \le C m_0^5, && |b_\perp| \le C m_0^8, \\ &|z| \le C m_0^{10}, && |k - \bar k(s)| \le C m_0^6, && |\dot m| \le C m_0^5, \end{aligned} \tag{135}$$

where $\bar k(s) = \lambda_*(1 - e^{-\nu s})$. At s_b,

$$|m(s_b) - m_0| \le C m_0^5 \log(1/m_0), \qquad \|\boldsymbol{h}(s_b) - \boldsymbol{h}_c(\varphi(s_b))\| \le C m_0^5. \tag{136}$$

All constants here depend on the fixed scalar target, not on m_0.

Proof. Put $\epsilon = m_0$ and initially stop the solution when $m \notin [\epsilon/2, 2\epsilon]$ or one of the scaled weight coordinates below leaves a fixed box. At $m = 0$ the trajectory starting from zero weights is $k = \bar{k}(s)$ with all other coordinates zero. Linearizing the (x, y) equations along this trajectory gives $\dot{\boldsymbol{X}} = -\boldsymbol{D}(\bar{k})\boldsymbol{X}$, where

$$\boldsymbol{X} = (x, y)^\top, \qquad \boldsymbol{D}(k) = \begin{pmatrix} 1 & 0 \\ 0 & \nu \end{pmatrix} + \frac{k\zeta}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} (1 - k \quad 1)\,. \tag{137}$$

For example, the part of $\boldsymbol{v}_{\boldsymbol{h}}^\top \boldsymbol{g}_{\boldsymbol{h}} + g_\varphi$ linear in x, y is $-\zeta[(1-k)x + y]$, with $k = \bar{k}(s)$ along this trajectory.

Use the positive definite matrix $\boldsymbol{H}_s = \mathrm{diag}(1 - \bar{k}(s), 1)$. Since $0 \le \bar{k}(s) \le \lambda_* < 1$ and $\bar{k}'(s) \ge 0$,

$$\nu^{-1}\boldsymbol{I} \preceq \boldsymbol{H}_s \preceq \boldsymbol{I}, \quad \dot{\boldsymbol{H}}_s \preceq 0, \quad \boldsymbol{H}_s\boldsymbol{D}(\bar{k}) = \mathrm{diag}(1 - \bar{k}, \nu) + \frac{\bar{k}\zeta}{2} \begin{pmatrix} 1 - \bar{k} \\ 1 \end{pmatrix} \begin{pmatrix} 1 - \bar{k} & 1 \end{pmatrix} \succeq \boldsymbol{H}_s. \tag{138}$$

Thus $\dot{\boldsymbol{X}} = -\boldsymbol{D}(\bar{k})\boldsymbol{X} + \boldsymbol{F}$ implies $d^+\|\boldsymbol{X}\|_{\boldsymbol{H}_s}/ds \le -\|\boldsymbol{X}\|_{\boldsymbol{H}_s} + \|\boldsymbol{F}\|_{\boldsymbol{H}_s}$. Define the scaled coordinates

$$B = |b|/\epsilon^3, \quad X_0 = \|\boldsymbol{X}\|_{\boldsymbol{H}_s}/\epsilon^5, \quad B_{\perp,0} = |b_\perp|/\epsilon^8, \quad Z_0 = |z|/\epsilon^{10}, \quad V_0 = |k - \bar{k}|/\epsilon^6. \tag{139}$$

On any fixed box of these coordinates with radius R, the exact formulas above give, for sufficiently small ϵ,

$$\begin{aligned} B' &\le -cB + C + C_R\epsilon^2, \\ X_0' &\le -cX_0 + CB + C_R\epsilon^2, \\ B_{\perp,0}' &\le -cB_{\perp,0} + C\{X_0 + B(X_0 + B)\} + C_R\epsilon^2, \\ Z_0' &\le -cZ_0 + C\{B_{\perp,0} + X_0(X_0 + B)\} + C_R\epsilon^2, \\ V_0' &\le -\nu V_0 + CB + C_R\epsilon^2, \qquad |\dot{m}| \le C\epsilon^5(X_0 + B) + C_R\epsilon^7. \end{aligned} \tag{140}$$

The derivatives of absolute values are upper Dini derivatives. The moment bounds used here are $|c_u| \le C\epsilon^3$, $|c_e| \le C\epsilon^2$, $|c_u'| \le C\epsilon^2$, $|c_e'| + |M'| \le C\epsilon$, and $|M - 6| \le C\epsilon^2$. The b forcing is bounded by $C\epsilon^3$; the leading (x, y) forcing is bounded by $C\epsilon^2|b|$ after removing equation 137. The $b_\perp$ forcing consists of yc_u and $-\omega b$, to leading order, and the z forcing of $b_\perp c_e$ and $-\omega(x + y)$. Finally $\dot{k} + \nu(k - \lambda_*) = -bc_u + \omega(x + y)$. These identities bound the leading terms in the displayed triangular system. Every remaining term, after its indicated normalization, is at most $C_R\epsilon^2$.

Choose upper bounds successively for $B, X_0, B_{\perp,0}, Z_0, V_0$, so each negative term strictly dominates its preceding forcing at its upper boundary. Then fix this box and reduce ϵ to make $C_R\epsilon^2$ smaller than these margins. All scaled coordinates start at zero and cannot exit the box. Integrating the angular bound gives $|m(s) - \epsilon| \le C\epsilon^5 s$. For $s \le L\log(1/\epsilon)$ this also prevents an exit from the alignment interval, proving equation 135.

To obtain the distance to the center graph, let $\eta_b = b - \kappa_* m^3/6$. The exact b equation and the preceding bounds give

$$|\dot{\eta}_b + 6\eta_b| \le C\epsilon^3 e^{-\nu s} + C\epsilon^5, \qquad |\eta_b(s)| \le C\epsilon^3(1 + s)e^{-\min(\nu,6)s} + C\epsilon^5. \tag{141}$$

Choose L so that both this decaying term and $e^{-\nu s_b}$ are at most $C\epsilon^5$. Together with equation 126 and equation 129, the coordinate bounds then give $\|\boldsymbol{h}(s_b) - \boldsymbol{h}_c(\varphi(s_b))\| \le C\epsilon^5$. The integrated angular estimate proves the other assertion. □

Lemma B.7 (Approximation by a center trajectory)**.** *Consider the OT endpoint flow from $\boldsymbol{K}_0 = \boldsymbol{a}_0 = 0$ with alignment $m_0 > 0$, and let s_b be as in Lemma B.6. For every sufficiently small fixed $\bar{m} > 0$ and all sufficiently small m_0, the trajectory reaches $m = \bar{m}$ in finite time. From s_b until that exit, it is approximated by a center-manifold trajectory Θ_c whose alignment at s_b is $\widetilde{m}_0$, with*

$$|\widetilde{m}_0 - m_0| \le Cm_0^5\log(1/m_0), \qquad \|\Theta(s) - \Theta_c(s)\| \le Cm_0^5 e^{-c(s - s_b)}. \tag{142}$$

The constants are independent of m_0. Moreover, $J(s_b) < J_{\mathrm{lin}}$.

Proof. By Lemma B.6, the distance to the center manifold at s_b is $O(m_0^5)$. In the coordinates $\boldsymbol{z}_c = \boldsymbol{h} - \boldsymbol{h}_c(\varphi)$, the equations take the form

$$\dot{\boldsymbol{z}}_c = -\boldsymbol{L}_0 \boldsymbol{z}_c + \mathcal{R}(\varphi, \boldsymbol{z}_c), \qquad \dot{\varphi} = f_c(\varphi) + b_c(\varphi, \boldsymbol{z}_c), \qquad \boldsymbol{L}_0 = (\mathsf{A}_0)_{\boldsymbol{h}\boldsymbol{h}} \boldsymbol{H}_{\boldsymbol{h}}(0). \tag{143}$$

Invariance of the graph gives $\mathcal{R}(\varphi, 0) = b_c(\varphi, 0) = 0$. Moreover, for any prescribed small $\eta > 0$, a sufficiently small fixed neighborhood satisfies

$$\|\mathcal{R}(\varphi, \boldsymbol{z}_c)\| \le \eta \|\boldsymbol{z}_c\|, \qquad |b_c(\varphi, \boldsymbol{z}_c)| \le C\|\boldsymbol{z}_c\|, \qquad |f_c'(\varphi)| \le \eta. \tag{144}$$

Here $\boldsymbol{h}_c'(0) = 0$, all eigenvalues of $\boldsymbol{L}_0$ are positive, and f_c has a fifth-order zero. Let $\boldsymbol{H}_{\mathrm{L}}$ be the positive solution of $\boldsymbol{L}_0^\top \boldsymbol{H}_{\mathrm{L}} + \boldsymbol{H}_{\mathrm{L}} \boldsymbol{L}_0 = \boldsymbol{I}$. Differentiating $\boldsymbol{z}_c^\top \boldsymbol{H}_{\mathrm{L}} \boldsymbol{z}_c$ and reducing η gives constants $C, c > 0$ with

$$\|\boldsymbol{z}_c(s)\| \le C\|\boldsymbol{z}_c(s_b)\| e^{-c(s-s_b)} \tag{145}$$

as long as the trajectory stays in this neighborhood.

Choose an outer angular interval $|\varphi| < 2\bar{\varphi}$ and a fixed normal neighborhood on which these estimates hold, and stop the trajectory at its first exit from $0 < \varphi < \bar{\varphi}$ or from that normal neighborhood. All neighborhoods are fixed independently of m_0. To control the angular phase, extend f_c from the outer interval to a globally Lipschitz function with Lipschitz constant $\eta < c$, reducing $\bar{\varphi}$ first if necessary. On any finite interval $[s_b, T]$ before exit, solve $\dot{\psi} = f_c(\psi)$ backward from $\psi(T) = \varphi(T)$. The scalar variation-of-constants inequality gives

$$|\varphi(s) - \psi(s)| \le \int_s^T e^{\eta(v-s)} |b_c(\varphi(v), \boldsymbol{z}_c(v))|\, dv \le \frac{C}{c-\eta} \|\boldsymbol{z}_c(s_b)\| e^{-c(s-s_b)}. \tag{146}$$

For a finite exit time, take T equal to that time by continuity. If the exit time is infinite, the initial phases belong to a compact interval of radius $C\|\boldsymbol{z}_c(s_b)\|$ about $\varphi(s_b)$. A convergent subsequence, together with continuous dependence of the extended scalar flow, supplies one center trajectory satisfying the same estimate on every finite interval. In either case the initial phase differs from $\varphi(s_b)$ by at most $C\|\boldsymbol{z}_c(s_b)\|$. The resulting center trajectory starts with positive alignment for small m_0 and has positive speed by equation 125. It therefore reaches any fixed small positive alignment in finite time. If the actual trajectory never exited, the center trajectory would reach $2\bar{\varphi}$ while the actual angle remained below $\bar{\varphi}$, contradicting the tracking estimate for small m_0. The normal bound excludes an exit through the boundary in $\boldsymbol{z}_c$. The center angle starts at order m_0 and increases, whereas the phase error is $O(m_0^5)$; hence exit through zero alignment is also excluded. The trajectory must therefore reach $\varphi = \bar{\varphi}$, or $m = \sin\bar{\varphi}$. This proves equation 142.

At s_b the optimal gain is at least $\kappa_*^2 m(s_b)^6/24$ and the loss due to the weight error is at most Cm_0^{10}. Hence the trajectory has also entered the strict sublevel $J < J_{\mathrm{lin}}$ for all sufficiently small m_0.

□

From angular speed to hitting time. Choose $\bar{m} > 0$ as in Lemma B.7, small enough that

$$\frac{\bar{m}^6}{\mathcal{M}_*(\bar{m})} < \frac{\theta}{D_*}. \tag{147}$$

By equation 80 and monotonicity of the optimal gain, no state with $|m| \le \bar{m}$ can reach the fractional target, even with optimal weights. Thus, for sufficiently small m_0, the target is reached after the initial fitting interval and the subsequent exit at $m = \bar{m}$. For any such fixed $\bar{m}$, equation 125 bounds $\dot{m}$ between $(c_*\kappa_*^2/2 - C\bar{m})m^5$ and $(c_*\kappa_*^2/2 + C\bar{m})m^5$. The leading equation $\dot{m} = (c_*\kappa_*^2/2)m^5$ takes time

$$\int_{\widetilde{m}_0}^{\bar{m}} \frac{2\, dm}{c_*\kappa_*^2 m^5} = \frac{\widetilde{m}_0^{-4} - \bar{m}^{-4}}{2c_*\kappa_*^2} \tag{148}$$

to reach $\bar{m}$. Applying the two speed bounds in the same way, then taking $m_0 \downarrow 0$ and $\bar{m} \downarrow 0$, gives

$$\lim_{m_0 \downarrow 0} \kappa_*^2 m_0^4 T_{\mathrm{OT},\theta}^{*,\mathrm{full}}(m_0) = \frac{1}{2c_*} = \frac{1+\lambda_*^2}{3+2\lambda_*^2}. \tag{149}$$

To justify using the entire hitting time here, the tracking bound places the actual exit time between the center times to two fixed nearby positive alignments, both with the same normalized limit. At exit the loss is below J_{lin} by a fixed positive amount. Boundedness of the weights and equation 101 bound the remaining time to the fractional target, uniformly as $m_0 \downarrow 0$: on that compact loss interval the gradient has a positive minimum. The same stationary-point classification proves convergence to the aligned optimum. Neither the initial interval s_b nor the final bounded interval affects the limit.

B.7 HITTING TIMES AND RANDOM INITIALIZATION

Proof of Proposition 2 and Theorem 2. We first take $t \uparrow 1$ with $m_0 > 0$ fixed, then let $m_0 \downarrow 0$. Let $m_\theta \in (0,1)$ be the unique solution of $m_\theta^6/\mathcal{M}_*(m_\theta) = \theta/D_*$, and put

$$\tau_\theta(m_0) = \int_{m_0}^{m_\theta} \frac{dm}{F_*(m)}. \tag{150}$$

Choose $m_* < m_\theta$. Positivity of F_* on $(0,1)$ follows from equation 112, or from the strict increase of the optimal gain. Its expansion at zero gives the inequality $|F_*(m) - m^5/2| \le Cm^7$ on a fixed small interval. Integrating this bound and treating the remaining compact interval separately yields

$$\left|\tau_\theta(m_0) - \frac{1}{2m_0^4}\right| \le C_\theta(1 + m_0^{-2}), \qquad 0 < m_0 \le m_*. \tag{151}$$

Independent coupling. Lemma B.3 implies, uniformly on compact positive slow-time intervals,

$$\frac{J_{\mathrm{lin,ind},t} - \mathcal{J}_{\mathrm{ind},t}(s)}{\kappa_{\mathrm{ind}}(t)^2} \longrightarrow \frac{m(\tau)^6}{2\mathcal{M}_*(m(\tau))}, \qquad \tau = \kappa_{\mathrm{ind}}(t)^2 s. \tag{152}$$

Indeed, the loss above the minimum over the weights is a bounded quadratic form in the tracking error. With $p = 6$ in Lemma B.2, the geometric loss divided by κ^2 is at most $C\alpha^8$ after entrance, and the remaining support perturbation tends to zero as well. The limiting gain crosses the fractional target with positive derivative. Values on either side of this crossing bracket the finite-t hitting time. Because the full loss decreases along gradient flow, it cannot cross the threshold earlier and then return above it. Consequently,

$$\lim_{t\uparrow 1} \kappa_{\mathrm{ind}}(t)^2 T^{\mathrm{full}}_{\mathrm{ind},\theta}(t) = \frac{2}{3}\tau_\theta(m_0). \tag{153}$$

At any fixed positive slow time before the crossing the limiting gain is strictly positive. For sufficiently large t, the trajectory has therefore entered the strict sublevel in Lemma B.1, which also proves its convergence.

OT coupling. For OT, $s_g \to 0$. The transported initial weights tend to zero, and the transported basis tends to $\boldsymbol{B}_\infty^{(0)}$. Smooth dependence of the internal vector field and equation 103 give convergence on every compact interval in s to the endpoint flow analyzed above. At the fractional target this limiting trajectory is not stationary, so its loss has strictly negative derivative. The same bracketing argument gives

$$\lim_{t\uparrow 1} T^{\mathrm{full}}_{\mathrm{OT},\theta}(t) = T^{*,\mathrm{full}}_{\mathrm{OT},\theta}(m_0). \tag{154}$$

The strict linear-loss sublevel again gives convergence for finite t sufficiently close to one. Combining the two limits with equation 149 and equation 151, we obtain

$$\begin{aligned} \lim_{m_0\downarrow 0}\lim_{t\uparrow 1} \kappa_{\mathrm{OT}}(t)^2 m_0^4 T^{\mathrm{full}}_{\mathrm{OT},\theta}(t) &= \frac{1+\lambda_*^2}{3+2\lambda_*^2}, \\ \lim_{m_0\downarrow 0}\lim_{t\uparrow 1} \kappa_{\mathrm{ind}}(t)^2 m_0^4 T^{\mathrm{full}}_{\mathrm{ind},\theta}(t) &= \frac{1}{3}. \end{aligned} \tag{155}$$

The endpoint trajectories involve only the plane spanned by the feature and its initial direction. Their bounds therefore have constants depending on the scalar target and θ, independently of D, r. Shrinking m_* if necessary, the positive limits give the two-sided inequalities in equation 92. The finite-t threshold can depend on the initialization.

Uniformity over compact initialization sets. Fix D, r and a compact set $\mathcal{I}$ of initializations satisfying $\sigma_{\min}(\boldsymbol{A}^\top \boldsymbol{B}_0) \ge \gamma > 0$ and $0 < m_- \le m_0 \le m_*$. The smooth phase map Π makes the limiting initial data continuous on $\mathcal{I}$. The support-entrance estimates, weight bounds, and vector-field derivative bounds above are uniform on this set. Thus, after support entrance, the internal trajectories converge uniformly to their limits on common compact time intervals bounded away from zero. For independent coupling this statement uses the normalized heads and slow time, and the loss

gain is divided by $\kappa_{\mathrm{ind}}(t)^2$. For OT it uses the original internal parameters and training time. The corresponding gains converge uniformly on these intervals as well.

Each limiting trajectory reaches the fractional target at a finite positive time and crosses it strictly: for independent coupling this follows from $F_*(m_\theta) > 0$, and for OT from equation 101 and the absence of a stationary point at that loss. Smooth dependence on initial data and the implicit function theorem therefore make the limiting crossing time continuous locally, and hence on $\mathcal{I}$. Compactness gives a common finite time interval containing all crossings and a positive lower bound on the crossing slopes of the limiting gains. Consequently the gain margins at times just before and after each crossing can be chosen uniformly. Uniform convergence of the gains then brackets every finite-t crossing for a single sufficiently large $t_0 < 1$. This proves the uniformity needed in Proposition 2. The resulting t_0 may depend on D, r, γ, m_-.

Random initialization. Finally, a uniform Stiefel initialization has invertible $\boldsymbol{A}^\top \boldsymbol{B}_0$ almost surely. For fixed D, r, δ, choose $\gamma_\delta > 0$ so that

$$\mathbb{P}\{\sigma_{\min}(\boldsymbol{A}^\top \boldsymbol{B}_0) \geq \gamma_\delta\} \geq 1 - \delta/2. \tag{156}$$

Conditionally on $\boldsymbol{B}_0$, the vector $\boldsymbol{w}_0$ is uniform on the sphere and $(\boldsymbol{B}_\infty^{(0)})^\top \boldsymbol{q}_*$ is a fixed unit vector. Spherical coordinate quantiles give constants $0 < \ell_\delta < u_\delta < \infty$ such that, for all sufficiently large r,

$$\mathbb{P}\left\{ \frac{\ell_\delta}{\sqrt{r}} \leq m_0 \leq \frac{u_\delta}{\sqrt{r}} \,\middle|\, \boldsymbol{B}_0 \right\} \geq 1 - \delta/2. \tag{157}$$

Increase r_δ so that $u_\delta/\sqrt{r} \leq m_*$. Intersecting these two events gives a compact initialization set of probability at least $1 - \delta$. Uniformity on this set supplies a deterministic $t_0 < 1$. The same event works for both couplings because their auxiliary geometry flow is identical. The nonzero endpoint correlation limits in equation 90, together with the two spherical quantiles, now give equation 22. This completes the proof. □

C Numerical Experiment Details

C.1 Numerical setup for joint support and feature learning

For Figure 1, we integrate the population gradient flow equation 86 with unit learning rates for all parameter blocks. We set $D = 3$, $r = 2$, $\boldsymbol{A} = (\boldsymbol{e}_1, \boldsymbol{e}_2)$, $\boldsymbol{u} = \boldsymbol{e}_1$, $H = \mathrm{He}_3$, and $\rho = 0.12$, and use the same initialization for both couplings:

$$\boldsymbol{B}_0 = \left(\frac{\boldsymbol{e}_1 + \boldsymbol{e}_3}{\sqrt{2}},\, \boldsymbol{e}_2 \right), \qquad \boldsymbol{w}_0 = \left(\frac{1}{8}, \frac{\sqrt{63}}{8} \right)^\top, \qquad \boldsymbol{K}_0 = 0, \quad \boldsymbol{a}_0 = 0. \tag{158}$$

Population expectations are computed by tensor-product Gauss–Hermite quadrature of order ten in the cubic latent coordinate and order four in every other Gaussian coordinate. The squared residual has degree at most 18 in the cubic latent coordinate and at most six in each other coordinate, so this rule integrates the loss and its parameter gradients exactly. The ODE is solved by BDF.

Panel (a) plots the support and feature errors

$$e_{\mathrm{sup}}(s) = \frac{1}{2} \|\boldsymbol{B}_s \boldsymbol{B}_s^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_F^2, \qquad e_{\mathrm{feat}}(s) = 1 - |(\boldsymbol{A}\boldsymbol{u})^\top \boldsymbol{B}_s \boldsymbol{w}_s|^2 \tag{159}$$

at $\varepsilon_t = 0.1$ for $0 \leq s \leq 10^4$. Panels (b,c) use $\varepsilon_t \in \{0.2, 0.1, 0.05\}$, with $t = 1/(1 + \varepsilon_t)$, and show the residual cubic correlation $|\kappa_C(t)|$ defined in Appendix B.1 and the hitting time $T_{C,1/2}^{\mathrm{full}}$ defined in equation 19, respectively. The reference lines are matched to the independent-coupling values at $\varepsilon_t = 0.05$.

C.2 Numerical setup for shared-time training

Figure 2 uses the same target, initialization, and unknown-support model as Figure 1. For each $a \in \{0, 0.5, 0.7, 0.8, 0.85, 0.9\}$, let $(\xi_i, \omega_i)_{i=1}^n$ be the n-point Gauss–Legendre nodes and weights

on $[-1, 1]$, with $n = 8$. Define $t_i = a + (1 - a)(\xi_i + 1)/2$ and $p_i = \omega_i/2$, so that $a < t_i < 1$ and $\sum_i p_i = 1$. The model minimizes the discrete-time population loss

$$\overline{\mathcal{J}}^{(n)}_{C,a} = \sum_{i=1}^{n} p_i \mathcal{J}_{C,t_i}. \tag{160}$$

The finite nodes and weights define the training distribution itself. Population expectations refer to endpoint pairs at each node, and every gradient update uses the full weighted sum. This objective is finite for the stated initialization. In contrast, the continuous uniform-time loss over $[a, 1]$ diverges whenever $\mathcal{G}(\boldsymbol{B}) > 0$, because its support term is $\mathcal{G}(\boldsymbol{B})/(1 - t)^2$. The discrete training distribution avoids this divergence. The support basis $\boldsymbol{B}$ and feature direction $\boldsymbol{w}$ are shared across all n flow times. To condition the remaining weights on time, let $P_0, \dots, P_3$ be the first four Legendre polynomials and set

$$\boldsymbol{K}(t) = \sum_{j=0}^{3} \boldsymbol{K}_j P_j(z_{a,t}), \qquad \boldsymbol{a}(t) = \sum_{j=0}^{3} \boldsymbol{a}_j P_j(z_{a,t}), \qquad z_{a,t} = \frac{2(t - a)}{1 - a} - 1. \tag{161}$$

All coefficients, together with $\boldsymbol{B}$ and $\boldsymbol{w}$, follow population gradient flow with unit learning rates.

As a separate diagnostic of the window, panel (a) reports the root mean squared residual cubic correlation

$$\overline{\kappa}_C(a) = \left(\frac{1}{1 - a} \int_a^1 \kappa_C(t)^2 \, dt \right)^{1/2}. \tag{162}$$

Panel (b) reports the first gradient-flow time at which the feature-direction error in equation 159 satisfies $e_{\mathrm{feat}} \le 1/2$. This criterion depends only on the shared direction $\boldsymbol{B}\boldsymbol{w}$ and does not require choosing an evaluation flow time.

For panel (c), we partition $(0, 1)$ into $K \in \{1, 2, 4, 8\}$ equal bins. The support basis $\boldsymbol{B}$ remains shared, while bin k has its own feature direction $\boldsymbol{w}_k$ and output weights $\boldsymbol{K}_k(t)$ and $\boldsymbol{a}_k(t)$, expanded in Legendre polynomials up to degree three in the local bin coordinate. Each bin is assigned mass $1/K$. We multiply the learning rates of its bin-specific parameters by K to remove the resulting $1/K$ update factor. The learning rate of the shared basis is unchanged. The reported time is the first s at which the direction in the final bin satisfies

$$1 - |(\boldsymbol{A}\boldsymbol{u})^\top \boldsymbol{B}_s \boldsymbol{w}_{K,s}|^2 \le \frac{1}{2}. \tag{163}$$

Thus $K = 1$ recovers the fully shared model. For $K > 1$, only the final bin contributes directly to the update of $\boldsymbol{w}_K$, while all bins update the shared basis $\boldsymbol{B}$. The ambient endpoint direction $\boldsymbol{B}\boldsymbol{w}_K$ therefore remains coupled to the other bins through $\boldsymbol{B}$.

We evaluate the finite sum in equation 160 directly and use twelve-point Gauss–Legendre quadrature for the diagnostic equation 162. At each flow time, population expectations are computed by tensor-product Gauss–Hermite quadrature of order ten in the cubic latent coordinate and order four in every other Gaussian coordinate. This Gaussian rule integrates the polynomial endpoint-pair expectation exactly at each flow time. The ODE is solved by BDF. For panel (c), we use at least two Gauss–Legendre nodes per bin and at least eight nodes in total. Feature directions are parameterized by angles, and a constraint-stabilization term maintains orthogonality of the support basis. Increasing the number of flow-time nodes changes the reported alignment times by less than one percent.

C.3 MNIST AND CIFAR-10 DECODER EXPERIMENTS

Data and models. We train a class-conditional variational autoencoder (VAE) on each dataset's training split with pixels scaled to $[-1, 1]$, then freeze its decoder $f_c : \mathbb{R}^r \to \mathbb{R}^D$. To generate targets, we sample classes uniformly and $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_r)$, replace z_0 by $z_0 + 0.35 \tanh(z_0)$, and decode the result. All methods use the same decoder for a given dataset. The VAE encoder uses two strided convolutions followed by fully connected layers that predict the latent Gaussian parameters. The decoder maps the latent vector and one-hot class label through fully connected layers and two transposed convolutions, ending with a $\tanh$ output. We train the VAE for 10 epochs with reconstruction and KL losses. The latent dimensions are 8 for MNIST and 16 for CIFAR-10.

Field predictors are residual CNNs that preserve image resolution. An input convolution maps the image to 32 hidden channels for MNIST or 128 for CIFAR-10, followed by two residual blocks and an output convolution that returns a vector field with the same shape as the image. Each block contains two 3×3 convolutions, each preceded by GroupNorm and SiLU, with an identity skip connection.

Training. For each method and noise scale, we train a separate predictor on

$$\boldsymbol{x} = \boldsymbol{y} + \epsilon \boldsymbol{g}, \qquad \boldsymbol{u} = -a_\epsilon \boldsymbol{g}, \qquad a_\epsilon = \begin{cases} 1+\epsilon, & \text{FM and OT-CFM,} \\ 1/\epsilon, & \text{diffusion,} \end{cases} \tag{164}$$

where $\boldsymbol{y}$ is a decoded image and $\boldsymbol{g}$ is standard Gaussian noise. Independent FM and diffusion sample them independently. OT-CFM pairs samples by Hungarian assignment within blocks of 64, minimizing squared pixel distance. The same coupling is used for training and evaluation. The FM label is the endpoint correction $\boldsymbol{U}_t - \boldsymbol{X}_t/t$ at the rescaled input $\boldsymbol{x} = \boldsymbol{X}_t/t$, with $t = (1+\epsilon)^{-1}$. The diffusion label estimates the additive-noise score. We minimize $\frac{1}{2}\mathbb{E}\|\boldsymbol{v}_\theta(\boldsymbol{x}) - \boldsymbol{u}\|^2$ using 3,000 AdamW updates with batch size 256, across ten noise scales from 10^{-4} to 10^{-1}. Learning rates are selected separately within each sweep, method, and noise scale by the lowest seed-averaged final training loss.

Reference fields and evaluation. For each coupling and noise scale, two independently trained U-Nets predict the Gaussian noise from the noisy image. Each U-Net has two downsampling stages, with 32, 64, and 128 channels at successive resolutions, and two corresponding upsampling stages. Strided convolutions reduce resolution and transposed convolutions restore it, with additive skip connections between matching resolutions. Two residual blocks process the lowest-resolution features, and one block refines each upsampling stage. All these blocks use the same structure as the field predictor. Each reference is trained for 5,000 AdamW updates. Their mean prediction, multiplied by $-a_\epsilon$, defines $\boldsymbol{v}_{\mathrm{ref}}$, which is used only for evaluation. We evaluate on 256 held-out pairs per run, fixed throughout training. Let $\boldsymbol{P}_N$ project onto the normal space obtained from the decoder Jacobian, and let $\boldsymbol{d}$ be the unit tangent gradient of the warped latent coordinate. We compute this direction using a regularized inverse of the Jacobian's Gram matrix. The density-direction metric uses only points where the coordinate is locally identifiable. With $\Delta\boldsymbol{v} = \boldsymbol{v}_\theta - \boldsymbol{v}_{\mathrm{ref}}$, the errors are

$$e_N = \frac{\mathbb{E}\|\boldsymbol{P}_N \Delta\boldsymbol{v}\|^2}{\mathbb{E}\|\boldsymbol{P}_N \boldsymbol{v}_{\mathrm{ref}}\|^2}, \tag{165}$$

$$e_d = \frac{\mathbb{E}\langle \boldsymbol{d}, \Delta\boldsymbol{v}\rangle^2}{\mathbb{E}\langle \boldsymbol{d}, \boldsymbol{v}_{\mathrm{ref}}\rangle^2}. \tag{166}$$

Expectations are empirical averages. These are relative errors against method-specific learned references.

D Extended related work

Training dynamics in diffusion models Shah et al. (2023) prove that gradient descent on the DDPM objective recovers spherical Gaussian mixture parameters under separation and initialization conditions. Han et al. (2024) use neural tangent kernel analysis and early stopping to establish optimization and generalization guarantees for score estimation with gradient-descent-trained networks. Han et al. (2025) compare diffusion and classification training and show that the denoising objective promotes more balanced feature learning in their data model.

Wang & Pehlevan (2025) analyze full-batch gradient flow for linear and convolutional diffusion denoisers. They combine the parameter dynamics with the probability-flow ODE to describe how the generated distribution changes during training. They characterize how data covariance determines the learning rates of different modes. Bonnaire et al. (2025) identify an early training timescale for generalization and a later one for memorization, supported by a high-dimensional random-feature analysis. Huang et al. (2026) analyze a two-stage training scheme in which small-noise score matching learns the manifold projection and moderate-noise training learns the intrinsic density.