FINITE-SAMPLE APPROXIMATION OF HESSIAN-GUIDED PERTURBED WASSERSTEIN GRADIENT FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wasserstein gradient flow extends gradient descent to probability measures. Its Hessian-guided perturbed variant (PWGF) adds Gaussian perturbations to escape saddle points in nonconvex problems. We investigate when its approximation by finitely many interacting particles remains accurate over growing time horizons. Our analysis retains the curvature accumulated along the population-driven reference path: negative curvature can amplify approximation errors, while subsequent positive curvature can damp their influence. This captures favorable scenarios in which temporary instability is compatible with accurate tracking over growing horizons. Under regularity assumptions and a prescribed common perturbation schedule, we prove particle and objective-value tracking bounds on a high-probability event for reference paths satisfying explicit conditions on accumulated curvature. To handle state-dependent Gaussian jumps, we construct a population-first coupling that preserves the reference particles' conditional independence and reduces jump errors to covariance comparison. We verify the conditions in a variance-plus-cosine model, where curvature recovery yields a growing-horizon tracking guarantee. We also establish local attraction, transverse descent, and positive second variation in two regions of a regularized matrix-factorization model, motivating a positive–negative–positive curvature pattern.

1 INTRODUCTION

We study finite-particle approximation for nonconvex optimization over probability measures,

$$\min_{\mu \in \operatorname{dom} F \subseteq \mathcal{P}_2(\mathbb{R}^d)} F(\mu), \tag{1}$$

where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the set of Borel probability measures with finite second moment; the domain $\operatorname{dom} F$ for our moment models is specified in Definition A.3. The canonical first-order dynamics is the Wasserstein gradient flow

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \qquad v_t(x) = -\nabla_W F(\mu_t, x), \tag{2}$$

which is gradient descent in the 2-Wasserstein geometry. In mean-field models, (2) is the population equation used to describe interacting-particle dynamics in the infinite-particle limit. This viewpoint underlies generative modeling (Arbel et al., 2019; Chu et al., 2019), particle variational inference and sampling (Jordan et al., 1998; Liu & Wang, 2016), and mean-field neural-network training (Chizat & Bach, 2018; Mei et al., 2018; Nitanda et al., 2022a). In several convex, displacement-convex, and entropy-regularized settings, WGF and its particle approximations admit strong convergence and propagation-of-chaos (Sznitman, 1991) guarantees (Chen et al., 2025; Suzuki et al., 2023; Nitanda et al., 2025; 2022b).

For nonconvex objectives, however, first-order stationarity does not rule out saddle points. Such second-order landscape structure already appears in mean-field models, including the attention model studied by Kim & Suzuki (2024). Inspired by perturbed gradient descent (Ge et al., 2015; Jin et al., 2018), Yamamoto et al. (2025) introduce Hessian-guided perturbed Wasserstein gradient flow (PWGF), which perturbs the population measure along Hessian-guided Gaussian vector fields to escape saddles. Their guarantees concern the population measure flow. A finite implementation replaces both the drift and the Hessian-guided jump covariance by empirical quantities, introducing errors along the flow and at each perturbation. The interacting particles are generally dependent, so standard sampling bounds for independent samples do not apply directly.

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.

We therefore ask:

When can a finite system of interacting particles accurately approximate population PWGF?

We answer this finite-sample question through the following contributions.

Contributions.

1. **Tracking and objective values.** When the reference path satisfies a controlled-curvature path condition, we obtain particle and objective-value tracking bounds throughout $[0, T]$ under regularity assumptions and a prescribed deterministic common jump schedule (Theorem 3.1). The bounds retain their T-dependence and weight ODE residuals and jump mismatches by separate curvature-response factors. By retaining the accumulated curvature, they capture how later positive curvature damps earlier errors. For fixed horizons, the ODE and jump prefactors are $N^{-1/2}$ and $N^{-1/4}$, respectively, up to arbitrarily small powers of N. A variance-plus-cosine example gives a growing-horizon tracking guarantee (Example 3.1).
2. **A population-first coupling for Gaussian jumps.** At each jump, we construct a finite-dimensional standard Gaussian vector that reproduces the population-field evaluations and drives the empirical update. This preserves the correct marginal dynamics and the reference particles' conditional i.i.d. law given the population fields, while reducing the jump comparison to a covariance-factor mismatch (Sections 2.3 and 5.2).
3. **Local geometry of matrix factorization.** Matrix factorization provides a product model for the nonlinear-moment geometry arising in matrix and tensor decompositions. In a regularized product-measure model, we establish attraction along an anti-symmetric branch, descent transverse to that branch, and positive second variation in two regions under spherical product pushforward perturbations (Section 4). These local results motivate the positive–negative–positive curvature scenario.

Relation to prior work. Yamamoto et al. (2025) analyze population PWGF escape and convergence, but do not bound its population-to-empirical approximation error. Glasgow et al. (2025) track continuous projected neural-network dynamics using local particle curvature, without Hessian-induced Gaussian jumps. Reference-particle couplings and Gaussian covariance-factor alignment have precedents (Sznitman, 1991; Masarotto et al., 2019). Here the coupling must preserve the prescribed population fields, because the conditional reference-particle law given those fields is the input to concentration. We establish this compatibility and control the resulting jump covariance residual through weighted features, then propagate continuous and jump errors through the sampled curvature.

Scope. The population PWGF proposed by Yamamoto et al. (2025) selects perturbation times adaptively. Prescribed schedules provide a reference point for studying how adaptive timing affects finite-particle approximation. Our results quantify finite-particle approximation errors under a common prescribed schedule while retaining state-dependent drift and jump covariance. The tracking bound applies along reference paths satisfying the controlled-curvature path condition in Theorem 3.1, which bounds error propagation along those paths. The variance-plus-cosine example verifies this condition with high probability and gives a growing-horizon tracking guarantee with a single prescribed jump (Example 3.1). Extending this verification to broader models and synchronizing adaptive jump times remain open directions.

2 Coupled population and particle PWGF

We define the population and finite-particle dynamics and the coupling used to compare them along a prescribed population PWGF path.

2.1 Population WGF and Hessian-guided perturbations

Let $F : \operatorname{dom} F \to \mathbb{R}$ be the objective functional. Between perturbation times, the population curve $(\mu_t)_{t\geq 0}$ follows the Wasserstein gradient flow

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \qquad v_t(x) := -\nabla_W F(\mu_t, x). \tag{3}$$

The original population PWGF triggers perturbations adaptively when $\|\nabla_W F(\mu_t)\|_{L^2(\mu_t)}$ falls below a prescribed first-order threshold; our finite-sample comparison uses deterministic jump times $0 < t_1 < \cdots < t_{L_{\text{jump}}} \le T$, fixed in advance and common to both systems. At each such time, Yamamoto et al. (2025) perturb the measure by pushing it forward along a Hessian-guided Gaussian vector field. The idea parallels perturbed gradient methods in Euclidean space (Ge et al., 2015; Jin et al., 2018): near a saddle, randomness should reach negative-curvature directions. In $\mathcal{P}_2(\mathbb{R}^d)$, (i) the definition of such perturbations and (ii) their alignment with descent directions are both nontrivial, motivating the Hessian-induced covariance. In Wasserstein space, the second variation along the pushforward $(\mathrm{Id} + hv)_{\#}\mu$ is represented by $\mathcal{H}_\mu = H_\mu + H'_\mu$, where

$$(H_\mu v)(x) := \int \nabla_W^2 F(\mu; x, y) v(y)\, \mu(dy), \qquad (H'_\mu v)(x) := \nabla_x \nabla_W F(\mu, x) v(x).$$

See Ambrosio et al. (2005) for detailed properties. The perturbation studied here uses the squared integral Hessian H_μ^2 as covariance. The tracking curvature uses the full operator $\mathcal{H}_\mu$; first-order stationarity alone does not imply $H'_\mu = 0$. The covariance kernel of H_μ^2 is

$$K_\mu(x, y) := \int \nabla_W^2 F(\mu; x, z) \nabla_W^2 F(\mu; z, y)\, \mu(\mathrm{d}z).$$

This covariance preserves the eigenspaces of H_μ and assigns variance proportional to its squared eigenvalues. Relating this covariance to descent directions of the full Hessian requires additional structural assumptions. The population jump is

$$\mu_{t_\ell} = (\mathrm{Id} + \eta_\ell \xi_{\mu_{t_\ell -}})_{\#}\mu_{t_\ell -}, \qquad \xi_{\mu_{t_\ell -}} \sim \mathrm{GP}(0, K_{\mu_{t_\ell -}}), \tag{4}$$

with perturbation size $\eta_\ell > 0$.

2.2 Empirical particle dynamics

For $\boldsymbol{X} = (X^1, \ldots, X^N) \in (\mathbb{R}^d)^N$, write $\rho_{\boldsymbol{X}} := \frac{1}{N}\sum_{i=1}^N \delta_{X^i}$. The empirical PWGF evolves by the particle ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t^i = -\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i), \qquad i = 1, \ldots, N, \tag{5}$$

between jump times. At t_ℓ, the particles are updated by

$$X_{t_\ell}^i = X_{t_\ell -}^i + \eta_\ell \xi_{\rho_{\mathbf{X}_{t_\ell -}}}(X_{t_\ell -}^i), \qquad i = 1, \ldots, N, \tag{6}$$

where $\xi_{\rho_{\mathbf{X}_{t_\ell -}}} \sim \mathrm{GP}(0, K_{\rho_{\mathbf{X}_{t_\ell -}}})$ is a Gaussian process induced by the empirical Hessian-guided kernel. The covariance of $\xi_{\rho_{\mathbf{X}}}(X^1), \ldots, \xi_{\rho_{\mathbf{X}}}(X^N)$ is a block matrix whose (i, j)-block is $K_{\rho_{\mathbf{X}}}(X^i, X^j)$. For a fixed configuration, this jump is finite dimensional: one applies an $Nd \times Nd$ covariance factor to a standard Gaussian vector in $\mathbb{R}^{Nd}$. It can therefore be simulated without access to the population path.

2.3 The auxiliary reference particles

The comparison is cleaner if we introduce a reference particle system $Y_t = (Y_t^1, \ldots, Y_t^N)$. Conditional on the population fields, $Y_t^1, \ldots, Y_t^N$ are i.i.d. with law μ_t. The reference particles start from the same initial positions as the empirical particles and evolve with the population drift and jump covariance. Their Gaussian increments are coupled to the empirical increments through the common Gaussian vector described below. Specifically, we set

$$X_0^i = Y_0^i, \qquad Y_0^i \overset{\text{i.i.d.}}{\sim} \mu_0. \tag{7}$$

Between jumps,

$$\frac{\mathrm{d}}{\mathrm{d}t} Y_t^i = -\nabla_W F(\mu_t, Y_t^i), \qquad i = 1, \ldots, N. \tag{8}$$

At t_ℓ, the reference jump is

$$Y_{t_\ell}^i = Y_{t_\ell -}^i + \eta_\ell \xi_{\mu_{t_\ell -}}(Y_{t_\ell -}^i), \qquad i = 1, \ldots, N, \tag{9}$$

At each jump, the population Gaussian field is generated from a fresh innovation independent of the initial sample and the previous innovations. Conditional on these fields, the flow and jump maps are deterministic; applying their common composition to the independent initial positions preserves the reference particles' conditional i.i.d. law (Lemma A.14 and eq. (115)). For the comparison, Lemma A.14 constructs a common standard Gaussian vector $\mathbf{g}_\ell^N \in \mathbb{R}^{Nd}$ from the given population-field evaluations and independent auxiliary Gaussian noise, leaving the population path unchanged. If B_{pop} denotes the aligned factor reproducing the population-field values at the reference particles and M_X denotes the empirical Hessian factor, then

$$\big(\xi_{\mu_{t_\ell-}}(Y^i_{t_\ell-})\big)_{i=1}^N = B_{\mathrm{pop}}\mathbf{g}_\ell^N, \qquad \big(\xi_{\rho_{\mathbf{X}_{t_\ell-}}}(X^i_{t_\ell-})\big)_{i=1}^N = M_X\mathbf{g}_\ell^N.$$

Both jumps have the prescribed conditional laws, and the Gaussian field-value vectors differ by $(M_X - B_{\mathrm{pop}})\mathbf{g}_\ell^N$. The precise construction, including singular covariances, and the conditional i.i.d. law of the reference particles are given in Lemmas A.14 and A.19.

2.4 Tracking metric and decomposition

We measure the pathwise discrepancy by $r_t := \Big(\frac{1}{N}\sum_{i=1}^N \|X_t^i - Y_t^i\|^2\Big)^{1/2} = \frac{1}{\sqrt{N}}\|X_t - Y_t\|_2$. Since the empirical coupling $N^{-1}\sum_i \delta_{(X_t^i, Y_t^i)}$ is admissible, $W_2(\rho_{\mathbf{X}_t}, \rho_{\mathbf{Y}_t}) \le r_t$. To compare objective values with the population flow, we combine the paired-particle estimate with concentration of the reference cloud's feature moments (Section A.7). The next section gives joint tracking and objective bounds under a prescribed common schedule.

3 Curvature-driven tracking theory

This section proves a finite-sample comparison theorem for perturbed WGF. The central object is the curvature propagator $G_\kappa^Y(t,s)$, defined below. It transports an error created at time s to time t, discounting or amplifying it according to the curvature accumulated after its creation.

3.1 Curvature-response factors

Definition 3.1 (Curvature-response factors). *Let H_t^Y denote the empirical discretization of $\mathcal{H}_{\mu_t}$ along the reference particles, acting on the indexed particle space $(\mathbb{R}^d)^N$:*

$$\mathrm{H}_t^Y := \left\{\frac{1}{N}\nabla_W^2 F\Big(\mu_t, Y_t^i, Y_t^j\Big) + \delta_{ij}\nabla_x\nabla_W F\big(\mu_t, Y_t^i\big) \in \mathbb{R}^{d\times d}\right\}_{i,j=1}^N. \tag{10}$$

We define the sampled curvature by

$$\kappa_t^{Y,N} := \lambda_{\min}(\mathrm{H}_t^Y) := \inf_{\mathbf{v}\in(\mathbb{R}^d)^N, \|\mathbf{v}\|_2=1} \langle \mathbf{v}, \mathrm{H}_t^Y\,\mathbf{v}\rangle \tag{11}$$

Positive $\kappa_t^{Y,N}$ corresponds to instantaneous contraction of the linearized discrepancy, while negative $\kappa_t^{Y,N}$ corresponds to local expansion. This is the same role played by a local strong-convexity parameter in ordinary gradient descent, except that here the parameter may change sign along the trajectory. For example, $F(\mu) = \int \frac{\|x\|^2}{2}\,\mathrm{d}\mu$ implies $\mathrm{H}_t^Y = \boldsymbol{I}_{dN}$ and $\kappa_t^{Y,N} = 1 > 0$.

On any reference path satisfying $\int_0^T |\kappa_u^{Y,N}|\,\mathrm{d}u < \infty$, we define the continuous curvature-response factor

$$\mathfrak{C}_t^Y := \int_0^t G_\kappa^Y(t,s)\mathrm{d}s, \quad G_\kappa^Y(t,s) := \exp\left\{-\int_s^t \kappa_u^{Y,N}\,\mathrm{d}u\right\}, \qquad 0 \le s \le t$$

and the jump curvature-response factor

$$\mathfrak{J}_t^Y := \sum_{\ell=1}^{L_t} \eta_\ell G_\kappa^Y(t, t_\ell), \quad L_t := \#\{\ell : t_\ell \le t\}.$$

Remark 3.1 (Conditioning used below.). *Let $\mathcal{G}_{\mathrm{pop}}$ be the sigma-field generated by the population Gaussian fields and the prescribed jump schedule. Conditional on $\mathcal{G}_{\mathrm{pop}}$, the reference particles are i.i.d. with law μ_t, while the finite-dimensional coupling vectors $\mathbf{g}_\ell^N$ are not included in this conditioning; see Section 2.3.*

3.2 MAIN THEOREM

We now quantify how curvature controls the accumulation of drift fluctuations and jump mismatches over a growing horizon, under the following assumptions:

- a structured moment representation of F, with weighted smoothness controlling its drift and Hessian (Assumption 5.1);
- uniform particle moments, branchwise Lipschitz paths, and the stated non-spikiness bound for the paired errors (Assumption 5.2); and
- a prescribed deterministic common schedule and the population-first Gaussian coupling (Assumption 5.3).

Under these conditions, we obtain the following comparison, valid simultaneously for all $t \in [0, T]$.

Theorem 3.1 (Particle and objective tracking along controlled-curvature paths)**.** *Assume Assumptions 5.1 to 5.3, with fixed structural parameters. Fix $\upsilon \geq 0$, $\zeta > 0$ with $\upsilon + \zeta < 1/4$, and constants $C_p, C_0, L_0 > 0$. Let $T \leq N^{1/4-\upsilon-\zeta}$, $p \in [N^{-C_p}, 1)$, and use a prescribed deterministic schedule with at most L_0 jumps and amplitudes $0 \leq \eta_\ell \leq C_0$. For brevity, write $\Xi_{N,T,p} := N(e+T)/p$. For each N, T, p, there is an event $\Omega_p = \Omega_{N,T,p}$, independent of ε, with probability at least $1-p$. For every $\varepsilon > 0$ and all sufficiently large N, the following holds on Ω_p. Suppose that the realized reference path satisfies*

$$\int_0^T |\kappa_t^{Y,N}| \, \mathrm{d}t < \infty, \qquad \sup_{t \leq T} \mathfrak{C}_t^Y \leq C_0 N^\upsilon (1+T), \qquad \sup_{t \leq T} \mathfrak{J}_t^Y \leq C_0 N^\upsilon.$$

Then, with $r_t = (N^{-1} \sum_{i=1}^N \|X_t^i - Y_t^i\|_{\mathbb{R}^d}^2)^{1/2}$, simultaneously for $t \in [0, T]$,

$$\begin{aligned} &r_t \vee |F(\rho_{\mathbf{X}_t}) - F(\mu_t)| \\ &\quad \leq C_\varepsilon \Xi_{N,T,p}^\varepsilon \left[(N^{-1/2} + T/N)(1 + \mathfrak{C}_t^Y) + (N^{-1/4} + \sqrt{T/N}) \mathfrak{J}_t^Y \right]. \end{aligned} \tag{12}$$

Here $\vee$ denotes the maximum. The constants may depend on the fixed parameters, including $\upsilon, \zeta, C_p, C_0, L_0$.

We also report numerical sanity checks of non-spikiness and adaptive PWGF in Section C.

Curvature determines how long finite-sample errors persist. The ODE residual has scale $N^{-1/2} + T/N$, while the jump mismatch has scale $N^{-1/4} + \sqrt{T/N}$, apart from the common subpolynomial factor $C_\varepsilon \Xi_{N,T,p}^\varepsilon$. The T/N and $\sqrt{T/N}$ terms arise from controlling the residuals over the entire interval $[0, T]$; under $T \leq N^{1/4-\upsilon-\zeta}$, they are dominated by $N^{-1/2}$ and $N^{-1/4}$, respectively. These scales are then weighted by the actual curvature propagator G_κ^Y: $\mathfrak{C}_t^Y$ accumulates errors generated continuously, whereas $\mathfrak{J}_t^Y$ propagates jump mismatches. Keeping this propagator avoids replacing the accumulated curvature by a worst-case bound that can grow exponentially in T. Positive curvature makes past errors decay, while negative curvature can preserve or amplify them (Figure 1). A jump mismatch may therefore dominate just after a perturbation and later be forgotten, whereas continuous errors are generated throughout the flow.

The following example makes this curvature-dependent tracking bound concrete.

Example 3.1 (Growing-horizon tracking in a variance-plus-cosine model)**.** *In dimension one, consider the model in Section A.8,*

$$F(\mu) = \frac{a}{2} \operatorname{Var}(\mu) + \beta \cos(\pi m_\mu), \qquad m_\mu = \int x \, \mu(dx), \qquad a, \beta > 0.$$

Take a fixed bounded mean-zero initial law and one prescribed Gaussian jump at a fixed time $\tau > 0$ with fixed amplitude $\eta > 0$, independently of N. Writing $m_t = m_{\mu_t}$ and setting $T = N^{1/8}$, Proposition A.1 gives an event of probability at least $1 - N^{-1/16}$ on which Lemma A.35 and eqs. (202) *and* (204) *give*

$$\kappa_t^{Y,N} = \min\{a, -\beta\pi^2 \cos(\pi m_t)\}, \qquad \sup_{0 \leq t \leq T} (\mathfrak{C}_t^Y + \mathfrak{J}_t^Y) \leq C N^{1/16}.$$

Theorem 3.1 then gives, on the same event,

$$\sup_{0 \leq t \leq T} [r_t + |F(\rho_{\mathbf{X}_t}) - F(\mu_t)|] \leq C_\varepsilon N^{-3/16+\varepsilon}.$$

The bounds hold for each fixed $\varepsilon > 0$ and sufficiently large N, with fixed model parameters.

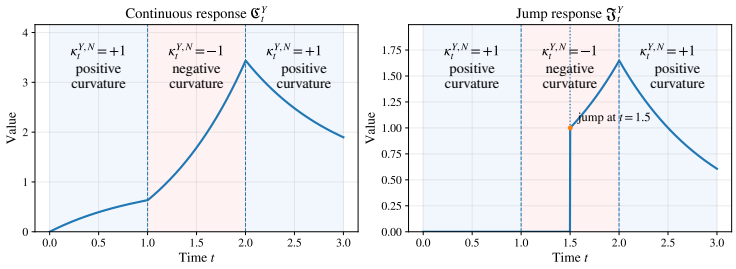


Figure 1: Illustration of the curvature-response factors in the tracking bound. We assume that $\kappa_t^{Y,N} = 1$ (positive curvature) if $t \in [0,1) \cup [2,3]$ and $\kappa_t^{Y,N} = -1$ (negative curvature) if $t \in [1,2)$. The factor $\mathfrak{C}_t^Y$ measures the remaining effect of unit ODE forcing, while $\mathfrak{J}_t^Y$ records a unit jump at $t = 1.5$, followed by amplification under negative curvature and decay under positive curvature.

4 Motivating geometry in a matrix factorization model

Matrix factorization is a product model for nonlinear-moment geometry (Kolda & Bader, 2009; Sidiropoulos et al., 2017; Anandkumar et al., 2014). Following the motivating matrix-factorization application of Yamamoto et al. (2025), we study a regularized product-measure extension and prove attraction along the anti-symmetric branch, transverse descent, and positivity of the second variation along spherical product pushforwards on an annulus and near the teacher. These population statements motivate the positive–negative–positive curvature-response scenario, but do not construct a connecting PWGF trajectory, transfer the second variations to the sampled curvature $\kappa_t^{Y,N}$, or verify its controlled-curvature path condition.

4.1 Product-measure model

Let $\mu = \mu_1 \otimes \mu_2$ have marginals on $\mathbb{S}^{d_w-1}$, and let $z \sim N(0, I_{d_w})$. For a vector-valued activation $\sigma : \mathbb{R} \to \mathbb{R}^{d_h}$, define the mean-field feature map $h_\nu(z) := \int \sigma(w^\top z)\, \nu(\mathrm{d}w)$, which averages features over weights $w \sim \nu$. Given a teacher $\mu^o = \mu_1^o \otimes \mu_2^o$, consider

$$F(\mu_1 \otimes \mu_2) := \frac{1}{2}\mathbb{E}_z\left[\left\| h_{\mu_1^o}(z) h_{\mu_2^o}(z)^\top - h_{\mu_1}(z) h_{\mu_2}(z)^\top \right\|_{\mathrm{F}}^2\right] + \frac{\lambda_{\mathrm{reg}}}{p_W} Z_\pm(\mu_1,\mu_2)^{p_W/2}, \tag{13}$$

with $\lambda_{\mathrm{reg}} > 0$, $p_W > 2$, $\nu_\star^\pm := \frac{1}{2}\delta_{a_\star} + \frac{1}{2}\delta_{-a_\star}$, and $Z_\pm(\mu) := Z_\pm(\mu_1,\mu_2) := \sum_{i=1}^2 W_{2,\mathbb{S}}^2(\mu_i, \nu_\star^\pm)$, where $a_\star \in \mathbb{S}^{d_w-1}$. The spherical distance convention is given in Section B.4.

4.2 Branches and standing conditions

For $0 < r_{\mathrm{cap}} < \pi/2$, let $\mathcal{P}_\pm(r_{\mathrm{cap}})$ denote the class of probability measures whose closed supports lie inside the two open spherical caps of radius r_{cap} centered at $\pm a_\star$, with mass $1/2$ in each cap. Throughout this section, impose the following restrictions.

Assumption 4.1 (Model restrictions; see Section B.4.3)**.** ***(i)*** *$\mu_i, \mu_i^o \in \mathcal{P}_\pm(r_{\mathrm{cap}})$ for $i = 1, 2$.* ***(ii)*** *The nonzero Hermite coefficients of σ occur at a finite set $\mathcal{O}$ of odd degrees.* ***(iii)*** *$\mu_1^o = \mu_2^o =: \nu^o$.*

For transverse descent in Proposition 4.1(iii), we additionally require a consistent average bias orthogonal to $a_\star$ across the two-sided latitude slices $|a_\star^\top w|$, with a positive margin away from the equator (Assumption B.3).

Using standard Gaussian-input moment coordinates (Ben Arous et al., 2021; Abbe et al., 2023; Bietti et al., 2022), set $T_{i,n} := \int w^{\otimes n} \mu_i(\mathrm{d}w)$, $S_n := (T_{1,n} + T_{2,n})/2$, and $A_n := (T_{1,n} - T_{2,n})/2$ for $n \in \mathcal{O}$. Define

$$\mathcal{M}_- := \{\mu_1 \otimes \mu_2 : \mu_2 = (-\mathrm{Id})_\# \mu_1\}, \qquad \mathcal{M}_+ := \{\mu_1 \otimes \mu_2 : \mu_1 = \mu_2\}. \tag{14}$$

Membership in $\mathcal{M}_-$ implies $S_n = 0$, and membership in $\mathcal{M}_+$ implies $A_n = 0$; the converses are not assumed. The central target is $\mu^\dagger := \nu_\star^\pm \otimes \nu_\star^\pm$.

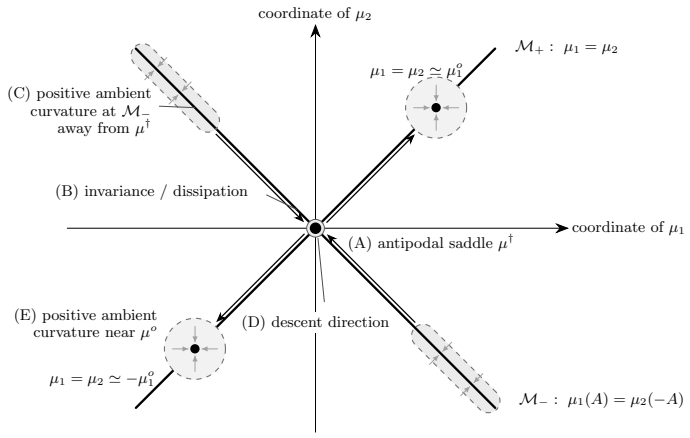


Figure 2: Schematic local geometry in Proposition 4.1: (A)–(B) the stationary target and its local attracting branch (i); (C) annular positive curvature (ii); (D) transverse descent (iii); and (E) positive curvature near the teacher (iv). The diagram's $-\mu_1^o$ denotes the reflected measure $(-\mathrm{Id})_{\#}\mu_1^o$. Pale arrows depict curvature, not gradient-flow directions. The arrows do not assert a connecting trajectory.

4.3 Local geometry

We use the spherical product pushforward second variation defined in Lemma B.6. Admissible fields are smooth product tangent fields preserving the caps for small perturbations; they need not be tangent to either branch. When the prescribed pushforward path is a product Wasserstein geodesic, this second variation is its usual geodesic second derivative.

Proposition 4.1 (Local geometry of the matrix factorization model)**.** *Under Assumption 4.1, with $\lambda_{\mathrm{reg}} > 0$ and $p_W > 2$, the following statements hold, each under the additional conditions stated in its item.*

- *(i)* ***Local attraction on the anti-symmetric branch (A–B).*** *For $2 < p_W < 3$, every globally well-posed characteristic product WGF on $\mathcal{M}_-$ that starts sufficiently close to $\mu^\dagger$ within the cap class remains in that class and converges to $\mu^\dagger$ in product Wasserstein distance. The precise initialization conditions and rate are given in Proposition B.5.*
- *(ii)* ***Positive ambient curvature on an annulus (C).*** *For every $\zeta_- > 0$, if λ_{reg} is sufficiently large, depending on ζ_- and the fixed model parameters, then at every $\mu \in \mathcal{M}_-$ with $Z_\pm(\mu) \geq \zeta_-^2$, the spherical product pushforward second variation is positive in every nonzero admissible direction. The quantitative threshold and lower bound are given in Corollary B.2.*
- *(iii)* ***Transverse descent at the target (D).*** *If $\mathcal{O} \neq \varnothing$ and the teacher satisfies Assumption B.3 at $a_\star$, ensuring a non-cancelling transverse teacher signal, there exists a same-density transverse perturbation μ_h from $\mu^\dagger$ such that, as $h \to 0$, $F(\mu_h) = F(\mu^\dagger) - \gamma_{\mathrm{esc}} h^2 + o(h^2)$, with $\gamma_{\mathrm{esc}} > 0$. Its construction is given in Proposition B.7. The regularizer contributes $O(|h|^{p_W}) = o(h^2)$.*
- *(iv)* ***Positive ambient curvature near the teacher (E).*** *If $Z_\pm(\mu_+^o) > 0$, where $\mu_+^o := \nu^o \otimes \nu^o$, the spherical product pushforward second variation has a positive uniform margin in a sufficiently small product Wasserstein neighborhood of μ_+^o, relative to the cap class. The existence of the margin and neighborhood is proved in Proposition B.9. The teacher need not be stationary for the regularized objective.*

5 Technical assumptions and proof overview

5.1 Technical Assumptions

We state the conditions used in the main theorem and proof sketch; Section A.2 gives their full analytic formulation, and Table 1 records how each model section relates to the theorem's assumptions and conclusions. The conditions below are verified for Example 3.1 in Lemma A.34.

Assumption 5.1 (Structured moment representation)**.** *The full conditions of Assumption A.1 hold. We use $F(\rho) = G(M(\rho))$, with $M(\rho) := (\int \phi_m \, d\rho)_{m\geq 1}$, as defined in Definition A.3. The features have controlled polynomial growth through third spatial derivatives, dominated by the decay of the moment weights $a_m = e^{-c_a m^{\kappa_{\mathrm{dec}}}}$, with $c_a > 0$ and $\kappa_{\mathrm{dec}} > 1$; see Equation* (16) *and assumption A.1. The functional G is C^3 on one open weighted-moment domain containing all finite-scale feasible moment regions, with a uniform weighted third-derivative bound. All structural parameters, including the anchor derivative norms and this uniform bound, are fixed independently of N, T, p.*

The localized drift and Hessian bounds derived in Lemmas A.4 and A.6 may depend on the spatial localization radius D.

Assumption 5.2 (Regularity of the coupled particle systems)**.** *Fix $T > 0$. The coupled particle systems $\mathbf{X}_t = (X_t^1, \ldots, X_t^N)$ and $\mathbf{Y}_t = (Y_t^1, \ldots, Y_t^N)$ satisfy the following conditions uniformly on $[0, T]$. The $O(1)$ bounds below are uniform in N over the horizons considered in Theorem 3.1.*

(i) Unconditional moment bound. *For fixed $\alpha \in (0, 2]$,*
$\sup_{t\leq T, q\geq 1} q^{-1/\alpha} \max\{(\mathbb{E}\|X_t^1\|^q)^{1/q}, (\mathbb{E}\|Y_t^1\|^q)^{1/q}\} = O(1)$.

(ii) Temporal regularity. *All particle paths are right-continuous with left limits at the jumps and share a deterministic $O(1)$-Lipschitz bound on every ODE branch, almost surely.*

(iii) Non-spikiness of r_t. *The coupling error r_t from Section 2.4 is assumed, almost surely, not to concentrate on a small number of particles: for some $A_{\mathrm{sp}} = O(1)$, $\frac{1}{N}\sum_{i=1}^N \|X_t^i - Y_t^i\|^3 \leq A_{\mathrm{sp}} r_t^3$ $(0 \leq t \leq T)$. This condition controls the cubic Taylor remainder in the particle positions by $O(r_t^3)$; see Lemma A.11 and eq.* (86)*. In Example 3.1, all paired errors coincide, giving $A_{\mathrm{sp}} = 1$ (Lemma A.34(v)).*

Assumption 5.3 (Prescribed schedule and population-first Gaussian jumps)**.** *The common times $(t_\ell)_{\ell=1}^{L_{\mathrm{jump}}}$ and amplitudes $\eta_\ell \geq 0$ are deterministic, with $L_{\mathrm{jump}} = O(1)$ in Theorem 3.1. The initial law μ_0 is deterministic; the common initial positions $X_0^i = Y_0^i$, $i = 1, \ldots, N$, are sampled independently from μ_0. Population innovations are independent of that sample and fresh at each jump. The population Gaussian fields and the empirical transitions satisfy Assumption A.2.*

5.2 Proof sketch

We explain how the ODE and jump comparisons produce the response factors in Definition 3.1. The calculations below are schematic: $\simeq$ retains the leading linearized drift, and $\lesssim$ suppresses localized coefficients and Gaussian probability factors quantified in the appendix.

The reference particles are i.i.d. with law μ_t conditional on $\mathcal{G}_{\mathrm{pop}}$ by Lemmas A.14 and A.19, enabling concentration (background: Sznitman (1991); Fournier & Guillin (2015)). We use Γ_t^N and $\widetilde{\Gamma}_t^N$ from Section A.1 for drift/Hessian fluctuations and covariance-sampling error at jumps, respectively.

ODE intervals (Section A.3). Write $\Delta_t := \mathbf{X}_t - \mathbf{Y}_t$, with particle blocks indexed by i, so that $r_t = N^{-1/2}\|\Delta_t\|_{\mathbb{R}^{dN}}$. For $r_t > 0$, the chain rule and the ODEs in eqs. (5) and (8) give the equalities below. After adding and subtracting the reference empirical drift, then linearizing and comparing Hessians (Lemmas A.7 and A.8), we informally have

$$r_t \dot{r}_t = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} r_t^2 = -\frac{1}{N}\sum_{i=1}^N \big\langle \Delta_t^i, \underbrace{\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i) - \nabla_W F(\mu_t, Y_t^i)}_{\simeq (\mathrm{H}_t^Y \Delta_t)^i \text{ (Taylor)}} \big\rangle \simeq -\frac{1}{N}\Delta_t^\top \mathrm{H}_t^Y \Delta_t.$$

By Definition 3.1, the quadratic form is bounded above by $-\kappa_t^{Y,N} r_t^2$; dividing by r_t gives the curvature term $-\kappa_t^{Y,N} r_t$. Restoring the errors, on the localized good events and while $r_t \leq R$, Lemma A.12, (106), gives $\dot{r}_t \leq -\kappa_t^{Y,N} r_t + c_D R r_t + \Gamma_t^N(1 + r_t)$. Here c_D is defined in (101); the same lemma also handles $r_t = 0$.

Jump times (Section A.4). Using the empirical and aligned population factors M_X, B_{pop} and shared Gaussian vector $\mathbf{g}_\ell^N$ from Section 2.3, we substitute the jump updates in eqs. (6) and (9) and apply the triangle inequality:

$$r_{t_\ell} = \frac{1}{\sqrt{N}}\big\|\Delta_{t_\ell -} + \eta_\ell (M_X - B_{\mathrm{pop}})\mathbf{g}_\ell^N\big\|_{\mathbb{R}^{dN}} \leq r_{t_\ell -} + \frac{\eta_\ell}{\sqrt{N}}\big\|(M_X - B_{\mathrm{pop}})\mathbf{g}_\ell^N\big\|_{\mathbb{R}^{dN}}.$$

Let M_Y be the empirical Hessian factor at the reference configuration just before the jump. We split $M_X - B_{\mathrm{pop}} = (M_X - M_Y) + (M_Y - B_{\mathrm{pop}})$. Smoothness controls the contribution of the first difference by a multiple of $r_{t_\ell -}$. For the remaining mismatch, Gaussian concentration and the aligned coupling give schematically, with all time-t quantities evaluated at $t_\ell -$,

$$\frac{\|(M_Y - B_{\mathrm{pop}})\mathbf{g}_\ell^N\|_{\mathbb{R}^{dN}}}{\sqrt{N}} \lesssim \underbrace{\sqrt{\frac{1}{N}\|M_Y^2 - B_{\mathrm{pop}}B_{\mathrm{pop}}^\top\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}}}_{\text{square root of covariance mismatch}} \lesssim \sqrt{\Gamma_t^N} + \sqrt{\widetilde{\Gamma}_t^N}.$$

Here $*$ denotes the nuclear norm. The (i,j)-block of the covariance mismatch is $K_{\rho_{\mathbf{Y}_t}}(Y_t^i, Y_t^j) - K_{\mu_t}(Y_t^i, Y_t^j)$. Using the kernel definition in Section 2.1, the difference $K_{\rho_{\mathbf{Y}_t}}(x,y) - K_{\mu_t}(x,y)$ consists schematically of

$$\underbrace{\text{Hessian approximation}}_{\text{as in the ODE step}} + \int \nabla_W^2 F(\mu_t; x, z)\, \nabla_W^2 F(\mu_t; z, y)\, (\rho_{\mathbf{Y}_t} - \mu_t)(dz).$$

Expanding the difference of the two Hessian products bounds the first term linearly in the Hessian approximation error, with localized coefficients suppressed. The second term is an empirical integration error. These comparisons are established in Lemmas A.13, A.16 and A.17. Combining the two contributions yields $r_{t_\ell} \lesssim (1+\eta_\ell) r_{t_\ell -} + \eta_\ell(\sqrt{\Gamma_{t_\ell -}^N} + \sqrt{\widetilde{\Gamma}_{t_\ell -}^N})$: a jump amplifies past error and adds a new error. The square-root dependence explains why an $N^{-1/2}$ covariance-sampling bound contributes $N^{-1/4}$ to the jump estimate. For background on Gaussian covariance alignment, see Masarotto et al. (2019).

From local errors to response factors. Between jumps, the homogeneous equation $\dot{r}_t = -\kappa_t^{Y,N} r_t$ has solution $r_t = G_\kappa^Y(t,s) r_s$, where $G_\kappa^Y(t,s) = \exp\Big(-\int_s^t \kappa_u^{Y,N}\, du\Big)$. Hence an error created at time s receives this weight at time t. Starting from $r_0 = 0$, integrating the continuous forcing and summing the jump forcing gives an integral and a sum, respectively. Taking uniform residual bounds outside them yields schematically

$$r_t \lesssim (\sup_{s\le T} \Gamma_s^N) \underbrace{\int_0^t G_\kappa^Y(t,s)\, ds}_{\mathfrak{C}_t^Y} + (\sqrt{\sup_{s\le T}\Gamma_s^N} + \sqrt{\max_{\ell \le L_{\mathrm{jump}}} \widetilde{\Gamma}_{t_\ell -}^N}) \underbrace{\textstyle\sum_{\ell: t_\ell \le t} \eta_\ell G_\kappa^Y(t, t_\ell)}_{\mathfrak{J}_t^Y}.$$

The maximum is zero if there are no jumps. The recursive comparison in Lemma A.31, eqs. (166) and (184), retains the extra nonlinear amplification and the products of jump multipliers.

Residual bounds and closure. Conditioning only on $\mathcal{G}_{\mathrm{pop}}$, moment localization and weighted-feature covariance concentration give, with probability at least $1-p$ and $\Xi = \Xi_{N,T,p}$ from Theorem 3.1,

$$\sup_{t\le T} \Gamma_t^N \le C_\varepsilon \Xi^\varepsilon (N^{-1/2} + T/N), \qquad \max_{\ell \le L_{\mathrm{jump}}} \widetilde{\Gamma}_{t_\ell -}^N \le C_\varepsilon \Xi^\varepsilon N^{-1/2};$$

see Lemma A.25 and corollary A.3. Under the theorem's conditions, Theorem A.1 and corollary A.4 close the localized estimate and absorb the extra amplification into $C_\varepsilon \Xi^\varepsilon$. Lemma A.33 and eq. (190) give the objective bound.

6 Conclusion

We quantified finite-particle tracking for Hessian-guided PWGF under prescribed common jump schedules. The coupling reduces Gaussian jump error to covariance comparison, while sampled curvature propagates residuals into tracking and objective bounds.

Limitations & future directions. Beyond the variance-plus-cosine example (Section A.8), deriving the controlled-curvature path condition and non-spikiness from model primitives remains open, as does synchronizing adaptive jump times.

AI USE STATEMENT

ChatGPT models assisted with brainstorming, writing and editing, locating candidate references, and checking mathematical arguments. This included formulating and refining mathematical claims, detailed proof calculations, revising author-written drafts, supplementing incomplete derivations, exploring proof repairs, preparing LaTeX and verification materials, editing schematic figures, designing the experimental setup, developing code for the synthetic experiments, and creating experimental figures. They are not components of the optimization method studied in this paper. The authors have checked the proofs. Responsibility for the final statements and proofs remains with the authors.

REFERENCES

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Birkhäuser Basel, 2005. URL https://link.springer.com/book/10.1007/b137080.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014. ISSN 1532-4435. URL https://jmlr.org/papers/v15/anandkumar14b.html.

Michael Arbel, Anna Korba, Adil SALIM, and Arthur Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/944a5ae3483ed5c1e10bbccb7942a279-Paper.pdf.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021. URL https://jmlr.org/papers/v22/20-1288.html.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3880–3888, 2016. URL https://proceedings.neurips.cc/paper/2016/file/b139e104214a08ae3f2ebcce149cdf6e-Paper.pdf.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9768–9783. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3fb6c52aeb11e09053c16eabee74dd7b-Paper-Conference.pdf.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. URL https://academic.oup.com/book/26549.

Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field Langevin dynamics. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 61(4): 2357–2404, November 2025. doi: 10.1214/24-AIHP1499. URL https://hal.science/hal-05349896.

Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Stochastic control, entropic interpolation and gradient flows on wasserstein product spaces, 2016. URL https://arxiv.org/abs/1601.04891.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

Casey Chu, Jose Blanchet, and Peter Glynn. Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1213–1222. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/chu19a.html.

Lauren Conger, Franca Hoffmann, Eric Mazumdar, and Lillian J. Ratliff. Coupled wasserstein gradient flows for min-max and cooperative games, 2025. URL https://arxiv.org/abs/2411.07403.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/abea47ba24142ed16b7d8fbf2c740e0d-Paper.pdf.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3–4):707–738, 2015. URL https://link.springer.com/article/10.1007/s00440-014-0583-7.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of the Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 797–842, 2015. URL https://proceedings.mlr.press/v40/Ge15.html.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1233–1242, 2017. URL https://proceedings.mlr.press/v70/ge17a.html.

Valeria Giunta, Thomas Hillen, Mark A. Lewis, and Jonathan R. Potts. Weakly nonlinear analysis of a two-species non-local advection–diffusion system. *Nonlinear Analysis: Real World Applications*, 78:104086, 2024. ISSN 1468-1218. doi: 10.1016/j.nonrwa.2024.104086. URL https://www.sciencedirect.com/science/article/pii/S1468121824000269.

Margalit Glasgow, Denny Wu, and Joan Bruna. Mean-field analysis of polynomial-width two-layer neural network beyond finite time horizon. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2461–2539. PMLR, 2025. URL https://proceedings.mlr.press/v291/glasgow25a.html.

Wolfgang Hackbusch. *Tensor Spaces and Numerical Tensor Calculus*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-28027-6. URL https://link.springer.com/book/10.1007/978-3-642-28027-6.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1042–1085. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/jin18a.html.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. 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*, pp. 24527–24561. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kim24af.html.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. doi: 10.1137/07070111X. URL https://doi.org/10.1137/07070111X.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Valentina Masarotto, Victor M. Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A*, 81(1):172–213, 2019. doi: 10.1007/s13171-018-0130-1. URL https://doi.org/10.1007/s13171-018-0130-1.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the NTK via a power series expansion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.07844.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9741–9757. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/nitanda22a.html.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Particle dual averaging: optimization of mean field neural network with global convergence rate analysis*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114010, nov 2022b. doi: 10.1088/1742-5468/ac98a8. URL https://doi.org/10.1088/1742-5468/ac98a8.

Atsushi Nitanda, Anzelle Lee, Damian Tan Xing Kai, Mizuki Sakaguchi, and Taiji Suzuki. Propagation of chaos for mean-field Langevin dynamics and its application to model ensemble. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46586–46610. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/nitanda25a.html.

Giovanni Peccati and Murad S Taqqu. *Wiener Chaos: Moments, Cumulants and Diagrams: A survey with computer implementation*, volume 1. Springer Science & Business Media, 2011. URL https://link.springer.com/book/10.1007/978-88-470-1679-8.

Robert T. Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, 1970. doi: 10.1007/BF01645492. URL https://doi.org/10.1007/BF01645492.

Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, July 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2690524. URL https://doi.org/10.1109/TSP.2017.2690524.

Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_JScUk9TBUn.

Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin (ed.), *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pp. 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg. ISBN 978-3-540-46319-1. URL https://link.springer.com/chapter/10.1007/BFb0085169.

Naoya Yamamoto, Juno Kim, and Taiji Suzuki. Hessian-guided perturbed wasserstein gradient flows for escaping saddle points. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.16974.

A A General Comparison Argument via the Auxiliary Particle System

Notations. For nonnegative quantities a, b, we write $a \lesssim b$ if $a \le Cb$ for a constant $C > 0$ independent of the main asymptotic parameters, such as the particle number N, horizon T, and failure probability p. Equivalently, $a = O(b)$ means $a \lesssim b$. We write $a = o(b)$ if $a/b \to 0$ in the prescribed asymptotic regime. Similarly, $a = \Omega(b)$ means $b = O(a)$, and $a = \omega(b)$ means $b = o(a)$. We write $a = \Theta(b)$ if both $a = O(b)$ and $a = \Omega(b)$ hold. Finally, $a = \widetilde{O}(b)$ hides polylogarithmic factors in the main asymptotic parameters. The hidden constants may depend on fixed problem parameters, the dimension, and structural smoothness constants, but not on N, T, or p.

For normed spaces E, F, $\mathcal{L}(E; F)$ denotes the space of bounded linear maps from E to F, and $\mathcal{L}(E_1, \dots, E_k; F)$ denotes the space of bounded k-linear maps from $E_1 \times \cdots \times E_k$ to F. When $E = F$, we write $\mathcal{L}(E)$. For a linear map $A : \mathbb{R}^d \to \mathbb{R}^d$, $\|A\|_{d\to d} := \sup_{\|v\|_2=1} \|Av\|_2$ denotes its Euclidean operator norm. We use $\|\cdot\|_{\mathrm{HS}}$ for the Hilbert–Schmidt norm and $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm; in finite-dimensional matrix spaces these coincide, but we reserve $\|\cdot\|_{\mathrm{HS}}$ for Hilbert-space operators or tensor-valued feature maps, and $\|\cdot\|_{\mathrm{F}}$ for ordinary finite matrices.

Vector fields, configurations, and empirical measures. For a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and vector fields $v, w : \mathbb{R}^d \to \mathbb{R}^d$, define

$$\|v\|^2_{L^2(\mu)} := \int \|v(x)\|^2 \, \mu(dx), \qquad \langle v, w\rangle_{L^2(\mu)} := \int \langle v(x), w(x)\rangle \, \mu(dx).$$

For a configuration $\mathbf{x} = (x^1, \dots, x^N) \in (\mathbb{R}^d)^N$, set

$$\|\mathbf{x}\|^2_N := \frac{1}{N}\sum_{i=1}^N \|x^i\|^2, \qquad \rho_{\mathbf{x}} := \frac{1}{N}\sum_{i=1}^N \delta_{x^i}, \qquad \xi^N(\mathbf{x}) := (\xi(x^1), \dots, \xi(x^N)).$$

We abbreviate $\rho_{X_t} := \rho_{\mathbf{X}_t}$ and $\rho_{Y_t} := \rho_{\mathbf{Y}_t}$. The coupled discrepancy is

$$\delta_{t,i} := X^i_t - Y^i_t, \qquad r^2_t := \frac{1}{N}\sum_{i=1}^N \|\delta_{t,i}\|^2 = \|\mathbf{X}_t - \mathbf{Y}_t\|^2_N.$$

The normalized configuration inner product is

$$\langle \mathbf{u}, \mathbf{v}\rangle_N := \frac{1}{N}\sum_{i=1}^N \langle u^i, v^i\rangle.$$

Gradient, Hessian, and jump covariance. The Wasserstein gradient is $\nabla_W F(\mu, x) \in \mathbb{R}^d$. We distinguish its measure derivative from its spatial Jacobian:

$$\nabla^2_W F(\mu; x, y) := \nabla^2_W F(\mu; x, y) \in \mathbb{R}^{d\times d},$$
$$\nabla_x \nabla_W F(\mu, x) := \nabla_x \nabla_W F(\mu, x) \in \mathbb{R}^{d\times d}.$$

Use the following operators on their natural domains in $L^2(\mu; \mathbb{R}^d)$:

$$(H_\mu v)(x) := \int \nabla^2_W F(\mu; x, y) v(y) \, \mu(dy),$$
$$(H'_\mu v)(x) := \nabla_x \nabla_W F(\mu, x) v(x), \qquad \mathcal{H}_\mu := H_\mu + H'_\mu.$$

In particular, the spatial multiplication operator has domain

$$\mathcal{D}(H'_\mu) := \{v \in L^2(\mu;\mathbb{R}^d) : \nabla_x \nabla_W F(\mu,\cdot)v(\cdot) \in L^2(\mu;\mathbb{R}^d)\}, \qquad \mathcal{D}(\mathcal{H}_\mu) = \mathcal{D}(H_\mu) \cap \mathcal{D}(H'_\mu).$$

Here $\mathcal{D}(H_\mu)$ consists of the fields for which the displayed kernel integral exists and defines an $L^2(\mu)$ field. For $v \in \mathcal{D}(\mathcal{H}_\mu)$ for which the pushforward perturbation and the indicated derivatives are defined, the full second-variation identity is

$$\frac{d^2}{dh^2}\Big|_{h=0} F((\mathrm{Id} + hv)_{\#}\mu) = \langle v, \mathcal{H}_\mu v\rangle_{L^2(\mu)}.$$

The tracking curvature retains both terms of $\mathcal{H}_\mu$. The additional spatial-Hessian bound in proposition A.2 makes H'_μ bounded; no such global bound is imposed here. In contrast, the Hessian-guided jump covariance uses the integral term:

$$K_\mu(x,y) := \int \nabla_W^2 F(\mu;x,z)\nabla_W^2 F(\mu;z,y)\,\mu(dz), \qquad \xi_\mu \sim \mathrm{GP}(0,K_\mu).$$

Here $\mathrm{GP}(0,K_\mu)$ denotes a centered Gaussian process with covariance kernel K_μ. The finite Gaussian coordinates used to compare two jumps are constructed from the population field in lemma A.14. For a matrix or trace-class operator A, we use the nuclear norm

$$\|A\|_* := \mathrm{tr}((A^*A)^{1/2}),$$

where A^* denotes the Hilbert adjoint and equals $A^\top$ for real matrices.

Conventions for the fixed-schedule comparison. The jump count is L_{jump}. We take $0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \le T$, put $t_0 = 0$ and $t_{L_{\mathrm{jump}}+1} = T$, and allow $L_{\mathrm{jump}} = 0$. Paths are right-continuous: Z_{t_ℓ} is the post-jump value and $Z_{t_\ell -} = \lim_{s\uparrow t_\ell} Z_s$ is the pre-jump value. Jump covariances and jump residuals are evaluated at $t_\ell-$. Empty sums and products are zero and one, respectively.

Population conditioning and comparison history. The population initial law and schedule are deterministic. Population Gaussian innovations are independent of the iid initial sample. Let $\mathcal{G}_\ell$ be the sigma-field generated by the population fields through jump ℓ, and let $\mathcal{G}_{\mathrm{pop}} = \mathcal{G}_{L_{\mathrm{jump}}}$. The population path satisfies $\mu_t = \mathrm{Law}(Y_t^1 \mid \mathcal{G}_{\mathrm{pop}})$; the conditional product law is proved in lemma A.19. The notation $\mathcal{H}_{\ell-}$ denotes the full pre-jump history, including $X_{t_\ell-}, Y_{t_\ell-}, \mu_{t_\ell-}$, all previous population innovations and auxiliary normals, and the initial sample. It does not include the current or future innovations. Finite Gaussian comparison coordinates are standard Gaussian conditionally on this history; they are not the conditioning object in the product law for Y.

A.1 Overview of the finite-particle comparison

This appendix expands the proof architecture summarized in section 5, especially the four-step sketch in section 5.2. The population PWGF mechanism itself is treated as the reference dynamics, in the sense of the Hessian-guided construction of Yamamoto et al. (2025). The purpose of the present appendix is the finite-particle empirical error analysis: we quantify how the implementable particle system tracks the prescribed mean-field PWGF path.

Classical globally Lipschitz propagation-of-chaos arguments close the reference-particle stability estimate by Gronwall with a worst-case Lipschitz constant (Sznitman, 1991). Here the same architecture is driven instead by the curvature propagator associated with the sampled curvature.

The comparison is organized through the auxiliary particle system $\boldsymbol{Y}_t = (Y_t^1, \ldots, Y_t^N)$:

$$\rho_{\boldsymbol{X}_t} \quad \longleftrightarrow \quad \rho_{\boldsymbol{Y}_t} \quad \longleftrightarrow \quad \mu_t.$$

Here $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^N)$ is the actual empirical system, whose drift and jump covariance are evaluated at $\rho_{\boldsymbol{X}_t}$; μ_t is the population PWGF path; and $\boldsymbol{Y}_t$ starts from the same initial sample and uses the population drift and the original population Gaussian field along μ_t. Procrustes/Wasserstein comparisons of covariance factors provide related mathematical background (Masarotto et al., 2019). The population-first conditional construction required here is proved in lemma A.14: it constructs the empirical Gaussian coordinates jointly with the field evaluations, preserving both marginal updates. The reference conditional product law conditions only on the population fields. Thus the comparison separates two effects: the pathwise stability of $\boldsymbol{X}_t$ relative to $\boldsymbol{Y}_t$, and the concentration of the conditionally i.i.d. reference cloud $\boldsymbol{Y}_t$ around μ_t.

Condition or conclusion	General comparison	Variance–cosine example	Matrix-factorization case study
Moment structure and weighted smoothness	Assumed in assumption A.1	Verified in lemma A.34(i)	Separate product-model conditions in assumption 4.1
Gaussian jumps and schedule	Prescribed by assumption A.2	One prescribed common jump, verified in lemma A.34(ii)	—
Moment growth and ODE regularity	Assumed in assumptions A.3 and A.4	Verified in lemma A.34(iii)–(iv)	—
Non-spikiness	Assumed in assumption A.5	Verified with $A_{\mathrm{sp}} = 1$ in lemma A.34(v)	—
Curvature-response mechanism	Nonlinear stability in theorem A.1; N^{υ}-scaled curvature-response bounds in corollary A.4	High-probability response bounds in lemma A.35; corollary A.4 applies with $\upsilon = \zeta = 1/16$	Population-level local geometry in proposition 4.1: attraction, transverse descent, and positive spherical product pushforward second variation in two regions
Finite-particle conclusion	Conditional tracking and objective comparison	Growing-horizon conclusion in proposition A.1	Open

Table 1: Relation of the model sections to the comparison theory. The matrix-factorization column records a separate population curvature-response case study; — denotes items outside its scope, and "Open" denotes the bridge to finite-particle tracking.

How this appendix matches the main proof sketch. The detailed proof follows the same four pieces as section 5.2. First, section A.2 defines the common initial coupling, the common-seed Gaussian jump construction, the moment-Lipschitz smoothness assumptions, and the localization event. Second, section A.3 proves the continuous-time contraction-versus-residual bound for

$$r_t \coloneqq \frac{1}{\sqrt{N}} \|\boldsymbol{X}_t - \boldsymbol{Y}_t\|_2.$$

Third, section A.4 controls the perturbation mismatch at prescribed jump times. Fourth, section A.5, together with section A.5.2 and section A.5.3, supplies the concentration estimates needed to close the deterministic inequality.

Core theorem: prescribed common jump schedule. The core result of this appendix is the fixed-schedule comparison in section A.6. In this layer the population and empirical systems are compared along a prescribed common jump schedule $0 = t_0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \le T$,

with $L_{\mathrm{jump}} = O(1)$ in Corollary A.4. On the population moment and sample localization events, and under anti-spikiness, the ODE estimate is

$$\dot{r}_t \le -\kappa_t^{Y,N} r_t + c_D r_t^2 + \Gamma_t^N r_t + \Gamma_t^N.$$

At each prescribed jump, the comparison uses only pre-jump residuals:

$$r_{t_\ell} \le (1 + a_\ell) r_{t_\ell -} + a_\ell \left(\sqrt{\Gamma^N_{t_\ell -}} + \sqrt{\widetilde{\Gamma}^N_{t_\ell -}}\right), \qquad a_\ell = c_J(D) q_p \eta_\ell.$$

The finite Gaussian coupling is constructed after generating the population field; the reference product law conditions only on the population fields. The exact result in Theorem A.1 retains the nonlinear condition $T(\delta + c_D R) \le 1/4$. Its failure probability is bounded within that nonlinear stability event and the regularity event. The sufficient horizon $T \le N^{1/4 - \upsilon - \zeta}$ in corollary A.4 also specifies the curvature-response bounds and uniform structural constants it uses.

The two stochastic residuals. The continuous residual is

$$\Gamma_t^N \coloneqq \sum_{m\geq 1} \gamma_m \Delta_{m,t}^N, \qquad \Delta_{m,t}^N \coloneqq \|M_m(\rho_{\boldsymbol{Y}_t}) - M_m(\mu_t)\|.$$

It controls the population-to-reference empirical error in the drift, Hessian, and functional terms. The jump covariance residual is

$$\widetilde{\Gamma}_t^N \coloneqq \frac{1}{N}\left\| \int \mathbf{U}_t(z)\mathbf{U}_t(z)^\top (\mu_t - \rho_{\boldsymbol{Y}_t})(\mathrm{d}z) \right\|_*,$$

where

$$\mathbf{U}_t(z) \coloneqq \left(\nabla_W^2 F(\mu_t; Y_t^1, z), \ldots, \nabla_W^2 F(\mu_t; Y_t^N, z)\right)^\top.$$

The first residual enters additively along ODE intervals, whereas the second enters through a square-root covariance perturbation at jumps. Consequently, even when both residuals concentrate at nearly $N^{-1/2}$ scale, the jump contribution generally produces an $N^{-1/4}$ bottleneck. Section A.5.2 and section A.5.3 supply the concentration estimates for Γ_t^N and $\widetilde{\Gamma}_t^N$, respectively.

From particle tracking to functional error. After the fixed-schedule r_t estimate is proved, section A.7 converts the tracking bound into a bound on

$$\left|F(\rho_{\boldsymbol{X}_t}) - F(\mu_t)\right|.$$

The key point is that no new stochastic object is introduced: the functional gap is bounded by the same pair (r_t, Γ_t^N). The common weights

$$\gamma_m(D) = \mathrm{L}_{K_{\mathrm{pop}},D} a_m$$

from lemma A.6 control both the dynamical comparison and the functional comparison.

An example on a growing horizon. Proposition A.1 treats a variance-plus-cosine model with bounded mean-zero initialization and one prescribed Gaussian jump. Its proof computes the sampled curvature, bounds the probability of escape, and checks finite-N stability. After positive curvature returns, the jump curvature-response factor decays, yielding tracking and a bound on the objective gap to the minimum for $C\log N \leq t \leq N^{1/8}$.

Fixed-time supplements. Lemma A.36 compares $\mathcal{G}_t = \int \|\nabla_W F(\mu_t, x)\|^2 \, \mu_t(dx)$ and $\mathcal{G}_t^N = N^{-1}\sum_i \|\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i)\|^2$ at one deterministic non-jump time. The margin condition (214) gives agreement of threshold decisions at that time. Separately, proposition A.2 transfers negative population curvature to the reference sample for finite-feature objectives under a spatial-Hessian condition. These results are not inputs to the tracking proof and do not control whole-path curvature-response factors or synchronize adaptive jump times.

A.2 Settings and Assumptions

Before defining the estimates, we fix the coupled population-reference-empirical dynamics and the common jump schedule used throughout the comparison.

Definition A.1. *Let $U_1, \ldots, U_N$ be i.i.d. with deterministic law μ_0, independently of the population innovations, and set $X_0^i = Y_0^i = U_i$. Fix deterministic jump times $0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \leq T$ and amplitudes $\eta_\ell \geq 0$. Write $t_0 = 0$ and $t_{L_{\mathrm{jump}}+1} = T$; zero jumps are allowed. On each nonempty interval $(t_\ell, t_{\ell+1})$, $0 \leq \ell \leq L_{\mathrm{jump}}$, let*

$$\partial_t X_t^i = -\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i), \qquad \partial_t Y_t^i = -\nabla_W F(\mu_t, Y_t^i), \qquad \partial_t \mu_t = \mathrm{div}(\mu_t \nabla_W F(\mu_t)).$$

The paths are right-continuous, and $\mathbf{X}_{t_\ell -}, \mathbf{Y}_{t_\ell -}$ denote their actual left limits. At a jump,

$$\mathbf{X}_{t_\ell} = \mathbf{X}_{t_\ell -} + \eta_\ell \xi_{\ell,X}^N(\mathbf{X}_{t_\ell -}), \qquad \mathbf{Y}_{t_\ell} = \mathbf{Y}_{t_\ell -} + \eta_\ell \xi_{\ell,Y}^N(\mathbf{Y}_{t_\ell -}), \tag{15}$$

where $\xi_{\ell,Y}^N$ evaluates the population field with covariance $K_{\mu_{t_\ell -}}$, and the empirical perturbation has the transition specified in Assumption A.2. The population law is pushed forward by the same population field. Fresh innovations are independent across jumps; the resulting fields can have covariances depending on earlier population randomness. Their finite coupling is constructed in

Lemma A.14. We assume that the population path admits the following characteristic realization: conditional on the population innovations, the ODE and jump maps compose into a common measurable map $\mathcal{T}_t$ (and the corresponding left-trace maps) such that

$$Y_t^i = \mathcal{T}_t(U_i), \qquad \mu_t = (\mathcal{T}_t)_{\#}\mu_0.$$

On each ODE branch, $\mathcal{T}_t$ is generated by the population characteristic ODE; at a jump, it includes the population pushforward stated above. This is the realization used in lemmas A.14 and A.19.

A.2.1 ASSUMPTIONS

The following assumptions specify the moment representation, Gaussian perturbations, and pathwise regularity used in the comparison. The feature parameters and the anchor norms below are fixed independently of N, T, p. The localization radius and conditional population scale will be chosen from N, T, p; their dependence remains in every local comparison coefficient.

Weighted moments and derivative norms. Fix $\alpha \in (0, 2]$. For Hilbert spaces H_m and features $\phi_m : \mathbb{R}^d \to \mathsf{H}_m$, write

$$\begin{aligned} M_m(\nu) &:= \int \phi_m(x)\, \nu(dx) \in \mathsf{H}_m, \qquad M(\nu) := (M_m(\nu))_{m\ge 1}, \\ a_m &:= e^{-c_a m^{\kappa_{\mathrm{dec}}}}, \qquad \|h\|_{\mathsf{H}_a}^2 := \sum_{m\ge 1} a_m^2 \|h_m\|_{\mathsf{H}_m}^2, \qquad \mathsf{H}_a := \{h : \|h\|_{\mathsf{H}_a} < \infty\}, \end{aligned} \tag{16}$$

where $c_a > 0$, $\kappa_{\mathrm{dec}} > 1$, and each feature integral is a Bochner integral whenever it exists. Scalar features are included by taking $\mathsf{H}_m = \mathbb{R}$ for every m. For finite $K > 0$ and $D \ge 1$, set

$$\begin{aligned} \mathcal{P}_{\alpha,K} &:= \left\{ \nu \in \mathcal{P}_2(\mathbb{R}^d) : \sup_{q\ge 1} q^{-1/\alpha} \left(\int \|x\|_{\mathbb{R}^d}^q \, \nu(dx) \right)^{1/q} \le K \right\}, \\ \mathcal{P}_D &:= \{\nu \in \mathcal{P}_2(\mathbb{R}^d) : \nu(\{x : \|x\|_{\mathbb{R}^d} \le D\}) = 1\}. \end{aligned} \tag{17}$$

Here K is a variable class scale; K_0 in assumption A.3 will bound unconditional particle moments.

Definition A.2 (Localized moment region)**.** *For finite $K > 0$ and $D \ge 1$, the feasible moment region is*

$$\mathcal{V}_{K,D} := \mathrm{conv}\{M(\nu) : \nu \in \mathcal{P}_{\alpha,K} \cup \mathcal{P}_D\}. \tag{18}$$

The feature condition below makes these moment vectors elements of H_a, as shown in lemma A.2(i). The convex hull here consists of finite convex combinations.

For an r-linear map $A : E_1 \times \cdots \times E_r \to E$, use the norm

$$\|A\|_{E_1\times\cdots\times E_r \to E} := \sup_{\|v_j\|_{E_j}\le 1,\ 1\le j\le r} \|A[v_1, \ldots, v_r]\|_E.$$

All Fréchet derivatives of G below are with respect to the weighted Hilbert norm in (16). If $\mathbf{e}_m u$ inserts $u \in \mathsf{H}_m$ in coordinate m and zero in the other coordinates, define the coordinate derivatives by

$$\begin{aligned} \langle \partial_m G(\theta), u\rangle_{\mathsf{H}_m} &:= \mathrm{D}G(\theta)[\mathbf{e}_m u], \\ \langle u, \partial_{mn}^2 G(\theta) v\rangle_{\mathsf{H}_m} &:= \mathrm{D}^2 G(\theta)[\mathbf{e}_m u, \mathbf{e}_n v]. \end{aligned} \tag{19}$$

Thus $\partial_m G(\theta) \in \mathsf{H}_m$ and $\partial_{mn}^2 G(\theta) : \mathsf{H}_n \to \mathsf{H}_m$.

Assumption A.1 (Structured moments and weighted smoothness)**.** *The following conditions hold, with fixed structural parameters independent of N, T, p.*

(i) ***Coordinate feature growth.*** *There are $C_\phi, c_{\mathrm{gr}} > 0$ and $\kappa_{\mathrm{gr}} \ge 0$ with $\kappa_{\mathrm{dec}} > \max\{1, \kappa_{\mathrm{gr}}\}$. Set $L_m := C_\phi m \exp(c_{\mathrm{gr}} m^{\kappa_{\mathrm{gr}}})$. Each ϕ_m is C^2, its third Fréchet derivative exists, and*

$$\begin{aligned} \|\phi_m(x)\|_{\mathsf{H}_m} &\le L_m(1 + \|x\|_{\mathbb{R}^d}^m), \\ \|\mathrm{D}^r \phi_m(x)\|_{(\mathbb{R}^d)^r \to \mathsf{H}_m} &\le L_m(1 + \|x\|_{\mathbb{R}^d}^{(m-r)_+}), \qquad r = 1, 2, 3. \end{aligned} \tag{20}$$

Here $(m - r)_+ := \max\{m - r, 0\}$, and $\mathrm{D}^0\phi_m = \phi_m$. For $r \ge 1$ the displayed norm is the multilinear operator norm. Continuity of $\mathrm{D}^3\phi_m$ is not required.

(ii) ***Weighted smoothness of the moment functional.*** *There is one deterministic open set $\mathcal{U} \subset \mathsf{H}_a$ and one $G \in C^3(\mathcal{U}; \mathbb{R})$ such that*

$$\bigcup_{K>0,\ D\geq 1} \mathcal{V}_{K,D} \subset \mathcal{U}, \qquad \sup_{\theta\in\mathcal{U}} \|\mathrm{D}^3 G(\theta)\|_{\mathsf{H}_a\times\mathsf{H}_a\times\mathsf{H}_a\to\mathbb{R}} \leq C_3 < \infty. \tag{21}$$

The set $\mathcal{U}$, the functional G, and C_3 do not vary with K, D, N, T, p. Convexity of $\mathcal{U}$ is not required: the moment segments used below lie in $\mathcal{V}_{K,D}$.

Definition A.3 (Objective induced by weighted moments)**.** *For $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ whose feature integrals exist and whose moment vector belongs to $\mathcal{U}$, define*

$$F(\nu) := G(M(\nu)). \tag{22}$$

This domain includes every $\mathcal{P}_{\alpha,K}$ with finite K and every $\mathcal{P}_D$ by lemma A.2(i) and (21)*.*

Local bounds on $\mathrm{D}G, \mathrm{D}^2 G$ and their coordinate derivatives will follow from a fixed anchor and C_3 in lemmas A.2 and A.4; no uniform bound on these lower derivatives over all of $\mathcal{U}$ is imposed.

Jump notation. We use the population and empirical jump models in (4) and (6). Let $\mathcal{H}_{\ell-}$ be the full pre-jump history and $\mathcal{G}_{\mathrm{pop}}$ the population-field sigma-field. For a law ν and a particle configuration $\mathbf{Z}$, set

$$K_\nu(x,y) := \int \nabla_W^2 F(\nu; x, z) \nabla_W^2 F(\nu; z, y)\, \nu(dz),$$
$$\mathbf{M}^N(\rho_{\mathbf{Z}}, \mathbf{Z})_{ij} := N^{-1/2} \nabla_W^2 F(\rho_{\mathbf{Z}}; Z^i, Z^j).$$

Assumption A.2 (Gaussian perturbation model)**.** *The population and empirical jumps follow the models specified above. At each prescribed jump, the population field is generated from a fresh innovation independent of the initial sample and all earlier innovations. Conditional on $\mathcal{H}_{\ell-}$, it is centered Gaussian with covariance $K_{\mu_{t_\ell-}}$. For every population law μ_t and empirical law $\rho_{\mathbf{X}_t}, \rho_{\mathbf{Y}_t}$ used in the comparison, including pre-jump laws, the Hessian kernel is symmetric under transposition of its two spatial slots, and the operator with kernel K_ν on $L^2(\nu; \mathbb{R}^d)$ is positive and trace class. A jointly measurable version of the field is used in the population pushforward.*

Finite representation and comparison coupling. Kernel symmetry gives the covariance identity

$$[K_{\rho_{\mathbf{Z}}}(Z^i, Z^j)]_{i,j} = \mathbf{M}^N(\rho_{\mathbf{Z}}, \mathbf{Z}) \mathbf{M}^N(\rho_{\mathbf{Z}}, \mathbf{Z})^\top.$$

This is the finite representation of the prescribed empirical Gaussian transition. Lemma A.14 constructs a coordinate that is standard Gaussian conditional on $\mathcal{H}_{\ell-}$ and realizes this transition together with the population evaluation vector. This is a construction, rather than an additional independence assumption on the coordinate and the reference system. Reference conditional independence is established under $\mathcal{G}_{\mathrm{pop}}$ in Lemma A.19. The covariance uses the integral Hessian term; the drift curvature uses the full Hessian.

The following tail condition ensures that the reference and empirical particles remain in a common sub-Weibull class throughout the fixed time horizon.

Assumption A.3 (Unconditional moment growth)**.** *There are deterministic $K_0 > 0$ and $\alpha \in (0, 2]$ such that*

$$\sup_{0\leq t\leq T} \sup_{q\geq 1} q^{-1/\alpha} \max\{(\mathbb{E}\|X_t^1\|_{\mathbb{R}^d}^q)^{1/q}, (\mathbb{E}\|Y_t^1\|_{\mathbb{R}^d}^q)^{1/q}\} \leq K_0.$$

Marginal and conditional bounds. The coupled construction is invariant under simultaneous particle permutations, as proved in Step 6 of lemma A.14; hence the same bound holds for every particle. Bounds at left limits follow by Fatou's lemma. The class scale K in assumption A.1 ranges over all finite values, including K_0. Conditional bounds with a deterministic enlarged scale K_{pop} hold on the population-measurable event of Lemma A.20; they are not assumed almost surely with scale K_0. In asymptotic statements a uniform bound on K_0 is specified explicitly.

We separately assume a common deterministic Lipschitz bound on each ODE branch. This pathwise condition is used for time-grid interpolation in lemma A.21 and for the statistics in section A.9; it is not a consequence of the moment assumption. The two traces at a jump belong to different branches, so the bound does not constrain the jump increment.

ODE intervals. For fixed $T > 0$, write $I_\ell = (t_\ell, t_{\ell+1})$, with $t_0 = 0$ and $t_{L_{\mathrm{jump}}+1} = T$. Empty intervals, including a terminal interval after a jump at T, impose no condition.

Assumption A.4 (Piecewise-Lipschitz flow)**.** *There is a deterministic constant $B \geq 0$ such that*

$$\sup_{1\leq i\leq N}\ \sup_{0\leq \ell\leq L_{\mathrm{jump}}}\ \sup_{\substack{s<t\\ s,t\in I_\ell}} \max\Big\{ \frac{\big\|Y_t^i - Y_s^i\big\|_{\mathbb{R}^d}}{|t-s|}, \frac{\big\|X_t^i - X_s^i\big\|_{\mathbb{R}^d}}{|t-s|}, \Big\} \leq B \qquad \text{a.s.} \tag{23}$$

The final pathwise condition rules out the degenerate case where the coupling error is carried by only a few particles.

Discrepancy and the anti-spikiness event. For

$$\delta_{t,i} \coloneqq X_t^i - Y_t^i, \qquad r_t^2 \coloneqq \frac{1}{N}\sum_{i=1}^N \|\delta_{t,i}\|_{\mathbb{R}^d}^2,$$

and a constant $A_{\mathrm{sp}} \geq 1$, define $E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \coloneqq \left\{ \sup_{0\leq t\leq T:\, r_t>0} \frac{\frac{1}{N}\sum_{i=1}^N \|\delta_{t,i}\|_{\mathbb{R}^d}^3}{r_t^3} \leq A_{\mathrm{sp}} \right\}$.

Assumption A.5 (Anti-spikiness on comparison paths)**.** *For a fixed $A_{\mathrm{sp}} \geq 1$, the pathwise comparison is restricted to $E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$. The almost-sure anti-spikiness condition in assumption 5.2 is the special case where this event has probability one; no such probability bound is asserted here.*

If $r_t = 0$, every discrepancy is zero. Thus, including these times, the event is equivalently specified by

$$\frac{1}{N}\sum_{i=1}^N \big\|X_t^i - Y_t^i\big\|_{\mathbb{R}^d}^3 \leq A_{\mathrm{sp}} r_t^3 \qquad (0 \leq t \leq T). \tag{24}$$

A.2.2 Local consequences used in the comparison

We first establish summability of the moment vectors and the integral formulas on the regions $\mathcal{V}_{K,D}$ in definition A.2. These formulas then give one common envelope for the localized comparison.

A common weighted-series estimate. Define

$$\chi(u) := [\log(e+u)]^{\kappa_{\mathrm{dec}}/(\kappa_{\mathrm{dec}}-1)} \qquad (u \geq 0). \tag{25}$$

Lemma A.1 (Weighted-series estimate)**.** *With a_m, L_m from assumption A.1(i), there is a constant C depending only on the fixed feature parameters and α such that, for $K > 0$, $D \geq 1$, and $m \geq 1$,*

$$a_m L_m\{2 + D^m + K^m m^{m/\alpha}\} \leq C\exp\{C\chi(K+D) - (c_a/4)m^{\kappa_{\mathrm{dec}}}\}. \tag{26}$$

In particular,

$$\sum_{m\geq 1} a_m L_m\{2 + D^m + K^m m^{m/\alpha}\} \leq Ce^{C\chi(K+D)}. \tag{27}$$

For each fixed $c > 0$ there is also a constant C_c such that

$$\chi(c(1+u)+1) \leq C_c\chi(u) \qquad (u \geq 0). \tag{28}$$

Proof. Write $\kappa = \kappa_{\mathrm{dec}}$ and $s = \kappa/(\kappa-1)$ in this proof. For $u, v \geq 0$ and $\varepsilon > 0$, Young's inequality with conjugate exponents κ, s gives

$$uv \leq \varepsilon u^\kappa + \frac{\kappa-1}{\kappa}(\kappa\varepsilon)^{-1/(\kappa-1)} v^s. \tag{29}$$

Indeed, substitute $a = (\kappa\varepsilon)^{1/\kappa}u$ and $b = (\kappa\varepsilon)^{-1/\kappa}v$ into $ab \le a^\kappa/\kappa + b^s/s$. Taking $u = m$, $v = \log(e + K + D)$, and $\varepsilon = c_a/4$ yields

$$m \log(e + K + D) \le \tfrac{c_a}{4} m^\kappa + C\chi(K + D).$$

The two remaining growth terms satisfy

$$c_{\mathrm{gr}} m^{\kappa_{\mathrm{gr}}} \le \tfrac{c_a}{4} m^\kappa + C, \qquad (m/\alpha) \log m + \log(4C_\phi m) \le \tfrac{c_a}{4} m^\kappa + C.$$

These inequalities hold for all $m \ge 1$ after increasing the fixed C: the ratios $m^{\kappa_{\mathrm{gr}}}/m^\kappa$ and $m \log m/m^\kappa$ tend to zero, and the finitely many remaining values have a finite maximum. Since $2 + D^m + K^m m^{m/\alpha} \le 4(e + K + D)^m m^{m/\alpha}$, substitution of these three bounds gives

$$\begin{aligned}
&\log\Big[a_m L_m \{2 + D^m + K^m m^{m/\alpha}\}\Big] \\
&\quad \le -c_a m^\kappa + c_{\mathrm{gr}} m^{\kappa_{\mathrm{gr}}} + m \log(e + K + D) + (m/\alpha) \log m + \log(4C_\phi m) \\
&\quad \le (-c_a + c_a/4 + c_a/4 + c_a/4) m^\kappa + C\chi(K + D) + C \\
&\quad = -(c_a/4) m^\kappa + C\chi(K + D) + C.
\end{aligned}$$

Exponentiating and absorbing the fixed factor e^C yields (26).

Summing leaves $\sum_{m \ge 1} e^{-(c_a/4)m^\kappa} < \infty$, proving (27). Finally,

$$\log(e + c(1 + u) + 1) \le \log(c + 2) + \log(e + u) \le (1 + \log(c + 2)) \log(e + u).$$

Raising to the power s proves (28). The constants here do not depend on the anchor norms or on K, D, N, T, p. □

Fixed anchor. Let δ_0 be the Dirac measure at the origin. Since $\delta_0 \in \mathcal{P}_D$, set

$$\begin{aligned}
\theta_0 &:= M(\delta_0) = (\phi_m(0))_{m \ge 1} \in \mathcal{V}_{K,D} \subset \mathcal{U}, \\
g_0 &:= \|\mathrm{D}G(\theta_0)\|_{\mathsf{H}_a \to \mathbb{R}}, \qquad h_0 := \|\mathrm{D}^2 G(\theta_0)\|_{\mathsf{H}_a \times \mathsf{H}_a \to \mathbb{R}}.
\end{aligned} \tag{30}$$

Moment summability in the next lemma justifies $\theta_0 \in \mathsf{H}_a$. The C^3 property in assumption A.1(ii) makes g_0, h_0 finite; they are fixed independently of K, D, N, T, p.

Lemma A.2 (Moment summability and integral formulas)**.** *Under assumption A.1, for finite $K > 0$ and $D \ge 1$ the following statements hold.*

(i) ***Moment vectors.*** *All feature integrals exist for $\nu \in \mathcal{P}_{\alpha,K} \cup \mathcal{P}_D$, and*

$$\begin{aligned}
\|M(\nu)\|_{\mathsf{H}_a} &\le \sum_{m \ge 1} a_m L_m (1 + K^m m^{m/\alpha}) < \infty \quad (\nu \in \mathcal{P}_{\alpha,K}), \\
\|M(\nu)\|_{\mathsf{H}_a} &\le \sum_{m \ge 1} a_m L_m (1 + D^m) < \infty \qquad\quad (\nu \in \mathcal{P}_D).
\end{aligned} \tag{31}$$

(ii) ***Integral formulas.*** *For $\theta, \vartheta \in \mathcal{V}_{K,D}$, put $\Delta := \vartheta - \theta$ and $\theta_s := (1 - s)\theta + s\vartheta$. Then*

$$G(\vartheta) - G(\theta) = \int_0^1 \mathrm{D}G(\theta_s)[\Delta]\, ds, \tag{32}$$

$$\mathrm{D}G(\vartheta) - \mathrm{D}G(\theta) = \int_0^1 \mathrm{D}^2 G(\theta_s)[\Delta, \cdot]\, ds, \tag{33}$$

$$\|\mathrm{D}^2 G(\vartheta) - \mathrm{D}^2 G(\theta)\|_{\mathsf{H}_a \times \mathsf{H}_a \to \mathbb{R}} \le C_3 \|\vartheta - \theta\|_{\mathsf{H}_a}. \tag{34}$$

(iii) ***Bounds from the anchor.*** *With θ_0, g_0, h_0 from* (30)*, for $\theta \in \mathcal{V}_{K,D}$,*

$$\begin{aligned}
\|\mathrm{D}^2 G(\theta)\|_{\mathsf{H}_a \times \mathsf{H}_a \to \mathbb{R}} &\le h_0 + C_3 \|\theta - \theta_0\|_{\mathsf{H}_a}, \\
\|\mathrm{D}G(\theta)\|_{\mathsf{H}_a \to \mathbb{R}} &\le g_0 + h_0 \|\theta - \theta_0\|_{\mathsf{H}_a} + \tfrac{1}{2} C_3 \|\theta - \theta_0\|_{\mathsf{H}_a}^2.
\end{aligned} \tag{35}$$

Proof. **Step 1: sum the weighted coordinate moments.** For $\nu \in \mathcal{P}_{\alpha,K}$, the $q = m$ case of (17) and the feature bound (20) give

$$\int \|\phi_m(x)\|_{\mathsf{H}_m}\,\nu(dx) \le L_m\left(1 + \int \|x\|_{\mathbb{R}^d}^m\,\nu(dx)\right) \le L_m(1 + K^m m^{m/\alpha}).$$

Thus the Bochner integrals exist and obey the same bound in norm. For their sum, apply lemma A.1, (27), with $D = 1$ and the given K. For every integer n,

$$\sum_{m\le n} a_m^2\|M_m(\nu)\|_{\mathsf{H}_m}^2 \le \left(\sum_{m\le n} a_m\|M_m(\nu)\|_{\mathsf{H}_m}\right)^2 \le \left(\sum_{m\ge 1} a_m L_m(1 + K^m m^{m/\alpha})\right)^2.$$

Letting $n \to \infty$ proves the first line of (31). For $\nu \in \mathcal{P}_D$, replace the moment bound by $\int \|x\|_{\mathbb{R}^d}^m\,\nu(dx) \le D^m$; apply (27) with $K = 1$ and the given D to get a finite weighted sum. The same finite-partial-sum inequality proves the second line of (31).

Step 2: integrate on a feasible moment segment. By definition A.2 and (21), $\theta_s \in \mathcal{V}_{K,D} \subset \mathcal{U}$ for every $s \in [0,1]$. Since $\dot\theta_s = \Delta$, the chain rule gives $\frac{d}{ds}G(\theta_s) = \mathrm{D}G(\theta_s)[\Delta]$ and $\frac{d}{ds}\mathrm{D}G(\theta_s) = \mathrm{D}^2G(\theta_s)[\Delta,\cdot]$. Integrating gives eqs. (32) and (33). For $u, v \in \mathsf{H}_a$, the same calculation and assumption (ii) give

$$(\mathrm{D}^2G(\vartheta) - \mathrm{D}^2G(\theta))[u,v] = \int_0^1 \mathrm{D}^3G(\theta_s)[\Delta,u,v]\,ds,$$
$$|(\mathrm{D}^2G(\vartheta) - \mathrm{D}^2G(\theta))[u,v]| \le C_3\|\Delta\|_{\mathsf{H}_a}\|u\|_{\mathsf{H}_a}\|v\|_{\mathsf{H}_a}.$$

The supremum over unit u, v is (34).

Step 3: integrate from the fixed anchor. Both θ_0 and θ lie in $\mathcal{V}_{K,D}$. Equation (34) first gives $\|\mathrm{D}^2G(\theta)\|_{\mathsf{H}_a\times\mathsf{H}_a\to\mathbb{R}} \le h_0 + C_3\|\theta - \theta_0\|_{\mathsf{H}_a}$. Apply this bound at $\theta_0 + s(\theta - \theta_0)$ in (33):

$$\begin{aligned}\|\mathrm{D}G(\theta)\|_{\mathsf{H}_a\to\mathbb{R}} &\le g_0 + \|\theta - \theta_0\|_{\mathsf{H}_a}\int_0^1 (h_0 + C_3 s\|\theta - \theta_0\|_{\mathsf{H}_a})\,ds \\ &= g_0 + h_0\|\theta - \theta_0\|_{\mathsf{H}_a} + \tfrac{1}{2}C_3\|\theta - \theta_0\|_{\mathsf{H}_a}^2.\end{aligned}$$

This proves (35). □

Particle localization and the conditional population scale. For $D \ge 1$, define

$$E_{T,D} := \left\{\max_{1\le i\le N}\sup_{0\le t\le T}(\|X_t^i\|_{\mathbb{R}^d} \vee \|Y_t^i\|_{\mathbb{R}^d}) \le D\right\}. \tag{36}$$

The population quantities used next are supplied by lemma A.20:

- K_{pop} is the deterministic conditional moment scale in (120). It depends on the probability budget and horizon, and need not equal the unconditional scale K_0.
- $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$ is the population-measurable event constructed there. For a budget $p_0 \in (0,1)$ it has probability at least $1 - p_0$, and (123) gives $\mu_t \in \mathcal{P}_{\alpha,K_{\mathrm{pop}}}$ on this event.

All time statements include pre-jump and post-jump values. The next lemma records how these conclusions place the comparison paths in the domain of G; the quantitative construction of K_{pop} and E_{pop} remains in lemma A.20.

Lemma A.3 (Moment points of the population, particle clouds, and interpolations)**.** *With $K = K_{\mathrm{pop}}$ in definition A.2:*

(i) ***Population.*** *On E_{pop}, for every $t \le T$,*

$$\sup_{q\ge 1} q^{-1/\alpha}\left(\int \|x\|_{\mathbb{R}^d}^q \mu_t(dx)\right)^{1/q} \le K, \qquad \mu_t \in \mathcal{P}_{\alpha,K}.$$

(ii) ***Empirical measures and particle interpolation.*** *On $E_{T,D}$, set $Z_t^{(s),i} := (1-s)Y_t^i + sX_t^i$. Then*

$$\rho_{\mathbf{X}_t}, \quad \rho_{\mathbf{Y}_t}, \quad \rho_{\mathbf{Z}_t^{(s)}} \in \mathcal{P}_D \qquad (t \le T,\ 0 \le s \le 1).$$

(iii) ***Segments in moment space.*** *On $E_{\mathrm{pop}} \cap E_{T,D}$, let ρ, ν be any two measures in*

$$\{\mu_t, \rho_{\mathbf{X}_t}, \rho_{\mathbf{Y}_t}\} \cup \{\rho_{\mathbf{Z}_t^{(s)}} : 0 \le s \le 1\}.$$

For every $u \in [0,1]$,

$$(1-u)M(\rho) + uM(\nu) \in \mathcal{V}_{K,D} \subset \mathcal{U}.$$

Thus the integral formulas and Hessian-difference bound in lemma A.2 apply on this segment.

Proof. For (i), use $\mu_t = \mathrm{Law}(Y_t^1 \mid \mathcal{G}_{\mathrm{pop}})$ and the conditional moment bound (123) of lemma A.20. For (ii), $E_{T,D}$ gives

$$\|Z_t^{(s),i}\|_{\mathbb{R}^d} \le (1-s)\|Y_t^i\|_{\mathbb{R}^d} + s\|X_t^i\|_{\mathbb{R}^d} \le D.$$

Each stated empirical measure is therefore supported in the closed ball. The same holds for left limits because that ball is closed. For (iii), each endpoint belongs to (18) by (i)–(ii), and

$$(1-u)M(\rho) + uM(\nu) = M((1-u)\rho + u\nu) \in \mathcal{V}_{K,D}.$$

The inclusion $\mathcal{V}_{K,D} \subset \mathcal{U}$ is exactly (21); no passage to a closure is needed. The particle interpolation in (ii) is handled by its support bound; it is not identified with the measure mixture in the last display. □

Definition A.4 (One local comparison envelope)**.** *For $K > 0$, $D \ge 1$, with a_m, L_m from eqs.* (16) *and* (20) *and the anchor norms g_0, h_0 from* (30)*, define*

$$\mathrm{L}_{K,D} := 4(1 + g_0 + h_0 + C_3)\left[1 + \sum_{m \ge 1} a_m L_m \{2 + D^m + K^m m^{m/\alpha}\}\right]^3. \tag{37}$$

In the localized comparison $K = K_{\mathrm{pop}}$ is deterministic, and we write $\mathrm{L}_D := \mathrm{L}_{K_{\mathrm{pop}},D}$. This abbreviation retains the dependence of K_{pop} on N, T, p; L_D is not assumed $O(1)$.

Lemma A.4 (Growth and derivative bounds for the local envelope)**.** *Assume assumption A.1. For $K > 0$, $D \ge 1$:*

(i) ***Growth in the localization scales.*** *There is a constant C depending only on the fixed feature parameters, α, g_0, h_0, C_3, and the fixed dimension, such that*

$$1 \le \mathrm{L}_{K,D} \le C \exp\{C\chi(K+D)\}. \tag{38}$$

In particular C is independent of K, D, N, T, p.

(ii) ***Moment derivatives on the feasible region.*** *For $\theta \in \mathcal{V}_{K,D}$,*

$$\|\mathrm{D}G(\theta)\|_{\mathsf{H}_a \to \mathbb{R}} \vee \|\mathrm{D}^2 G(\theta)\|_{\mathsf{H}_a \times \mathsf{H}_a \to \mathbb{R}} \le \mathrm{L}_{K,D}. \tag{39}$$

Proof. Use the following quantities only within this proof:

$$R := 1 + \sum_{m \ge 1} a_m L_m \{2 + D^m + K^m m^{m/\alpha}\}, \qquad Q := 1 + g_0 + h_0 + C_3.$$

Thus $\mathrm{L}_{K,D} = 4QR^3$.

Step 1: apply the common series estimate. By lemma A.1, (27), $R \le Ce^{C\chi(K+D)}$. Substitution into (37) gives $\mathrm{L}_{K,D} = 4QR^3 \le Ce^{C\chi(K+D)}$, with the fixed anchor factor absorbed into C. This proves (i) without changing the defined finite-sample coefficient $\mathrm{L}_{K,D}$.

Step 2: apply the anchor estimates on the specified region. By (20), $\|\phi_m(0)\|_{\mathsf{H}_m} \le L_m$. For either class in (17),

$$\begin{aligned}\|M_m(\nu) - \phi_m(0)\|_{\mathsf{H}_m} &\le \int \|\phi_m(x)\|_{\mathsf{H}_m}\,\nu(dx) + \|\phi_m(0)\|_{\mathsf{H}_m} \\ &\le L_m\left(2 + \int \|x\|_{\mathbb{R}^d}^m\,\nu(dx)\right) \\ &\le \begin{cases} L_m(2+D^m), & \nu \in \mathcal{P}_D, \\ L_m(2+K^m m^{m/\alpha}), & \nu \in \mathcal{P}_{\alpha,K}. \end{cases}\end{aligned}$$

For $\theta \in \mathcal{V}_{K,D}$, the finite-convex-hull definition (18) gives $\theta = \sum_{j=1}^J \lambda_j M(\nu_j)$, where $\lambda_j \ge 0$, $\sum_j \lambda_j = 1$, and $\nu_j \in \mathcal{P}_{\alpha,K} \cup \mathcal{P}_D$. Using the weighted $\ell^2 \le \ell^1$ estimate from lemma A.2(i), we get

$$\begin{aligned}\|\theta - \theta_0\|_{\mathsf{H}_a} &\le \sum_{j=1}^J \lambda_j \sum_{m\ge 1} a_m \|M_m(\nu_j) - \phi_m(0)\|_{\mathsf{H}_m} \\ &\le \sum_{j=1}^J \lambda_j \sum_{m\ge 1} a_m L_m\{2 + D^m + K^m m^{m/\alpha}\} \\ &= R - 1 \le R.\end{aligned}$$

Substitute this bound into (35). Since $R \ge 1$, $h_0 + C_3 R \le (h_0 + C_3)R \le QR$ and $g_0 + h_0 R + (C_3/2)R^2 \le (g_0 + h_0 + C_3)R^2 \le QR^2$. Thus

$$\begin{aligned}\sup_{\theta\in\mathcal{V}_{K,D}} \|\mathrm{D}G(\theta)\|_{\mathsf{H}_a\to\mathbb{R}} &\le QR^2, \\ \sup_{\theta\in\mathcal{V}_{K,D}} \|\mathrm{D}^2G(\theta)\|_{\mathsf{H}_a\times\mathsf{H}_a\to\mathbb{R}} &\le QR.\end{aligned} \tag{40}$$

Both are at most $4QR^3 = \mathrm{L}_{K,D}$, proving (ii). To obtain coordinate decay, recall (19) and $\|\mathbf{e}_m u\|_{\mathsf{H}_a} = a_m\|u\|_{\mathsf{H}_m}$. For unit $u \in \mathsf{H}_m$, $v \in \mathsf{H}_n$,

$$\begin{aligned}|\langle \partial_m G(\theta), u\rangle| &= |\mathrm{D}G(\theta)[\mathbf{e}_m u]| \le \mathrm{L}_{K,D} a_m, \\ |\langle u, \partial^2_{mn} G(\theta) v\rangle| &= |\mathrm{D}^2G(\theta)[\mathbf{e}_m u, \mathbf{e}_n v]| \le \mathrm{L}_{K,D} a_m a_n.\end{aligned}$$

Taking the unit suprema gives $\|\partial_m G(\theta)\|_{\mathsf{H}_m} \le \mathrm{L}_{K,D} a_m$ and $\|\partial^2_{mn} G(\theta)\|_{\mathsf{H}_n\to\mathsf{H}_m} \le \mathrm{L}_{K,D} a_m a_n$. □

Lemma A.5 (Weighted feature regularity from coordinate bounds)**.** *Under item (i) of assumption A.1, the map $J_0(x) := (\phi_m(x))_{m\ge1}$ is a C^2 map from $\mathbb{R}^d$ to H_a, with*

$$\mathrm{D}J_0(x)[h] = (\mathrm{D}\phi_m(x)[h])_{m\ge1}, \qquad \mathrm{D}^2J_0(x)[h,v] = (\mathrm{D}^2\phi_m(x)[h,v])_{m\ge1}.$$

For $D \ge 1$,

$$\sup_{\|x\|_{\mathbb{R}^d}\le D} \|\mathrm{D}^r J_0(x)\|_{(\mathbb{R}^d)^r\to\mathsf{H}_a} \le \sum_{m\ge1} a_m L_m(1+D^m) < \infty \quad (r = 0,1,2), \tag{41}$$

$$\|\mathrm{D}^r J_0(x) - \mathrm{D}^r J_0(y)\|_{(\mathbb{R}^d)^r\to\mathsf{H}_a} \le \left(\sum_{m\ge1} a_m L_m(1+D^m)\right)\|x-y\|_{\mathbb{R}^d} \quad (r = 1,2). \tag{42}$$

The difference bound holds for $\|x\|_{\mathbb{R}^d}, \|y\|_{\mathbb{R}^d} \le D$. The $r = 0$ subscript denotes the H_a norm; for $r = 1, 2$ it is the multilinear operator norm from $(\mathbb{R}^d)^r$ to H_a.

Proof. **Step 1: bound the coordinate maps and their differences.** For fixed D, summability of $\sum_m a_m L_m(1+D^m)$ is (31) for bounded support. For $r = 0, 1, 2$, (20) and $D \ge 1$ give, for unit directions,

$$\|(\mathrm{D}^r\phi_m(x)[v_1,\dots,v_r])_m\|_{\mathsf{H}_a} \le \left(\sum_m a_m^2 L_m^2 (1+D^m)^2\right)^{1/2} \le \sum_m a_m L_m(1+D^m).$$

For $r = 0$ the left-hand side means $\|J_0(x)\|_{\mathsf{H}_a}$. Thus the first- and second-derivative candidate sequences define bounded linear and bilinear maps, respectively. If $\|x\|_{\mathbb{R}^d}, \|y\|_{\mathbb{R}^d} \le D$, then $\|(1-s)y + sx\|_{\mathbb{R}^d} \le (1-s)\|y\|_{\mathbb{R}^d} + s\|x\|_{\mathbb{R}^d} \le D$. The mean-value inequality, applied to $\mathrm{D}^r\phi_m$ for $r = 1, 2$, gives

$$\begin{aligned}\|\mathrm{D}^r\phi_m(x) - \mathrm{D}^r\phi_m(y)\|_{(\mathbb{R}^d)^r \to \mathsf{H}_m} &\le \sup_{\|z\|_{\mathbb{R}^d} \le D} \|\mathrm{D}^{r+1}\phi_m(z)\|_{(\mathbb{R}^d)^{r+1} \to \mathsf{H}_m} \|x - y\|_{\mathbb{R}^d} \\ &\le L_m(1 + D^m)\|x - y\|_{\mathbb{R}^d}.\end{aligned} \tag{43}$$

The norms here are from $(\mathbb{R}^d)^r$ to H_m. To apply the scalar mean-value theorem, evaluate on unit directions and pair the output with a unit vector of H_m; then take the supremum. This uses the bound in (20), including $r = 3$, but no continuity of $\mathrm{D}^3\phi_m$.

Step 2: sum the coordinate Taylor remainders. Fix $D > \max\{1, \|x\|_{\mathbb{R}^d}\}$ and $\|h\|_{\mathbb{R}^d} < D - \|x\|_{\mathbb{R}^d}$; the segment $x + sh$ is in the ball by the convexity calculation in Step 1. Coordinate-wise integration gives

$$\phi_m(x+h) - \phi_m(x) - \mathrm{D}\phi_m(x)[h] = \int_0^1 \{\mathrm{D}\phi_m(x+sh) - \mathrm{D}\phi_m(x)\}[h]\, ds,$$

$$\{\mathrm{D}\phi_m(x+h) - \mathrm{D}\phi_m(x)\}[v] - \mathrm{D}^2\phi_m(x)[h, v] = \int_0^1 \{\mathrm{D}^2\phi_m(x+sh) - \mathrm{D}^2\phi_m(x)\}[h, v]\, ds.$$

Apply (43) with points $x + sh, x$, whose distance is $s\|h\|_{\mathbb{R}^d}$. For unit v, each right-hand side has norm at most $L_m(1 + D^m)\|h\|^2_{\mathbb{R}^d} \int_0^1 s\, ds = \frac{1}{2}L_m(1 + D^m)\|h\|^2_{\mathbb{R}^d}$. The weighted norm of a sequence is at most the weighted sum of its coordinate norms, by the finite-partial-sum inequality $(\sum_{m \le n} b_m^2)^{1/2} \le \sum_{m \le n} b_m$ for $b_m \ge 0$ and then $n \to \infty$. Consequently,

$$\begin{aligned}&\|J_0(x+h) - J_0(x) - (\mathrm{D}\phi_m(x)[h])_m\|_{\mathsf{H}_a} \\ &\qquad \le \tfrac{1}{2}\left(\sum_m a_m L_m(1 + D^m)\right)\|h\|^2_{\mathbb{R}^d}, \\ &\sup_{\|v\|_{\mathbb{R}^d} \le 1} \|(\{\mathrm{D}\phi_m(x+h) - \mathrm{D}\phi_m(x)\}[v] - \mathrm{D}^2\phi_m(x)[h, v])_m\|_{\mathsf{H}_a} \\ &\qquad \le \tfrac{1}{2}\left(\sum_m a_m L_m(1 + D^m)\right)\|h\|^2_{\mathbb{R}^d}.\end{aligned}$$

Dividing by $\|h\|_{\mathbb{R}^d}$ proves the two derivative identities. Summing (43) in the same weighted norm gives (42), so these derivatives are continuous and J_0 is C^2. □

Lemma A.6 (Local comparison bounds with common moment weights)**.** *Assume assumption A.1. Fix $K > 0$, $D \ge 1$, and define*

$$\gamma_m(K, D) := \mathrm{L}_{K,D} a_m = \mathrm{L}_{K,D} e^{-c_a m^{\kappa_{\mathrm{dec}}}}. \tag{44}$$

Let ρ, ν be in the domain of F in eq. (22), with $M(\rho), M(\nu) \in \mathcal{V}_{K,D}$, and let x, x', y, y' be in the ball of radius D. Write $\Delta_m := \|M_m(\rho) - M_m(\nu)\|_{\mathsf{H}_m}$.

(i) ***Dependence on the measure.***

$$|F(\rho) - F(\nu)| \vee \|\nabla_W F(\rho, x) - \nabla_W F(\nu, x)\|_{\mathbb{R}^d} \le \sum_m \gamma_m(K, D)\Delta_m, \tag{45}$$

$$\begin{aligned}\max\big\{&\|\nabla_x \nabla_W F(\rho, x) - \nabla_x \nabla_W F(\nu, x)\|_{\mathbb{R}^d \to \mathbb{R}^d}, \\ &\|\nabla^2_W F(\rho; x, y) - \nabla^2_W F(\nu; x, y)\|_{\mathbb{R}^d \to \mathbb{R}^d}\big\} \le \tfrac{1}{2}\sum_m \gamma_m(K, D)\Delta_m.\end{aligned} \tag{46}$$

In particular the sum *of the two Hessian differences in* (46) *is at most $\sum_m \gamma_m(K, D)\Delta_m$.*

(ii) ***Magnitude and spatial dependence.***

$$\|\nabla_W F(\nu, x)\|_{\mathbb{R}^d} \vee \|\nabla_x \nabla_W F(\nu, x)\|_{\mathbb{R}^d \to \mathbb{R}^d} \vee \|\nabla^2_W F(\nu; x, y)\|_{\mathbb{R}^d \to \mathbb{R}^d} \le \mathrm{L}_{K,D}. \tag{47}$$

The spatial differences satisfy

$$\|\nabla_W F(\nu,x)-\nabla_W F(\nu,x')\|_{\mathbb{R}^d}\le \mathrm{L}_{K,D}\|x-x'\|_{\mathbb{R}^d}, \tag{48}$$

$$\|\nabla_x\nabla_W F(\nu,x)-\nabla_x\nabla_W F(\nu,x')\|_{\mathbb{R}^d\to\mathbb{R}^d}\le \mathrm{L}_{K,D}\|x-x'\|_{\mathbb{R}^d}, \tag{49}$$

$$\|\nabla^2_W F(\nu;x,y)-\nabla^2_W F(\nu;x',y')\|_{\mathbb{R}^d\to\mathbb{R}^d}\le \mathrm{L}_{K,D}(\|x-x'\|_{\mathbb{R}^d}+\|y-y'\|_{\mathbb{R}^d}). \tag{50}$$

For the localized comparison we abbreviate $\gamma_m(D):=\gamma_m(K_{\mathrm{pop}},D)=\mathrm{L}_D a_m$.

Proof. **Step 1: recall the derivative representations and their bounds.** Within this proof set

$$Q:=1+g_0+h_0+C_3,\qquad R:=1+\sum_m a_m L_m\{2+D^m+K^m m^{m/\alpha}\},\qquad \mathrm{L}_{K,D}=4QR^3.$$

On the stated moment and spatial regions, eqs. (40) and (41) give

$$\begin{gathered}\|\mathrm{D}G\|_{\mathsf{H}_a\to\mathbb{R}}\le QR^2,\qquad \|\mathrm{D}^2G\|_{\mathsf{H}_a\times\mathsf{H}_a\to\mathbb{R}}\le QR,\\ \|\mathrm{D}^r J_0(x)\|_{(\mathbb{R}^d)^r\to\mathsf{H}_a}\le R\qquad (r=1,2).\end{gathered} \tag{51}$$

For $\theta_\nu:=M(\nu)$ and $u,v\in\mathbb{R}^d$, the moment chain rule for eq. (22) is

$$\langle\nabla_W F(\nu,x),v\rangle=\mathrm{D}G(\theta_\nu)[\mathrm{D}J_0(x)v], \tag{52}$$

$$\langle u,\nabla_x\nabla_W F(\nu,x)v\rangle=\mathrm{D}G(\theta_\nu)[\mathrm{D}^2J_0(x)[u,v]], \tag{53}$$

$$\langle u,\nabla^2_W F(\nu;x,y)v\rangle=\mathrm{D}^2G(\theta_\nu)[\mathrm{D}J_0(x)u,\mathrm{D}J_0(y)v]. \tag{54}$$

These are the coordinate chain-rule sums, with the directions collected in H_a. For example the first sum is $\sum_m\langle\partial_m G(\theta_\nu),\mathrm{D}\phi_m(x)v\rangle$. Finite coordinate sums converge to the displayed forms because the directions converge in H_a by lemma A.5, and $\mathrm{D}G,\mathrm{D}^2G$ are bounded forms on H_a by assumption A.1(ii).

Step 2: change the measure along its moment segment. Set $\theta_\rho:=M(\rho)$, $\Delta:=\theta_\rho-\theta_\nu$, and $\theta_s:=\theta_\nu+s\Delta\in\mathcal{V}_{K,D}$. By eqs. (32), (33) and (51),

$$|F(\rho)-F(\nu)|\le\int_0^1 QR^2\|\Delta\|_{\mathsf{H}_a}\,ds=QR^2\|\Delta\|_{\mathsf{H}_a},$$

$$\|\mathrm{D}G(\theta_\rho)-\mathrm{D}G(\theta_\nu)\|_{\mathsf{H}_a\to\mathbb{R}}\le\int_0^1 QR\|\Delta\|_{\mathsf{H}_a}\,ds=QR\|\Delta\|_{\mathsf{H}_a}.$$

Substituting the second inequality in eqs. (52) and (53) and taking suprema over unit directions gives

$$\|\nabla_W F(\rho,x)-\nabla_W F(\nu,x)\|_{\mathbb{R}^d}\le QR\|\Delta\|_{\mathsf{H}_a}\|\mathrm{D}J_0(x)\|_{\mathbb{R}^d\to\mathsf{H}_a}\le QR^2\|\Delta\|_{\mathsf{H}_a},$$

$$\|\nabla_x\nabla_W F(\rho,x)-\nabla_x\nabla_W F(\nu,x)\|_{\mathbb{R}^d\to\mathbb{R}^d}\le QR\|\Delta\|_{\mathsf{H}_a}\|\mathrm{D}^2J_0(x)\|_{\mathbb{R}^d\times\mathbb{R}^d\to\mathsf{H}_a}\le QR^2\|\Delta\|_{\mathsf{H}_a}.$$

For the integral Hessian, use eqs. (34) and (54):

$$\begin{aligned}&\|\nabla^2_W F(\rho;x,y)-\nabla^2_W F(\nu;x,y)\|_{\mathbb{R}^d\to\mathbb{R}^d}\\ &\quad\le C_3\|\Delta\|_{\mathsf{H}_a}\|\mathrm{D}J_0(x)\|_{\mathbb{R}^d\to\mathsf{H}_a}\|\mathrm{D}J_0(y)\|_{\mathbb{R}^d\to\mathsf{H}_a}\le QR^2\|\Delta\|_{\mathsf{H}_a}.\end{aligned}$$

Finally,

$$\|\Delta\|_{\mathsf{H}_a}\le\sum_m a_m\Delta_m,\qquad QR^2\le\tfrac12\mathrm{L}_{K,D}\quad\text{because }\mathrm{L}_{K,D}=4QR^3,\ R\ge1.$$

This proves both inequalities in (i), with room to sum the two Hessian errors.

Step 3: change the spatial variables. The three representations and (51) give

$$\|\nabla_W F(\nu,x)\|_{\mathbb{R}^d}\le QR^2R,$$

$$\|\nabla_x\nabla_W F(\nu,x)\|_{\mathbb{R}^d\to\mathbb{R}^d}\le QR^2R,$$

$$\|\nabla^2_W F(\nu;x,y)\|_{\mathbb{R}^d\to\mathbb{R}^d}\le QRR^2.$$

All three are at most $QR^3 \le \mathrm{L}_{K,D}$. Equation (42) gives $\|\mathrm{D}^r J_0(x) - \mathrm{D}^r J_0(x')\|_{(\mathbb{R}^d)^r \to \mathsf{H}_a} \le R\|x - x'\|_{\mathbb{R}^d}$ for $r = 1, 2$. Inserting these bounds in the first two representations gives eqs. (48) and (49), since $QR^2 R \le \mathrm{L}_{K,D}$. For the kernel, change one slot at a time:

$$\begin{aligned}&\mathrm{D}^2 G(\theta_\nu)[\mathrm{D}J_0(x)u, \mathrm{D}J_0(y)v] - \mathrm{D}^2 G(\theta_\nu)[\mathrm{D}J_0(x')u, \mathrm{D}J_0(y')v]\\ &= \mathrm{D}^2 G(\theta_\nu)[(\mathrm{D}J_0(x) - \mathrm{D}J_0(x'))u, \mathrm{D}J_0(y)v]\\ &\quad + \mathrm{D}^2 G(\theta_\nu)[\mathrm{D}J_0(x')u, (\mathrm{D}J_0(y) - \mathrm{D}J_0(y'))v].\end{aligned}$$

For unit u, v, its absolute value is at most $QR\{R^2\|x-x'\|_{\mathbb{R}^d} + R^2\|y-y'\|_{\mathbb{R}^d}\}$. Taking the supremum proves (50). □

In the local comparison estimates below we work on E_{pop} and use the deterministic conditional moment scale and coefficient convention from the assumptions. Spatial localization and anti-spikiness are imposed where stated.

A.3 Bound $r_t = \frac{1}{\sqrt{N}}\|\boldsymbol{X}_t - \boldsymbol{Y}_t\|_2$ between jumps

Throughout this subsection, use $\mathrm{L}_D = \mathrm{L}_{K_{\mathrm{pop}},D}$ and $\gamma_m(D) = \mathrm{L}_D a_m$ from lemma A.6. We suppress the argument D of γ_m when the localization radius is fixed.

On each ODE interval, the discrepancy between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ is read through the segment joining the two configurations.

For $s \in [0, 1]$, define the interpolating particle configuration

$$\boldsymbol{Z}_t^{(s)} \coloneqq (1-s)\boldsymbol{Y}_t + s\boldsymbol{X}_t \in \mathbb{R}^{dN}, \quad (Z_t^{(s)})^i \coloneqq (1-s)Y_t^i + sX_t^i \tag{55}$$

The linearization of the empirical vector field produces the following block Hessian matrix on $(\mathbb{R}^d)^N$.

For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\boldsymbol{v} \in (\mathbb{R}^d)^N$, define a block matrix

$$\mathrm{H}(\mu, \boldsymbol{v}) \coloneqq \left\{\frac{1}{N}\nabla_W^2 F(\mu, v_i, v_j) + \delta_{ij}\nabla_x \nabla_W F(\mu, v_i) \in \mathbb{R}^{d\times d}\right\}_{i,j=1}^{N}. \tag{56}$$

For each (t, s), define a $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ block matrix

$$\mathrm{H}_t^{(s)} \coloneqq \mathrm{H}(\rho_{\boldsymbol{Z}_t^{(s)}}, \boldsymbol{Z}_t^{(s)}). \tag{57}$$

For each $t \ge 0$, define the population-sampled Hessian matrix along the coupled infinite-particle trajectory by

$$\mathrm{H}_t^Y \coloneqq \mathrm{H}(\mu_t, \boldsymbol{Y}_t). \tag{58}$$

The smallest eigenvalue of the population-sampled block Hessian is the local curvature quantity that drives contraction or expansion of the coupling error.

Let

$$\kappa_t^{Y,N} \coloneqq \lambda_{\min}\big(\mathrm{H}_t^Y\big). \tag{59}$$

Bound the deterministic dynamics The first lemma identifies the exact Jacobian of the empirical vector field, so that the particlewise difference can be expressed through a block Hessian.

Lemma A.7 (Mean-value theorem for the empirical vector field)**.** *Let $V_N(\boldsymbol{X}) := \big(V(\rho_{\boldsymbol{X}}, X_i)\big)_{i=1}^N$, $\rho_{\boldsymbol{X}} := \frac{1}{N}\sum_{k=1}^N \delta_{X_k}$, where $V : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ is C^1 in the spatial variable and Wasserstein differentiable in the measure variable, with continuous derivatives $\nabla_x V$ and $\nabla_W V$. For $\boldsymbol{X}_t, \boldsymbol{Y}_t \in (\mathbb{R}^d)^N$, define $\Delta_t := \boldsymbol{X}_t - \boldsymbol{Y}_t$, $\boldsymbol{Z}_t^{(s)} := \boldsymbol{Y}_t + s\Delta_t$. Then*

$$V_N(\boldsymbol{X}_t) - V_N(\boldsymbol{Y}_t) = \int_0^1 DV_N(\boldsymbol{Z}_t^{(s)})\,\Delta_t \,\mathrm{d}s,$$

where $DV_N(\boldsymbol{Z}) \in \mathbb{R}^{Nd\times Nd}$ is the block matrix whose (i, j)-block is given by

$$\big[DV_N(\boldsymbol{Z})\big]_{ij} = \delta_{ij}\,\nabla_x V(\rho_{\boldsymbol{Z}}, Z_i) + \frac{1}{N}\,\nabla_W V(\rho_{\boldsymbol{Z}}, Z_i)(Z_j).$$

For the moment objective F in definition A.3, the same mean-value identity holds for $V_N(\boldsymbol{X}) = (-\nabla_W F(\rho_{\boldsymbol{X}}, X_i))_{i=1}^N$, with $\big[DV_N(\boldsymbol{Z})\big]_{ij} = -\delta_{ij}\,\nabla_x \nabla_W F(\rho_{\boldsymbol{Z}}, Z_i) - \frac{1}{N}\,\nabla_W^2 F(\rho_{\boldsymbol{Z}}, Z_i, Z_j)$.

Proof. **Step 1: integrate the configuration derivative.** Fix t and define $\Phi(s) := V_N(\boldsymbol{Z}_t^{(s)}) \in (\mathbb{R}^d)^N$. Since V_N is C^1 on $(\mathbb{R}^d)^N$, the fundamental theorem of calculus gives

$$V_N(\boldsymbol{X}_t) - V_N(\boldsymbol{Y}_t) = \Phi(1) - \Phi(0) = \int_0^1 \Phi'(s)\,\mathrm{d}s = \int_0^1 DV_N(\boldsymbol{Z}_t^{(s)})\,\Delta_t\,\mathrm{d}s.$$

Step 2: compute the spatial and measure derivatives. It remains to identify the Jacobian DV_N. Assume that $V(\mu, x)$ is C^1 in x and differentiable in μ in the Wasserstein sense, with derivative denoted by $\nabla_W V(\mu, x)(\cdot)$. Then, by the chain rule for empirical measures,

$$\partial_{X_j}\big[V(\rho_{\boldsymbol{X}}, x)\big] = \frac{1}{N}\nabla_W V(\rho_{\boldsymbol{X}}, x)(X_j).$$

Applying the usual chain rule to $V(\rho_{\boldsymbol{X}}, X_i)$ yields

$$\big[DV_N(\boldsymbol{X})\big]_{ij} = \partial_{X_j} V(\rho_{\boldsymbol{X}}, X_i) = \delta_{ij}\,\nabla_x V(\rho_{\boldsymbol{X}}, X_i) + \frac{1}{N}\,\nabla_W V(\rho_{\boldsymbol{X}}, X_i)(X_j).$$

Finally, if $V = -\nabla_W F$, then $\nabla_x V = -\nabla_x \nabla_W F,\ \nabla_W V = -\nabla_W^2 F$, which yields the stated block formula. For the moment objective in definition A.3, every empirical law lies in its domain. The identity $M(\rho_{\boldsymbol{X}}) = N^{-1}\sum_k J_0(X_k)$, lemma A.5, and (52) give a C^1 empirical field V_N. Its block derivative follows directly as below. Moving only particle j in direction h gives

$$\left.\frac{d}{d\varepsilon}\right|_0 M_m\left(\frac{1}{N}\sum_{k\neq j}\delta_{X_k} + \frac{1}{N}\delta_{X_j+\varepsilon h}\right) = \frac{1}{N}\mathrm{D}\phi_m(X_j)h.$$

In the weighted space, the corresponding derivative of $M(\rho_{\boldsymbol{X}})$ is $N^{-1}\mathrm{D}J_0(X_j)h$ by lemma A.5. Keep the evaluation point x fixed and pair with $v \in \mathbb{R}^d$ in (52):

$$\begin{aligned}\partial_{X_j}\langle \nabla_W F(\rho_{\boldsymbol{X}}, x), v\rangle_{\mathbb{R}^d}[h] &= \mathrm{D}^2 G(M(\rho_{\boldsymbol{X}}))[N^{-1}\mathrm{D}J_0(X_j)h, \mathrm{D}J_0(x)v] \\ &= N^{-1}\langle v, \nabla_W^2 F(\rho_{\boldsymbol{X}}; x, X_j)h\rangle_{\mathbb{R}^d}.\end{aligned}$$

The last equality uses symmetry of $\mathrm{D}^2 G$ and (54). Since v is arbitrary, set $x = X_i$ to obtain the measure contribution $N^{-1}\nabla_W^2 F(\rho_{\boldsymbol{X}}; X_i, X_j)h$. Moving the evaluation point contributes $\delta_{ij}\nabla_x \nabla_W F(\rho_{\boldsymbol{X}}, X_i)h$. Both contributions change sign when $V = -\nabla_W F$. □

Using this linearization, the next lemma isolates the three forces in the ODE error: sampled curvature, Hessian mismatch, and empirical drift residual.

Lemma A.8 (Upper bound of $\dot{r}_t$ under the ODE flow given error terms ι_t^N and ε_t^N.)**.** *Let $V(\mu, x) := -\nabla_W F(\mu, x)$. Consider the coupled systems*

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t^i = V(\rho_{\boldsymbol{X}_t}, X_t^i), \qquad \frac{\mathrm{d}}{\mathrm{d}t}Y_t^i = V(\mu_t, Y_t^i). \tag{60}$$

Define

$$\Delta_t := \boldsymbol{X}_t - \boldsymbol{Y}_t \in (\mathbb{R}^d)^N, \qquad r_t := \frac{1}{\sqrt{N}}\|\Delta_t\|_2, \tag{61}$$

$$\iota_t^N := \frac{1}{N}\sup_{s\in[0,1]}|\Delta_t^\top(\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y)\Delta_t|, \quad \varepsilon_t^N := \left(\frac{1}{N}\sum_{i=1}^N \big|V(\rho_{\boldsymbol{Y}_t}, Y_t^i) - V(\mu_t, Y_t^i)\big|^2\right)^{1/2}. \tag{62}$$

Assume that the map $\boldsymbol{x} \mapsto (V(\rho_{\boldsymbol{x}}, x^i))_{i=1}^N$ is differentiable along the segment joining $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. Then, at non-jump times with $r_t > 0$,

$$\dot{r}_t \le -\kappa_t^{Y,N} r_t + r_t^{-1}\iota_t^N + \varepsilon_t^N. \tag{63}$$

Proof. **Step 1: differentiate and split the squared discrepancy.**

First,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}r_t^2 = \frac{1}{N}\left\langle \Delta_t, V_N(\boldsymbol{X}_t) - V_\infty(\boldsymbol{Y}_t)\right\rangle, \tag{64}$$

where

$$V_N(\boldsymbol{X}_t) \coloneqq \big(V(\rho_{\boldsymbol{X}_t}, X_t^i)\big)_{i=1}^N, \qquad V_\infty(\boldsymbol{Y}_t) \coloneqq \big(V(\mu_t, Y_t^i)\big)_{i=1}^N. \tag{65}$$

We decompose

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}r_t^2 = \underbrace{\frac{1}{N}\left\langle \Delta_t, V_N(\boldsymbol{X}_t) - V_N(\boldsymbol{Y}_t)\right\rangle}_{\mathrm{(A)}} + \underbrace{\frac{1}{N}\left\langle \Delta_t, V_N(\boldsymbol{Y}_t) - V_\infty(\boldsymbol{Y}_t)\right\rangle}_{\mathrm{(B)}}. \tag{66}$$

Step 2: isolate sampled curvature and the Hessian error. For (A), by the mean-value theorem (lemma A.7),

$$V_N(\boldsymbol{X}_t) - V_N(\boldsymbol{Y}_t) = -\int_0^1 \mathrm{H}_t^{(s)}\Delta_t \,\mathrm{d}s. \tag{67}$$

Hence

$$\mathrm{(A)} = -\int_0^1 \frac{1}{N}\Delta_t^\top \mathrm{H}_t^{(s)}\Delta_t \,\mathrm{d}s \tag{68}$$

$$= -\int_0^1 \frac{1}{N}\Delta_t^\top \mathrm{H}_t^Y \Delta_t \,\mathrm{d}s - \int_0^1 \frac{1}{N}\Delta_t^\top \big(\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y\big)\Delta_t \,\mathrm{d}s. \tag{69}$$

By the definition of κ_t,

$$\begin{aligned} -\frac{1}{N}\Delta_t^\top \mathrm{H}_t^Y \Delta_t &= -\left\langle \frac{1}{\sqrt{N}}\Delta_t, \mathrm{H}_t^Y \frac{1}{\sqrt{N}}\Delta_t\right\rangle_2 \\ &\le -\kappa_t^{Y,N}\left\|\frac{1}{\sqrt{N}}\Delta_t\right\|_{\mathbb{R}^{dN}}^2 = -\kappa_t^{Y,N} r_t^2 \qquad (\forall s \in [0,1]), \end{aligned} \tag{70}$$

and

$$\left|\frac{1}{N}\Delta_t^\top \big(\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y\big)\Delta_t\right| \le \iota_t^N. \tag{71}$$

Therefore,

$$\mathrm{(A)} \le -\kappa_t^{Y,N} r_t^2 + \iota_t^N. \tag{72}$$

Step 3: bound the drift residual and divide for positive error. For (B), by the Cauchy–Schwarz inequality,

$$\mathrm{(B)} \le \left(\frac{1}{N}\|\Delta_t\|_2^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^N \big|V(\rho_{\boldsymbol{Y}_t}, Y_t^i) - V(\mu_t, Y_t^i)\big|^2\right)^{1/2} \tag{73}$$

$$= r_t \varepsilon_t^N. \tag{74}$$

Combining the estimates for (A) and (B), we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}r_t^2 \le -\kappa_t^{Y,N} r_t^2 + \iota_t^N + r_t\varepsilon_t^N. \tag{75}$$

Hence, whenever $r_t > 0$,

$$\dot{r}_t \le -\kappa_t^{Y,N} r_t + \iota_t^N r_t^{-1} + \varepsilon_t^N. \tag{76}$$

At $r_t = 0$ the quotient is not used; the norm derivative is treated directly in Lemma A.12. □

Bound the error terms ι_t^N and ε_t^N by moment errors. We will repeatedly convert block-operator estimates into scalar matrix estimates by bounding each block norm separately.

Lemma A.9 (operator norm bound of blocked matrices.)**.** *Let*

$$A = \begin{bmatrix} B_{1,1} & \dots & B_{1,N} \\ \vdots & \ddots & \vdots \\ B_{N,1} & \dots & B_{N,N} \end{bmatrix}, B_{i,j} \in \mathbb{R}^{d\times d} \tag{77}$$

be a $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ block matrix and

$$C = (\|B_{i,j}\|_{\mathbb{R}^d\to\mathbb{R}^d})_{i,j=1}^N \tag{78}$$

be a corresponding $\mathbb{R}^N \times \mathbb{R}^N$ matrix. then,

$$\|A\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \|C\|_{\mathbb{R}^N\to\mathbb{R}^N} \tag{79}$$

Proof. For any vector $x \in \mathbb{R}^{dN}$ such that $\|x\|_2 = 1$, we write $x = (x_1, \dots, x_N)$, $x_i \in \mathbb{R}^d$. First note that

$$1 = \|x\|_2^2 = \sum_i (\sum_j x_{i,j}^2) = \sum_i z_i^2 = \|z\|_2^2, \quad z := (\|x_1\|_2, \dots, \|x_N\|_2) \in \mathbb{R}^N. \tag{80}$$

Next, we define $[Ax]_i := ((Ax)_j)_{j=(i-1)d+1}^{id}$, $i = 1, \dots, N$. For all i, $[Ax]_i = \sum_{j=1}^N B_{ij}x_j$. The norm is bounded as $\|[Ax]_i\|_2 \le \sum_j \|B_{ij}x_j\|_{\mathbb{R}^d} \le \sum_j \|B_{i,j}\|_{\mathbb{R}^d\to\mathbb{R}^d}\|x_j\|_{\mathbb{R}^d}$. Therefore,

$$\|Ax\|_2^2 = \sum_{i=1}^N \|[Ax]_i\|_2^2 \le \sum_{i=1}^N (Cz)_i^2 = \|Cz\|_2^2 \le \|C\|_{\mathbb{R}^N\to\mathbb{R}^N}^2 \|z\|_2^2 = \|C\|_{\mathbb{R}^N\to\mathbb{R}^N}^2 \|x\|_2^2. \tag{81}$$

Taking square roots and then the supremum over $x \neq 0$ proves the claim. □

The first Hessian mismatch comes from replacing the empirical measure $\rho_{\boldsymbol{Z}_t^{(s)}}$ by the population law μ_t while keeping the particle locations fixed.

Lemma A.10 (Directional measure-slot error)**.** *Use the common weights $\gamma_m(D) = \mathrm{L}_D a_m$ from lemma A.6, and define*

$$\begin{aligned} \widetilde{\mathrm{H}}_t^{(s)} &:= \mathrm{H}(\mu_t, \boldsymbol{Z}_t^{(s)}), \\ \Delta_{m,t}^N &:= \|M_m(\rho_{\boldsymbol{Y}_t}) - M_m(\mu_t)\|_{\mathsf{H}_m}, \\ \Gamma_t^N(D) &:= \sum_{m\ge1} \gamma_m(D)\Delta_{m,t}^N, \qquad \Lambda(D) := \sum_{m\ge1} \gamma_m(D)L_m(1+2D^{m-1}). \end{aligned} \tag{82}$$

On $E_{\mathrm{pop}} \cap E_{T,D}$, for every $s \in [0,1]$,

$$\left| \frac{1}{N} \Delta_t^\top (\mathrm{H}_t^{(s)} - \widetilde{\mathrm{H}}_t^{(s)}) \Delta_t \right| \le \{\Lambda(D) r_t + \Gamma_t^N(D)\} r_t^2. \tag{83}$$

Proof. **Step 1: bound the Hessian difference by moment displacement.** Fix $s \in [0,1]$ and abbreviate $z_i := Z_t^{(s),i}$, $y_i := Y_t^i$. Let $b_s := \sum_m \gamma_m(D)\|M_m(\rho_{\boldsymbol{Z}_t^{(s)}}) - M_m(\mu_t)\|$. The two Hessian bounds in eq. (46) give

$$\begin{aligned} \|\nabla_W^2 F(\rho_{\boldsymbol{Z}_t^{(s)}}; z_i, z_j) - \nabla_W^2 F(\mu_t; z_i, z_j)\|_{\mathbb{R}^d\to\mathbb{R}^d} &\le b_s/2, \\ \|\nabla_x \nabla_W F(\rho_{\boldsymbol{Z}_t^{(s)}}, z_i) - \nabla_x \nabla_W F(\mu_t, z_i)\|_{\mathbb{R}^d\to\mathbb{R}^d} &\le b_s/2. \end{aligned}$$

The first term of the block Hessian has the additional factor $1/N$; the second is block diagonal. Applying lemma A.9 to each part therefore yields

$$\|\mathrm{H}_t^{(s)} - \widetilde{\mathrm{H}}_t^{(s)}\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \left\| \frac{b_s}{2N} \mathbf{1}\mathbf{1}^\top \right\|_{\mathbb{R}^N\to\mathbb{R}^N} + \left\| \frac{b_s}{2} I_N \right\|_{\mathbb{R}^N\to\mathbb{R}^N} = b_s.$$

Step 2: compare the empirical moments along the interpolation. For the moment displacement, integrate $\mathrm{D}\phi_m$ on the segment joining y_i and z_i, which stays in the ball of radius D. The feature bound in assumption A.1 and eq. (20) gives

$$\begin{aligned}\|\phi_m(z_i) - \phi_m(y_i)\|_{\mathsf{H}_m} &\le \int_0^1 L_m(1 + D^{(m-1)_+})\|z_i - y_i\|_{\mathbb{R}^d}\, du \\ &\le L_m(1 + 2D^{m-1})\|z_i - y_i\|_{\mathbb{R}^d}.\end{aligned}$$

Since $z_i - y_i = s(X_t^i - Y_t^i)$, Cauchy–Schwarz gives

$$\begin{aligned}\|M_m(\rho_{\boldsymbol{Z}_t^{(s)}}) - M_m(\rho_{\boldsymbol{Y}_t})\|_{\mathsf{H}_m} &\le \frac{L_m(1 + 2D^{m-1})}{N}\sum_i \|z_i - y_i\|_{\mathbb{R}^d} \\ &\le L_m(1 + 2D^{m-1})\left(\frac{1}{N}\sum_i \|z_i - y_i\|_{\mathbb{R}^d}^2\right)^{1/2} \\ &\le L_m(1 + 2D^{m-1})r_t.\end{aligned} \tag{84}$$

The triangle inequality, followed by summation with weights $\gamma_m(D)$, now gives $b_s \le \Lambda(D)r_t + \Gamma_t^N(D)$. Finally $N^{-1}\|\Delta_t\|_{\mathbb{R}^{dN}}^2 = r_t^2$, so the bound on the block operator norm proves eq. (83). □

The second Hessian mismatch comes from moving the particle locations from $\boldsymbol{Y}_t$ to the interpolation $\boldsymbol{Z}_t^{(s)}$ while keeping the population law fixed.

Lemma A.11 (Directional point-slot error under anti-spikiness)**.** *Let*

$$\delta_{t,i} := X_t^i - Y_t^i, \qquad r_t^2 := \frac{1}{N}\sum_{i=1}^N \|\delta_{t,i}\|_{\mathbb{R}^d}^2.$$

Recall from lemma A.10 the population-measure Hessian

$$\widetilde{\mathrm{H}}_t^{(s)} := \mathrm{H}(\mu_t, \boldsymbol{Z}_t^{(s)}). \tag{85}$$

On the event $E_{\mathrm{pop}} \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap E_{T,D}$, for every $s \in [0,1]$,

$$\left|\frac{1}{N}\Delta_t^\top\big(\widetilde{\mathrm{H}}_t^{(s)} - \mathrm{H}_t^Y\big)\Delta_t\right| \le \big((2 + A_{\mathrm{sp}})\mathrm{L}_D\big)r_t^3. \tag{86}$$

Proof. Fix $s \in [0,1]$, and write $z_i := Z_{t,i}^{(s)}$, $y_i := Y_t^i$, and $\delta_i := X_t^i - Y_t^i$. Since $z_i - y_i = s\delta_i$, we have $\|z_i - y_i\|_{\mathbb{R}^d} \le \|\delta_i\|_{\mathbb{R}^d}$.

Step 1: bound the integral-Hessian contribution. Split the quadratic form into the integral-Hessian contribution Q_{off}, whose sum includes $i = j$, and the spatial-Jacobian contribution Q_{diag}:

$$Q_{\mathrm{off}} := \frac{1}{N^2}\sum_{i,j=1}^N \left\langle \delta_i, \left[\nabla_W^2 F(\mu_t, z_i, z_j) - \nabla_W^2 F(\mu_t, y_i, y_j)\right]\delta_j\right\rangle, \tag{87}$$

$$Q_{\mathrm{diag}} := \frac{1}{N}\sum_{i=1}^N \left\langle \delta_i, \left[\nabla_x\nabla_W F(\mu_t, z_i) - \nabla_x\nabla_W F(\mu_t, y_i)\right]\delta_i\right\rangle. \tag{88}$$

The spatial Lipschitzness (lemma A.6) gives

$$\begin{aligned}\left\|\nabla_W^2 F(\mu_t, z_i, z_j) - \nabla_W^2 F(\mu_t, y_i, y_j)\right\|_{d\to d} &\le \mathrm{L}_D\|z_i - y_i\|_{\mathbb{R}^d} + \mathrm{L}_D\|z_j - y_j\|_{\mathbb{R}^d} \\ &\le \mathrm{L}_D\|\delta_i\|_{\mathbb{R}^d} + \mathrm{L}_D\|\delta_j\|_{\mathbb{R}^d}.\end{aligned} \tag{89}$$

Hence

$$|Q_{\mathrm{off}}| \le \frac{1}{N^2}\sum_{i,j=1}^N \|\delta_i\|_{\mathbb{R}^d}\|\delta_j\|_{\mathbb{R}^d}\big(\mathrm{L}_D\|\delta_i\|_{\mathbb{R}^d} + \mathrm{L}_D\|\delta_j\|_{\mathbb{R}^d}\big) \tag{90}$$

$$= \frac{\mathrm{L}_D}{N^2}\left(\sum_i \|\delta_i\|_{\mathbb{R}^d}^2\right)\left(\sum_j \|\delta_j\|_{\mathbb{R}^d}\right) + \frac{\mathrm{L}_D}{N^2}\left(\sum_i \|\delta_i\|_{\mathbb{R}^d}\right)\left(\sum_j \|\delta_j\|_{\mathbb{R}^d}^2\right). \tag{91}$$

By Cauchy–Schwarz,

$$\sum_i \|\delta_i\|_{\mathbb{R}^d} \le \sqrt{N}\left(\sum_i \|\delta_i\|_{\mathbb{R}^d}^2\right)^{1/2} = Nr_t, \qquad \sum_i \|\delta_i\|_{\mathbb{R}^d}^2 = Nr_t^2.$$

Therefore

$$|Q_{\mathrm{off}}| \le 2\mathrm{L}_D r_t^3. \tag{92}$$

Step 2: bound the spatial-Jacobian contribution. The spatial Lipschitzness (lemma A.6) gives

$$\|\nabla_x\nabla_W F(\mu_t, z_i) - \nabla_x\nabla_W F(\mu_t, y_i)\|_{d\to d} \le \mathrm{L}_D\|z_i - y_i\|_{\mathbb{R}^d} \le \mathrm{L}_D\|\delta_i\|_{\mathbb{R}^d}. \tag{93}$$

Consequently,

$$|Q_{\mathrm{diag}}| \le \frac{\mathrm{L}_D}{N}\sum_{i=1}^N \|\delta_i\|_{\mathbb{R}^d}^3 \le \mathrm{L}_D A_{\mathrm{sp}} r_t^3 \tag{94}$$

on $E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$. Combining the two estimates proves eq. (86). □

Combining the measure-slot and point-slot estimates gives the directional control of the Hessian mismatch that appears in the differential inequality for r_t.

Corollary A.1 (Combined directional bound for ι_t^N)**.** *On $E_{\mathrm{pop}} \cap E_{T,D} \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$, with the population moment event of lemma A.20, sample localization* (36)*, and anti-spikiness* (24)*, it holds that*

$$\iota_t^N \le \Big(\Lambda(D) + (2 + A_{\mathrm{sp}})\mathrm{L}_D\Big) r_t^3 + \Gamma_t^N r_t^2. \tag{95}$$

The same deterministic radius D is used in the event and all coefficients. Define $q_{4,t} := (N^{-1}\sum_i \|X_t^i - Y_t^i\|_{\mathbb{R}^d}^4)^{1/4}$. In particular, if $q_{4,t} \le A_4 r_t$ for all t, with $A_4 \ge 1$, then the same bound holds with $A_{\mathrm{sp}} = A_4^2$.

Proof. The directional Hessian discrepancy, recalled from lemma A.8, (62), is $\iota_t^N = N^{-1}\sup_{s\in[0,1]}|\Delta_t^\top(\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y)\Delta_t|$, where $\Delta_t = \mathbf{X}_t - \mathbf{Y}_t$. **Step 1: combine the two Hessian contributions.** For each $s \in [0,1]$, decompose

$$\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y = \big(\mathrm{H}_t^{(s)} - \widetilde{\mathrm{H}}_t^{(s)}\big) + \big(\widetilde{\mathrm{H}}_t^{(s)} - \mathrm{H}_t^Y\big). \tag{96}$$

Taking the associated quadratic form against Δ_t, dividing by N, and applying the triangle inequality gives

$$\left|\frac{1}{N}\Delta_t^\top\big(\mathrm{H}_t^{(s)} - \mathrm{H}_t^Y\big)\Delta_t\right| \le \left|\frac{1}{N}\Delta_t^\top\big(\mathrm{H}_t^{(s)} - \widetilde{\mathrm{H}}_t^{(s)}\big)\Delta_t\right| \tag{97}$$

$$+ \left|\frac{1}{N}\Delta_t^\top\big(\widetilde{\mathrm{H}}_t^{(s)} - \mathrm{H}_t^Y\big)\Delta_t\right|. \tag{98}$$

The first term is bounded by lemma A.10 as $(\Lambda(D)r_t + \Gamma_t^N)r_t^2$. The second term is bounded by lemma A.11 as $((2 + A_{\mathrm{sp}})\mathrm{L}_D)r_t^3$. Taking the supremum over $s \in [0,1]$ proves eq. (95).

Step 2: deduce anti-spikiness from the fourth moment. For $\delta_{t,i} = X_t^i - Y_t^i$, Cauchy–Schwarz gives

$$\frac{1}{N}\sum_i \|\delta_{t,i}\|_{\mathbb{R}^d}^3 \le \left(\frac{1}{N}\sum_i \|\delta_{t,i}\|_{\mathbb{R}^d}^2\right)^{1/2}\left(\frac{1}{N}\sum_i \|\delta_{t,i}\|_{\mathbb{R}^d}^4\right)^{1/2} = r_t q_{4,t}^2 \le A_4^2 r_t^3.$$

The calculation also holds for $r_t = 0$, when every $\delta_{t,i} = 0$. Thus (24) holds with $A_{\mathrm{sp}} = A_4^2$. □

Bound r_t by moment residual. The final ODE estimate packages all continuous-time residuals into the common moment fluctuation Γ_t^N, yielding the stopped curvature-response inequality used in the main theorem.

Lemma A.12 (Common empirical moment residual and directional differential inequality for r_t)**.** *Work on the population moment event of Lemma A.20, and on the events $E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$ and $E_{T,D}$, and fix an open ODE interval $I_\ell = (t_\ell, t_{\ell+1})$. Recall from eq.* (82) *the coordinate residuals and their weighted sum:*

$$\Delta_{m,t}^N := \|M_m(\rho_{\boldsymbol{Y}_t}) - M_m(\mu_t)\|_{\mathsf{H}_m}, \qquad \Gamma_t^N(D) := \sum_{m\geq 1} \gamma_m(D)\Delta_{m,t}^N. \tag{99}$$

Here, as in eq. (44)*,*

$$\gamma_m(D) := \mathrm{L}_D a_m. \tag{100}$$

Let

$$c_D := \Lambda(D) + (2 + A_{\mathrm{sp}})\mathrm{L}_D. \tag{101}$$

Then

$$\varepsilon_t^N \leq \Gamma_t^N, \tag{102}$$

and, for a.e. $t \in I_\ell$ with $r_t > 0$,

$$\dot{r}_t \leq -\kappa_t^{Y,N} r_t + c_D r_t^2 + \Gamma_t^N(D) r_t + \Gamma_t^N(D). \tag{103}$$

The right upper Dini derivative is

$$\mathrm{D}^+ r_t := \limsup_{h\downarrow 0} \frac{r_{t+h} - r_t}{h}, \qquad t, t+h \in I_\ell. \tag{104}$$

It agrees with $\dot{r}_t$ where the ordinary derivative exists. The same upper bound holds with $\mathrm{D}^+ r_t$, including at $r_t = 0$ as justified below. This definition is used within one ODE branch, not across a jump.

Moreover, for the pathwise first-exit time

$$\tau_R^\ell := \inf\{t \in I_\ell : r_t \geq R\}, \tag{105}$$

we have, for $t < \tau_R^\ell$,

$$\dot{r}_t \leq -\big(\kappa_t^{Y,N} - c_D R - \Gamma_t^N(D)\big) r_t + \Gamma_t^N(D). \tag{106}$$

Proof. **Step 1: control the residual and positive discrepancy.** For each i, the gradient bound in eq. (45) gives

$$\begin{aligned} \big\|V(\rho_{\boldsymbol{Y}_t}, Y_t^i) - V(\mu_t, Y_t^i)\big\|_{\mathbb{R}^d} &= \big\|\nabla_W F(\rho_{\boldsymbol{Y}_t}, Y_t^i) - \nabla_W F(\mu_t, Y_t^i)\big\|_{\mathbb{R}^d} \\ &\leq \sum_{m\geq 1} \gamma_m(D)\Delta_{m,t}^N = \Gamma_t^N. \end{aligned} \tag{107}$$

Therefore

$$\varepsilon_t^N = \left(\frac{1}{N}\sum_{i=1}^N \big\|V(\rho_{\boldsymbol{Y}_t}, Y_t^i) - V(\mu_t, Y_t^i)\big\|_{\mathbb{R}^d}^2\right)^{1/2} \leq \Gamma_t^N. \tag{108}$$

Next, by corollary A.1,

$$\iota_t^N \leq c_D r_t^3 + \Gamma_t^N r_t^2. \tag{109}$$

Before applying lemma A.8, we verify its differentiability premise. On $E_{\mathrm{pop}} \cap E_{T,D}$, convexity of the radius-D ball gives $Z_t^{(s),i} \in \overline{B}_D(0)$ for every $s \in [0,1]$, and lemma A.3 gives $M(\rho_{\boldsymbol{Z}_t^{(s)}}) \in \mathcal{V}_{K_{\mathrm{pop}},D} \subset \mathcal{U}$. The C^2 feature map of lemma A.5, the condition $G \in C^3(\mathcal{U})$ in assumption A.1, and the derivative formulas eqs. (52) to (54) therefore make the empirical configuration field C^1 along this segment. Applying lemma A.8, for $r_t > 0$,

$$\begin{aligned} \dot{r}_t &\leq -\kappa_t^{Y,N} r_t + \frac{\iota_t^N}{r_t} + \varepsilon_t^N && (110)\\ &\leq -\kappa_t^{Y,N} r_t + c_D r_t^2 + \Gamma_t^N r_t + \varepsilon_t^N && (111)\\ &\leq -\kappa_t^{Y,N} r_t + c_D r_t^2 + \Gamma_t^N r_t + \Gamma_t^N. && (112) \end{aligned}$$

Step 2: handle zero discrepancy without division. It remains to justify the inequality at $r_t = 0$, where division by r_t was not used. First take s, t on the same ODE branch. The pathwise bound $\|Y_t^1 - Y_s^1\|_{\mathbb{R}^d} \le B|t-s|$ and the first-derivative bound in assumption A.1 give

$$\begin{aligned}\|M_m(\mu_t) - M_m(\mu_s)\|_{\mathsf{H}_m} &\le \mathbb{E}_{\mathcal{G}}\|\phi_m(Y_t^1) - \phi_m(Y_s^1)\|_{\mathsf{H}_m} \\ &\le BL_m|t-s|\,\mathbb{E}_{\mathcal{G}}(1 + \|Y_t^1\|_{\mathbb{R}^d}^{m-1} + \|Y_s^1\|_{\mathbb{R}^d}^{m-1}) \\ &\le BL_m|t-s|[1 + 2\{K_{\mathrm{pop}}m^{1/\alpha}\}^{m-1}].\end{aligned}$$

Here $\mathbb{E}_{\mathcal{G}}$ means expectation conditional on the population fields; the conditional moment estimate is (123). Summing with weights a_m bounds the H_a norm, since a nonnegative sequence has ℓ^2 norm at most its ℓ^1 norm:

$$\|M(\mu_t) - M(\mu_s)\|_{\mathsf{H}_a} \le B|t-s| \sum_{m\ge 1} a_m L_m[1 + 2\{K_{\mathrm{pop}}m^{1/\alpha}\}^{m-1}] < \infty.$$

Apply lemma A.1, (27), with $K = 1 + K_{\mathrm{pop}}$ and $D = 1$: $\{K_{\mathrm{pop}}m^{1/\alpha}\}^{m-1} \le (1+K_{\mathrm{pop}})^m m^{m/\alpha}$, so the series is finite. For either particle system $S = X$ or $S = Y$, write $\mathbf{S}_t = (S_t^1, \dots, S_t^N)$. The event $E_{T,D}$, the branchwise B-Lipschitz bound, and convexity of the radius-D ball give

$$\begin{aligned}\|M_m(\rho_{\mathbf{S}_t}) - M_m(\rho_{\mathbf{S}_s})\|_{\mathsf{H}_m} &\le \frac{1}{N}\sum_{i=1}^N \|\phi_m(S_t^i) - \phi_m(S_s^i)\|_{\mathsf{H}_m} \\ &\le BL_m(1 + D^{(m-1)_+})|t-s|.\end{aligned}$$

Consequently, using $D \ge 1$ and the bounded-support summability in eq. (31),

$$\|M(\rho_{\mathbf{S}_t}) - M(\rho_{\mathbf{S}_s})\|_{\mathsf{H}_a} \le B|t-s| \sum_{m\ge 1} a_m L_m(1 + D^m) < \infty.$$

Thus the population and both empirical moment paths are continuous. Together with continuity of $\mathrm{D}G, \mathrm{D}^2G$ on $\mathcal{U}$ from assumption A.1(ii), the particle drifts in (52) are continuous on the branch. Recall $\Delta_t = \mathbf{X}_t - \mathbf{Y}_t$ and $V(\nu, x) = -\nabla_W F(\nu, x)$. If $r_t = 0$, then $\Delta_t = 0$ and

$$\dot\Delta_t = (V(\rho_{\mathbf{Y}_t}, Y_t^i) - V(\mu_t, Y_t^i))_{i=1}^N, \qquad \Delta_{t+h} = h\dot\Delta_t + o(h).$$

By the Dini definition (104),

$$\mathrm{D}^+ r_t = \lim_{h\downarrow 0} \frac{\|\Delta_{t+h}\|_2}{\sqrt{N}h} = \frac{\|\dot\Delta_t\|_2}{\sqrt{N}} = \varepsilon_t^N \le \Gamma_t^N.$$

For positive r_t, the preceding calculation gives the same upper Dini inequality (or its a.e. version under absolute continuity). This proves eq. (103). If $t < \tau_R^\ell$, then $r_t < R$, so

$$c_D r_t^2 \le c_D R r_t, \tag{113}$$

which yields eq. (106). □

A.4 Bound of r_t including jump errors

The jump analysis has one purpose: compare two Gaussian perturbation vectors using the same seed, so that the error is measured by covariance mismatch.

At a jump time t_ℓ, the reference particles are defined by evaluating the population Gaussian field:

$$Y_{t_\ell}^i = Y_{t_\ell -}^i + \eta_\ell \xi_{\mu_{t_\ell -}}(Y_{t_\ell -}^i), \qquad \xi_{\mu_{t_\ell -}} \sim GP(0, K_{\mu_{t_\ell -}}).$$

Thus, conditional on the population Gaussian fields and the jump schedule, the reference particles are deterministic pushforwards of their i.i.d. initial particles and remain conditionally i.i.d. with common law μ_t.

We now pass from the Gaussian field to its finite evaluation vector at the reference particles.

For the jump comparison only, we use the following finite-dimensional representation. Conditional on the full pre-jump history $\mathcal{H}_{\ell-}$, the vector

$$\big(\xi_{\mu_{t_\ell-}}(Y^1_{t_\ell-}),\dots,\xi_{\mu_{t_\ell-}}(Y^N_{t_\ell-})\big)$$

is a centered Gaussian vector with covariance $C_{\mu_{t_\ell-}}(Y_{t_\ell-})$. Hence it can be represented as $B_{\mu_{t_\ell-}}(Y_{t_\ell-})g^N_\ell$, where $B_{\mu_{t_\ell-}}(Y_{t_\ell-})B_{\mu_{t_\ell-}}(Y_{t_\ell-})^\top = C_{\mu_{t_\ell-}}(Y_{t_\ell-})$. This representation is used only inside the covariance-mismatch estimate.

Covariance-factor couplings have a Procrustes/Wasserstein precedent in Masarotto et al. (2019). Here we choose a factor that preserves the population covariance and admits a direct square-root bound. We do not need to solve the Procrustes minimization problem.

Lemma A.13 (Aligned construction of the finite square root)**.** *Let μ be a law in the Gaussian perturbation model assumption A.2, and let $\mathbf{Y} = (Y^1,\dots,Y^N) \in (\mathbb{R}^d)^N$ have finite covariance evaluations as in that model. Write $M := \mathbf{M}^N(\rho^N_{\mathbf{Y}},\mathbf{Y}) \in \mathbb{R}^{Nd\times Nd}$, $M_{ij} := \frac{1}{\sqrt{N}}\nabla^2_W F(\rho^N_{\mathbf{Y}};Y^i,Y^j)$, as well as $C := \big[K_\mu(Y^i,Y^j)\big]_{i,j=1}^N \in \mathbb{R}^{Nd\times Nd}$. Then M and C are symmetric, C is positive semidefinite, and there exists an orthogonal matrix $Q \in \mathbb{R}^{Nd\times Nd}$ such that*

$$M = |M|Q, \qquad |M| := (M^2)^{1/2}.$$

If we define

$$B := C^{1/2}Q,$$

then

$$BB^\top = C,$$

and

$$\|B-M\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} = \|C^{1/2}-|M|\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.$$

In particular,

$$\|B-M\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \|C-M^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}},$$

where $\|\cdot\|_*$ is the nuclear norm (sum of singular values).

Proof. **Step 1: check symmetry and positivity.** Since each block of M is given by

$$M_{ij} = \frac{1}{\sqrt{N}}\nabla^2_W F(\rho^N_{\mathbf{Y}};Y^i,Y^j),$$

the assumed symmetry of $\nabla^2_W F$ yields symmetry of M. Similarly, the kernel matrix C is also symmetric.

Moreover, for any $v = (v_1,\dots,v_N) \in \mathbb{R}^{Nd}$,

$$v^\top Cv = \sum_{i,j=1}^N v_i^\top K_\mu(Y^i,Y^j)v_j.$$

By the definition of K_μ,

$$v^\top Cv = \int \sum_{i,j=1}^N v_i^\top \nabla^2_W F(\mu;Y^i,z)\nabla^2_W F(\mu;z,Y^j)v_j\,\mu(dz).$$

Since $\nabla^2_W F(\mu;z,Y^j) = \nabla^2_W F(\mu;Y^j,z)^\top$, this becomes

$$v^\top Cv = \int \Big\|\sum_{i=1}^N \nabla^2_W F(\mu;Y^i,z)^\top v_i\Big\|^2_{\mathbb{R}^d}\mu(dz) \ge 0.$$

Hence C is positive semidefinite.

Step 2: construct the aligned factor. For the spectral decomposition $M = U\,\mathrm{diag}(\lambda_j)U^\top$, define

$$Q := U\,\mathrm{diag}(\mathrm{sign}_+(\lambda_j))U^\top, \qquad \mathrm{sign}_+(a) := \begin{cases} -1, & a<0,\\ 1, & a\ge 0.\end{cases}$$

Then $QQ^\top = U\,\mathrm{diag}(1)U^\top = I$, including on the zero eigenspace, and

$$|M|Q = U\,\mathrm{diag}(|\lambda_j|\,\mathrm{sign}_+(\lambda_j))U^\top = M.$$

Define

$$B := C^{1/2}Q.$$

Then

$$BB^\top = C^{1/2}QQ^\top C^{1/2} = C.$$

On the other hand,

$$B - M = C^{1/2}Q - |M|Q = (C^{1/2} - |M|)Q.$$

Right orthogonal invariance follows from $\|(C^{1/2} - |M|)Q\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} = \mathrm{tr}((C^{1/2} - |M|)QQ^\top(C^{1/2} - |M|)^\top) = \|C^{1/2} - |M|\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}$. Therefore

$$\|B - M\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} = \|(C^{1/2} - |M|)Q\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} = \|C^{1/2} - |M|\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.$$

Step 3: bound the factor difference by the covariance difference. The estimate proved in this step is the finite-dimensional Powers–Størmer inequality (Powers & Størmer, 1970); we include the calculation to keep the norms and constants explicit. For completeness, set $S = C^{1/2} - |M|$ in the following calculation, and write $S = \sum_j s_j e_j e_j^\top$ in an orthonormal eigenbasis and define $\mathrm{sign}(S) = \sum_j \mathrm{sign}(s_j) e_j e_j^\top$, where $\mathrm{sign}(a)$ is 1, 0, or -1 according as $a > 0$, $a = 0$, or $a < 0$. This differs from $\mathrm{sign}_+(0) = 1$ used to make Q orthogonal. The product identity follows by cancellation:

$$C^{1/2}S + S|M| = C - C^{1/2}|M| + C^{1/2}|M| - |M|^2 = C - M^2.$$

Thus

$$C - M^2 = C^{1/2}S + S|M|, \qquad \|\,\mathrm{sign}(S)\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le 1, \qquad S\,\mathrm{sign}(S) = \mathrm{sign}(S)S = |S|.$$

For any symmetric matrix $A = \sum_j \lambda_j e_j e_j^\top$ and any matrix J with $\|J\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le 1$, $\mathrm{tr}(JA) = \sum_j \lambda_j \langle e_j, Je_j\rangle \le \sum_j |\lambda_j| = \|A\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}$. Apply this fact to $A = C - M^2$ and $J = \mathrm{sign}(S)$, then use cyclicity of the trace:

$$\begin{aligned}\|C - M^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} &\ge \mathrm{tr}(\mathrm{sign}(S)(C^{1/2}S + S|M|)) \\ &= \mathrm{tr}(S\,\mathrm{sign}(S)C^{1/2}) + \mathrm{tr}(\mathrm{sign}(S)S|M|) \\ &= \mathrm{tr}(|S|(C^{1/2} + |M|)).\end{aligned}$$

For each e_j, both $\langle e_j, C^{1/2}e_j\rangle$ and $\langle e_j, |M|e_j\rangle$ are nonnegative. The scalar inequality $a + b \ge |a - b|$ for $a, b \ge 0$ therefore gives

$$\langle e_j, (C^{1/2} + |M|)e_j\rangle \ge |\langle e_j, (C^{1/2} - |M|)e_j\rangle| = |s_j|.$$

Consequently,

$$\begin{aligned}\|C - M^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} &\ge \sum_j |s_j|\langle e_j, (C^{1/2} + |M|)e_j\rangle \ge \sum_j |s_j|^2 \\ &= \|C^{1/2} - |M|\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} = \|B - M\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.\end{aligned}$$

No nonsingularity or optimality of the chosen factor is used. □

Lemma A.14 (Population-first finite coupling)**.** *Fix a prescribed jump t_ℓ in definition A.1 and assumption A.2. Recall that $\mathcal{H}_{\ell-}$ is the pre-jump history: it includes the initial sample, previous population innovations, previous auxiliary normals, and the pre-jump states, but not the current or future innovations. Generate the prescribed population field and define*

$$v := \big(\xi_{\mu_{t_\ell-}}(Y^i_{t_\ell-})\big)_{i=1}^N, \qquad C := \big[K_{\mu_{t_\ell-}}(Y^i_{t_\ell-}, Y^j_{t_\ell-})\big]_{i,j=1}^N.$$

Thus $v \mid \mathcal{H}_{\ell-} \sim N(0, C)$. Recall the Hessian factors:

$$(M_Z)_{ij} := \frac{1}{\sqrt{N}}\nabla^2_W F(\rho_{\mathbf{Z}_{t_\ell-}}; Z^i_{t_\ell-}, Z^j_{t_\ell-}), \quad Z \in \{X, Y\},$$
$$Q := \mathrm{sign}_+(M_Y), \qquad B := C^{1/2}Q.$$

This local factor B is denoted by B_{pop} in the main text. On an extension of the probability space, take $h_\ell \sim N(0, I_{Nd})$ independent of the original population system, the initial sample, and all other auxiliary normals. Define

$$\mathbf{g}_\ell^N = B^\dagger v + (I - B^\dagger B)h_\ell,$$

where $B^\dagger$ is the Moore–Penrose inverse, defined below. Then:

1. *The finite coordinate is conditionally standard Gaussian and reproduces the original population evaluation:*

$$\mathbf{g}_\ell^N \mid \mathcal{H}_{\ell-} \sim N(0, I_{Nd}), \qquad B\mathbf{g}_\ell^N = v \quad \textit{a.s.}$$

2. *The factor error satisfies*

$$\|B - M_Y\|^2_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \|C - M_Y^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.$$

3. *The perturbation $M_X\mathbf{g}_\ell^N$ has the prescribed empirical conditional covariance $[K_{\rho_{\mathbf{x}_{t_\ell-}}}(X^i_{t_\ell-}, X^j_{t_\ell-})]_{i,j}$. Repeating the construction with fresh h_ℓ couples all prescribed jumps while retaining the original population path and the conditional product law for the reference particles.*

Proof. **Step 1: the pre-jump matrices and the inverse.** The evaluation points and $\mu_{t_\ell-}$ are fixed after conditioning on $\mathcal{H}_{\ell-}$. The fresh field therefore has the asserted finite Gaussian law by assumption A.2; joint measurability is part of that model. All matrices C, M_X, M_Y, Q, B are measurable before the jump. For Q no measurable choice of eigenvectors is needed: the sign in lemma A.13 is the Borel map

$$Q = I - 2\lim_{k\to\infty}(|M_Y| - M_Y)(|M_Y| - M_Y + k^{-1}I)^{-1}.$$

On an eigenvalue λ, the ratio tends to 1 if $\lambda < 0$ and to 0 if $\lambda \ge 0$. The positive square root is continuous.

For a singular-value decomposition $B = U\,\mathrm{diag}(s_j)V^\top$, with U, V orthogonal and $s_j \ge 0$, define its Moore–Penrose inverse by replacing each positive s_j with s_j^{-1} and each zero s_j with zero:

$$B^\dagger = V\,\mathrm{diag}(s_j^{-1}\mathbf{1}_{\{s_j>0\}})U^\top, \qquad P_B := B^\dagger B = V\,\mathrm{diag}(\mathbf{1}_{\{s_j>0\}})V^\top.$$

The displayed diagonal convention avoids evaluating 0^{-1}. It gives $P_B^\top = P_B$ and

$$P_B^2 = V\,\mathrm{diag}(\mathbf{1}^2_{\{s_j>0\}})V^\top = P_B, \qquad BB^\dagger = U\,\mathrm{diag}(\mathbf{1}_{\{s_j>0\}})U^\top,$$
$$B(I - P_B) = U\,\mathrm{diag}(s_j\mathbf{1}_{\{s_j=0\}})V^\top = 0, \qquad (I - BB^\dagger)B = 0.$$

The last equality has the same diagonal entries $s_j\mathbf{1}_{\{s_j=0\}}$. The inverse is Borel measurable even when the rank changes, because

$$B^\dagger = \lim_{k\to\infty}(B^\top B + k^{-1}I)^{-1}B^\top;$$

the singular-value coefficient on the right is $s_j/(s_j^2 + k^{-1})$.

Step 2: the conditional standard Gaussian law. The preceding lemma gives $BB^\top = C$. Conditional on $\mathcal{H}_{\ell-}$, the terms $B^\dagger v$ and $(I - P_B)h_\ell$ are independent centered Gaussians, since h_ℓ is independent of $\sigma(\mathcal{H}_{\ell-}, v)$. Since both conditional means are zero, their covariance matrices are

$$\begin{aligned}\mathbb{E}[B^\dagger v v^\top (B^\dagger)^\top \mid \mathcal{H}_{\ell-}] &= B^\dagger C (B^\dagger)^\top = (B^\dagger B)(B^\dagger B)^\top = P_B,\\ \mathbb{E}[(I - P_B)h_\ell h_\ell^\top (I - P_B)^\top \mid \mathcal{H}_{\ell-}] &= (I - P_B)I(I - P_B)^\top = I - 2P_B + P_B^2 = I - P_B.\end{aligned} \tag{114}$$

Conditional independence makes the cross covariance zero, so the sum has covariance $P_B + (I - P_B) = I$. For every $\theta \in \mathbb{R}^{Nd}$, their characteristic functions give

$$\mathbb{E}[e^{i\theta^\top \mathbf{g}_\ell^N} \mid \mathcal{H}_{\ell-}] = e^{-\theta^\top P_B\theta/2}e^{-\theta^\top(I-P_B)\theta/2} = e^{-\|\theta\|^2_{\mathbb{R}^{dN}}/2}.$$

This proves the conditional $N(0, I_{Nd})$ law, including singular C.

Step 3: retain the original population update. Since $(I - BB^\dagger)B = 0$ and $C = BB^\top$,

$$\begin{aligned}\mathbb{E}[\|(I - BB^\dagger)v\|^2_{\mathbb{R}^{dN}} \mid \mathcal{H}_{\ell-}] &= \operatorname{tr}((I - BB^\dagger)BB^\top(I - BB^\dagger)^\top) \\ &= \operatorname{tr}([(I - BB^\dagger)B][(I - BB^\dagger)B]^\top) = 0.\end{aligned}$$

The squared norm is nonnegative, so $BB^\dagger v = v$ almost surely and

$$B\mathbf{g}^N_\ell = BB^\dagger v + B(I - P_B)h_\ell = v.$$

If $C = 0$, then $v = B = B^\dagger = 0$ and $\mathbf{g}^N_\ell = h_\ell$, which verifies this case directly. The factor error is the bound in lemma A.13 with $M = M_Y$.

Step 4: the empirical transition. Since M_X is pre-jump measurable, $M_X\mathbf{g}^N_\ell \mid \mathcal{H}_{\ell-} \sim N(0, M_X M_X^\top)$. For this calculation write $X^i = X^i_{t_\ell-}$ and $\rho_X = N^{-1}\sum_k \delta_{X^k}$. Kernel symmetry and the definition of K_ν in assumption A.2 give

$$\begin{aligned}(M_X M_X^\top)_{ij} &= \frac{1}{N}\sum_{k=1}^N \nabla^2_W F(\rho_X; X^i, X^k)\nabla^2_W F(\rho_X; X^j, X^k)^\top \\ &= \frac{1}{N}\sum_{k=1}^N \nabla^2_W F(\rho_X; X^i, X^k)\nabla^2_W F(\rho_X; X^k, X^j) \\ &= \int \nabla^2_W F(\rho_X; X^i, z)\nabla^2_W F(\rho_X; z, X^j)\rho_X(dz) = K_{\rho_X}(X^i, X^j).\end{aligned}$$

This is exactly the required empirical conditional law.

Step 5: iterate without changing the reference law. Keep the original population innovations and iid initial sample $U_1, \dots, U_N \sim \mu_0$, and add independent $h_1, \dots, h_{L_{\text{jump}}}$. Suppose the two pre-jump configurations at t_ℓ have been constructed from the initial sample and the first $\ell - 1$ innovations and auxiliary normals. Steps 1–4 give the next coordinate, update X by $\eta_\ell M_X \mathbf{g}^N_\ell$, and update Y by its original field value $\eta_\ell v$. The ODEs then give configurations measurable with respect to the next pre-jump history. This is a finite induction over the prescribed schedule.

The conditional standard Gaussian law is constant in the history. For $A \in \mathcal{H}_{\ell-}$ and a Borel set $D \subset \mathbb{R}^{Nd}$,

$$\mathbb{P}(A \cap \{\mathbf{g}^N_\ell \in D\}) = \mathbb{E}[\mathbf{1}_A \mathbb{P}(\mathbf{g}^N_\ell \in D \mid \mathcal{H}_{\ell-})] = \mathbb{P}(A)\, N(0, I_{Nd})(D).$$

Thus each coordinate is independent of its pre-jump history, which includes previous coordinates. Independence of the current field or post-jump reference particles is not asserted. The population system is unaltered: conditional on its fields, the ODE and jump maps compose into a common map $\mathcal{T}_t$ with

$$Y^i_t = \mathcal{T}_t(U_i), \qquad \mu_t = (\mathcal{T}_t)_\# \mu_0.$$

The initial sample is independent of all population innovations, so $\text{Law}((U_1, \dots, U_N) \mid \mathcal{G}_{\text{pop}}) = \mu_0^{\otimes N}$. For bounded measurable $f_1, \dots, f_N$, integration against this product law and the pushforward identity give

$$\begin{aligned}\mathbb{E}\left[\prod_i f_i(Y^i_t) \,\middle|\, \mathcal{G}_{\text{pop}}\right] &= \int \prod_i f_i(\mathcal{T}_t(u_i)) \prod_i \mu_0(du_i) \\ &= \prod_i \int f_i(\mathcal{T}_t(u_i))\mu_0(du_i) = \prod_i \int f_i(y)\mu_t(dy).\end{aligned} \tag{115}$$

This is the product law used in lemma A.19; $\mathbf{g}^N_\ell$ is not included in its conditioning sigma-field.

Step 6: preserve exchangeability of particle pairs. Let P be the orthogonal matrix of any permutation of the N particle blocks. Permute the initial sample and every auxiliary normal by P, so $h'_\ell = Ph_\ell \sim N(0, PIP^\top) = N(0, I)$. These transformed inputs have the same joint law, with the population innovations held fixed. Under simultaneous permutation of $\mathbf{X}, \mathbf{Y}$, the empirical measures

are unchanged and the evaluation vector becomes Pv. The covariance and Hessian factors therefore transform as

$$C' = PCP^\top, \qquad M'_Z = PM_ZP^\top, \qquad Q' = PQP^\top,$$
$$B' = PBP^\top, \qquad (B')^\dagger = PB^\dagger P^\top.$$

For the square root, $(PC^{1/2}P^\top)^2 = PCP^\top$ and the first matrix is positive semidefinite; uniqueness of the positive square root gives the identity. The formulas for Q and $B^\dagger$ in Step 1 give the other identities by conjugating each matrix and its inverse. Hence

$$(B')^\dagger Pv + [I - (B')^\dagger B']Ph_\ell = P\{B^\dagger v + (I - B^\dagger B)h_\ell\} = P\mathbf{g}_\ell^N.$$

The empirical increment transforms to $PM_X\mathbf{g}_\ell^N$, and the reference increment to Pv. The drift formulas also commute with P. The iid initial sample and independent standard normals have permutation-invariant joint law, with the population innovations unchanged. Induction over ODE intervals and jumps therefore shows that $(X_t^i, Y_t^i)_{i=1}^N$ is exchangeable. This gives identical one-particle moment bounds; it does not assert independence of the empirical particles. □

With the square roots aligned, the Gaussian perturbation error reduces to a kernel covariance mismatch.

Lemma A.15 (Reduction to kernel mismatch)**.** *Recall $M = M_Y = \mathbf{M}^N(\rho_{\mathbf{Y}}, \mathbf{Y})$ and $B = C^{1/2}\operatorname{sign}_+(M_Y)$ from lemmas A.13 and A.14, where $\mathbf{Y} = \mathbf{Y}_{t_\ell -}$. Let $\mathbf{g}^N \mid \mathcal{H}_{\ell-} \sim N(0, I_{Nd})$ be the coordinate constructed above, with B, M, C measurable before the jump. Then*

$$\mathbb{E}\big[\|(B-M)\mathbf{g}^N\|_N^2 \mid \mathcal{H}_{\ell-}\big] = \frac{1}{N}\|B-M\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}^2 \le \frac{1}{N}\|C-M^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}},$$

where $\|\cdot\|_N := \frac{1}{\sqrt{N}}\|\cdot\|_{\mathbb{R}^{dN}}$. Equivalently,

$$\mathbb{E}\big[\|(B-M)\mathbf{g}^N\|_N^2 \mid \mathcal{H}_{\ell-}\big] \le \frac{1}{N}\Big\|\big[K_\mu(Y^i,Y^j) - K_{\rho_{\mathbf{Y}}^N}(Y^i,Y^j)\big]_{i,j=1}^N\Big\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.$$

Proof. Writing $A = B - M$ in this calculation, conditional standard Gaussianity gives

$$\begin{aligned}\mathbb{E}[\|A\mathbf{g}^N\|_N^2 \mid \mathcal{H}_{\ell-}] &= \frac{1}{N}\sum_{a,b,c} A_{ab}A_{ac}\mathbb{E}[g_b^N g_c^N \mid \mathcal{H}_{\ell-}] \\ &= \frac{1}{N}\sum_{a,b,c} A_{ab}A_{ac}\delta_{bc} = \frac{1}{N}\sum_{a,b} A_{ab}^2 = \frac{1}{N}\|A\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}^2.\end{aligned}$$

Apply lemma A.13. The block identity

$$(M^2)_{ij} = \frac{1}{N}\sum_k \nabla_W^2 F(\rho_{\mathbf{Y}}; Y^i, Y^k)\nabla_W^2 F(\rho_{\mathbf{Y}}; Y^k, Y^j) = K_{\rho_{\mathbf{Y}}}(Y^i, Y^j)$$

then gives the stated kernel difference. □

The kernel mismatch is split into a smooth measure-dependence term and the empirical covariance fluctuation of the reference cloud.

Lemma A.16 (Kernel mismatch bound)**.** *Write $H_\nu(x,y) = \nabla_W^2 F(\nu; x, y)$ for the kernel in this lemma. Work on $E_{\mathrm{pop}} \cap E_{T,D}$ from lemma A.20 and eq. (36), with $\mathrm{L}_D = \mathrm{L}_{K_{\mathrm{pop}},D}$ from definition A.4. By lemma A.3, the moments of $\mu_t, \rho_{\mathbf{Y}_t}$ and the segment between them lie in the reachable moment region. For $\|x\|_{\mathbb{R}^d}, \|y\|_{\mathbb{R}^d} \le D$ and these measures, Lemma A.6 gives $\|H_\nu(x,y)\|_{\mathbb{R}^d\to\mathbb{R}^d} \le \mathrm{L}_D$. Then the block matrix*

$$\Delta K^N(\mu, \mathbf{Y}) := \big[K_\mu(Y^i,Y^j) - K_{\rho_{\mathbf{Y}}^N}(Y^i,Y^j)\big]_{i,j=1}^N \in \mathbb{R}^{Nd\times Nd}$$

satisfies

$$\frac{1}{N}\|\Delta K^N(\mu_t, \mathbf{Y})\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le 2d\,\mathrm{L}_D\Gamma_t^N(D) + \tilde{\Gamma}_t^N$$

where

$$\mathbf{U}_\nu(z) := \begin{pmatrix} H_\nu(Y^1, z) \\ \vdots \\ H_\nu(Y^N, z) \end{pmatrix} \in \mathbb{R}^{Nd\times d},$$

$$\tilde{\Gamma}_t^N := \frac{1}{N}\left\| \int \mathbf{U}_{\mu_t}(z)\mathbf{U}_{\mu_t}(z)^\top (\mu_t - \rho_{\mathbf{Y}_t})(dz) \right\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}$$

The explicit factor N cancels from the coefficient $2d\mathrm{L}_D$. Dependence through $\mathrm{L}_D = \mathrm{L}_{K_{\mathrm{pop}},D}$ is retained, as specified in definition A.4.

Proof. **Step 1: separate the covariance and measure changes.** Fix t and write $\mathbf{Y} = \mathbf{Y}_t$, $\mu = \mu_t$, and $\rho = \rho_{\mathbf{Y}_t}$. Since

$$H_\mu(Y^i, z)^\top = H_\mu(z, Y^i),$$

we may write

$$K_\mu(Y^i, Y^j) = \int H_\mu(Y^i, z)H_\mu(z, Y^j)\,\mu(dz) = \int H_\mu(Y^i, z)H_\mu(Y^j, z)^\top\,\mu(dz).$$

Therefore, with

$$\mathbf{U}_\mu(z) = \begin{pmatrix} H_\mu(Y^1, z) \\ \vdots \\ H_\mu(Y^N, z) \end{pmatrix} \in \mathbb{R}^{Nd\times d},$$

we have the matrix identity

$$\big[K_\mu(Y^i, Y^j)\big]_{i,j=1}^N = \int \mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top\,\mu(dz).$$

Hence

$$\begin{aligned}\Delta K^N(\mu, \mathbf{Y}) = &\int \mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top\,(\mu - \rho)(dz) \\ &+ \int \Big(\mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top - \mathbf{U}_\rho(z)\mathbf{U}_\rho(z)^\top\Big)\,\rho(dz).\end{aligned}$$

Thus, by the triangle inequality,

$$\|\Delta K^N(\mu, \mathbf{Y})\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le I_1 + I_2,$$

where

$$I_1 := \int \big\|\mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top - \mathbf{U}_\rho(z)\mathbf{U}_\rho(z)^\top\big\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}\,\rho(dz),$$

and

$$I_2 := \left\| \int \mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top\,(\mu - \rho)(dz) \right\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}.$$

The two terms require different estimates. The measure in I_1 is the radius-D empirical law ρ, so the localized kernel bounds apply to its integrand. The signed measure in I_2 contains the unbounded population law μ; it is retained as the covariance residual $\widetilde{\Gamma}_t^N$ of eq. (150) and is controlled by the weighted-feature concentration argument.

Step 2: bound the sample-supported measure change. For $X, Y \in \mathbb{R}^{dN\times d}$ with columns $X = [x_1\ \cdots\ x_d]$, $Y = [y_1\ \cdots\ y_d]$, the sole possible nonzero singular value of $x_k y_k^\top$ is $\|x_k\|_{\mathbb{R}^{dN}}\|y_k\|_{\mathbb{R}^{dN}}$. The triangle and Cauchy–Schwarz inequalities give

$$\begin{aligned}\|XY^\top\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} &= \left\|\sum_{k=1}^d x_k y_k^\top\right\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \sum_{k=1}^d \|x_k\|_{\mathbb{R}^{dN}}\|y_k\|_{\mathbb{R}^{dN}} \\ &\le \left(\sum_k \|x_k\|_{\mathbb{R}^{dN}}^2\right)^{1/2}\left(\sum_k \|y_k\|_{\mathbb{R}^{dN}}^2\right)^{1/2} = \|X\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}\|Y\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}.\end{aligned} \tag{116}$$

The same argument works for any number of columns. For Hilbert–Schmidt maps $X, Y : E \to F$ between Hilbert spaces, truncate the common input orthonormal basis. The nuclear norm of the omitted sum is at most the product of the two square-summable column tails, which tends to zero. Thus $\|XY^*\|_{*;\,F\to F} \le \|X\|_{\mathrm{HS};\,E\to F}\|Y\|_{\mathrm{HS};\,E\to F}$ also holds in that setting. For I_1, use this bound and $AA^\top - BB^\top = (A-B)A^\top + B(A-B)^\top$. Hence

$$\begin{aligned}&\|\mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top - \mathbf{U}_\rho(z)\mathbf{U}_\rho(z)^\top\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}\\ &\quad\le \|\mathbf{U}_\mu(z) - \mathbf{U}_\rho(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}\|\mathbf{U}_\mu(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}\\ &\qquad + \|\mathbf{U}_\rho(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}\|\mathbf{U}_\mu(z) - \mathbf{U}_\rho(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}}.\end{aligned}$$

For ρ-almost every z, (36) gives $\|z\|_{\mathbb{R}^d} \le D$ and $\|Y^i\|_{\mathbb{R}^d} \le D$. The following local bounds therefore apply to the integrand of I_1:

$$\|\mathbf{U}_\eta(z)\|^2_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}} = \sum_{i=1}^N \|H_\eta(Y^i,z)\|^2_{F;\,\mathbb{R}^d\to\mathbb{R}^d} \le d\sum_{i=1}^N \|H_\eta(Y^i,z)\|^2_{\mathbb{R}^d\to\mathbb{R}^d} \le Nd\,\mathrm{L}_D^2,$$

so

$$\|\mathbf{U}_\eta(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}} \le \sqrt{Nd}\,\mathrm{L}_D \qquad (\eta\in\{\mu,\rho\}).$$

Moreover, eq. (46) gives the (weaker) bound

$$\|H_\mu(Y^i,z) - H_\rho(Y^i,z)\|_{\mathbb{R}^d\to\mathbb{R}^d} \le \Gamma_t^N(D), \qquad \Gamma_t^N(D) = \sum_{m\ge1}\gamma_m(D)\|M_m(\mu) - M_m(\rho)\|_{\mathsf{H}_m},$$

where the residual is defined in (82); hence

$$\|\mathbf{U}_\mu(z) - \mathbf{U}_\rho(z)\|^2_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}} \le d\sum_{i=1}^N \|H_\mu(Y^i,z) - H_\rho(Y^i,z)\|^2_{\mathbb{R}^d\to\mathbb{R}^d} \le Nd\,\Gamma_t^N(D)^2.$$

Thus

$$\|\mathbf{U}_\mu(z) - \mathbf{U}_\rho(z)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^{dN}} \le \sqrt{Nd}\,\Gamma_t^N(D).$$

Combining the above bounds,

$$\big\|\mathbf{U}_\mu(z)\mathbf{U}_\mu(z)^\top - \mathbf{U}_\rho(z)\mathbf{U}_\rho(z)^\top\big\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le 2Nd\,\mathrm{L}_D\,\Gamma_t^N(D).$$

Integrating in z yields

$$I_1 \le 2Nd\,\mathrm{L}_D\,\Gamma_t^N(D).$$

Dividing by N gives

$$N^{-1}I_1 \le 2d\mathrm{L}_D\Gamma_t^N(D).$$

The remaining term is exactly $N^{-1}I_2 = \widetilde{\Gamma}_t^N$. Its integral includes the population law μ_t; no bound $\|z\|_{\mathbb{R}^d} \le D$ is applied to that integral. This proves the claim. □

We now combine the Lipschitz comparison of empirical perturbations with the covariance mismatch bound to obtain the jump recursion for r_t.

Lemma A.17 (One-jump comparison)**.** *Use the coupling of Lemma A.14 at t_ℓ, and let $u \in (0,1)$. Define*

$$\Lambda(D) := \sum_{m\ge1}\gamma_m(D)L_m(1+2D^{m-1}), \qquad q(u) := 1 + \sqrt{2\log(2/u)},$$

$$c_J(D) := 1 + \sqrt{d}\{\Lambda(D) + 2\mathrm{L}_D\} + \sqrt{2d\mathrm{L}_D}.$$

There is an event $\Omega^{\mathrm{jump}}_{\ell,u}$ with $\mathbb{P}((\Omega^{\mathrm{jump}}_{\ell,u})^c \mid \mathcal{H}_{\ell-}) \le u$ such that, on this event intersected with the population moment and sample localization events,

$$r_{t_\ell} \le (1 + c_J(D)q(u)\eta_\ell)r_{t_\ell-} + c_J(D)q(u)\eta_\ell\big(\sqrt{\Gamma^N_{t_\ell-}(D)} + \sqrt{\widetilde{\Gamma}^N_{t_\ell-}}\big).$$

All residuals and kernel covariances in this statement are evaluated before the jump. Constants may depend on the deterministic population moment scale as specified in the localized-envelope convention.

Proof. **Step 1: control the two conditional Gaussian errors.** By eq. (15) and lemma A.14,

$$\mathbf{X}_{t_\ell} - \mathbf{Y}_{t_\ell} = \mathbf{X}_{t_\ell -} - \mathbf{Y}_{t_\ell -} + \eta_\ell\{(M_X - M_Y) + (M_Y - B)\}\mathbf{g}_\ell^N.$$

Divide its triangle inequality by $\sqrt{N}$ to get

$$r_{t_\ell} \le r_{t_\ell -} + \frac{\eta_\ell}{\sqrt{N}}\|(M_X - M_Y)\mathbf{g}_\ell^N\|_{\mathbb{R}^{dN}} + \frac{\eta_\ell}{\sqrt{N}}\|(M_Y - B)\mathbf{g}_\ell^N\|_{\mathbb{R}^{dN}}. \tag{117}$$

For a pre-jump measurable matrix A, the conditional standard Gaussian law gives $\mathbb{E}[\|A\mathbf{g}_\ell^N\|_{\mathbb{R}^{dN}}^2 \mid \mathcal{H}_{\ell-}] = \|A\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}^2$ as in lemma A.15. Hence its mean norm is at most $\|A\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}$ by Cauchy–Schwarz. Also

$$|\|Ag\|_{\mathbb{R}^{dN}} - \|Ag'\|_{\mathbb{R}^{dN}}| \le \|A(g-g')\|_{\mathbb{R}^{dN}} \le \|A\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}}\|g-g'\|_{\mathbb{R}^{dN}}.$$

The Gaussian concentration inequality for a Lipschitz function (Boucheron et al., 2013), applied conditionally with deviation $\|A\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}}\sqrt{2\log(2/u)}$, has failure probability $e^{-\log(2/u)} = u/2$. Using $\|A\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \|A\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}$ gives

$$\mathbb{P}\Big(\|A\mathbf{g}_\ell^N\|_{\mathbb{R}^{dN}} > (1+\sqrt{2\log(2/u)})\|A\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \,\Big|\, \mathcal{H}_{\ell-}\Big) \le u/2.$$

Apply this to the two matrices in (117) and take a union bound. These Gaussian events are defined before imposing any sample localization event, so no conditional independence after localization is used.

Step 2: compare the two empirical Hessian factors. Write $X^i = X^i_{t_\ell -}$, $Y^i = Y^i_{t_\ell -}$, $\rho_X = \rho_{\mathbf{X}_{t_\ell -}}$, $\rho_Y = \rho_{\mathbf{Y}_{t_\ell -}}$, and $\delta_i = X^i - Y^i$. Use the paired empirical moment calculation (84) in lemma A.10, with $s = 1$ and the actual pre-jump configurations. On $E_{\mathrm{pop}} \cap E_{T,D}$ it gives

$$\sum_m \gamma_m(D)\|M_m(\rho_X) - M_m(\rho_Y)\|_{\mathsf{H}_m} \le \sum_m \gamma_m(D)L_m(1+2D^{m-1})r_{t_\ell -} = \Lambda(D)r_{t_\ell -},$$

using (82). Insert $\nabla_W^2 F(\rho_Y; X^i, X^j)$ between $\nabla_W^2 F(\rho_X; X^i, X^j)$ and $\nabla_W^2 F(\rho_Y; Y^i, Y^j)$. Apply (46) to the measure difference and (50) to the spatial difference. The two differences are bounded by $\frac{1}{2}\Lambda(D)r_{t_\ell -}$ and $\mathrm{L}_D(\|\delta_i\|_{\mathbb{R}^d} + \|\delta_j\|_{\mathbb{R}^d})$, respectively. Loosening $1/2$ to 1, for every i, j this yields

$$\|\nabla_W^2 F(\rho_X; X^i, X^j) - \nabla_W^2 F(\rho_Y; Y^i, Y^j)\|_{\mathbb{R}^d\to\mathbb{R}^d} \le \Lambda(D)r_{t_\ell -} + \mathrm{L}_D(\|\delta_i\|_{\mathbb{R}^d} + \|\delta_j\|_{\mathbb{R}^d}).$$

Here the same L_D bounds each spatial slot, by eq. (50). The block Frobenius norm is at most $\sqrt{d}$ times its operator norm. The $N^{-1/2}$ in each block gives exactly

$$\frac{1}{N}\|M_X - M_Y\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}^2 = \frac{1}{N^2}\sum_{i,j}\|\nabla_W^2 F(\rho_X; X^i, X^j) - \nabla_W^2 F(\rho_Y; Y^i, Y^j)\|_{F;\,\mathbb{R}^d\to\mathbb{R}^d}^2.$$

Minkowski's inequality for this normalized double sum gives

$$\begin{aligned}
N^{-1/2}\|M_X - M_Y\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} &\le \sqrt{d}\left[\frac{1}{N^2}\sum_{i,j}\{\Lambda(D)r_{t_\ell -} + \mathrm{L}_D\|\delta_i\|_{\mathbb{R}^d} + \mathrm{L}_D\|\delta_j\|_{\mathbb{R}^d}\}^2\right]^{1/2} \\
&\le \sqrt{d}\,\Lambda(D)r_{t_\ell -} + \sqrt{d}\,\mathrm{L}_D\left(\frac{1}{N^2}\sum_{i,j}\|\delta_i\|_{\mathbb{R}^d}^2\right)^{1/2} \\
&\quad + \sqrt{d}\,\mathrm{L}_D\left(\frac{1}{N^2}\sum_{i,j}\|\delta_j\|_{\mathbb{R}^d}^2\right)^{1/2} \\
&= \sqrt{d}\{\Lambda(D) + 2\mathrm{L}_D\}r_{t_\ell -},
\end{aligned}$$

since each squared double average equals $N^{-1}\sum_i \|\delta_i\|_{\mathbb{R}^d}^2 = r_{t_\ell -}^2$.

Step 3: use the population covariance residual. For the second matrix, the aligned square-root bound and the kernel mismatch lemma give

$$\begin{aligned} N^{-1/2}\|M_Y - B\|_{F;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}} &\le \sqrt{N^{-1}\|C - M_Y^2\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}} \\ &\le \sqrt{2d\mathrm{L}_D\Gamma^N_{t_\ell-}(D) + \widetilde{\Gamma}^N_{t_\ell-}} \\ &\le \sqrt{2d\mathrm{L}_D}\sqrt{\Gamma^N_{t_\ell-}(D)} + \sqrt{\widetilde{\Gamma}^N_{t_\ell-}}. \end{aligned}$$

The middle inequality is lemma A.16; the last uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Substitute into (117) on the Gaussian event from Step 1:

$$\begin{aligned} r_{t_\ell} \le r_{t_\ell-} + \eta_\ell q(u)\big[\sqrt{d}\{\Lambda(D) + 2\mathrm{L}_D\}r_{t_\ell-} \\ + \sqrt{2d\mathrm{L}_D}\sqrt{\Gamma^N_{t_\ell-}(D)} + \sqrt{\widetilde{\Gamma}^N_{t_\ell-}}\big]. \end{aligned}$$

The defined $c_J(D)$ bounds each of the three coefficients in brackets, which gives the asserted recursion. More explicitly, its leading 1 bounds the coefficient of $\sqrt{\widetilde{\Gamma}^N_{t_\ell-}}$, while the other two summands bound the coefficients of $r_{t_\ell-}$ and $\sqrt{\Gamma^N_{t_\ell-}(D)}$, respectively. □

Remark A.1. *The aligned matrix and the whitening coordinate are used to compare the empirical transition with a finite evaluation of the population field. Conditional i.i.d. of the reference particles is a separate population pushforward statement, proved in Lemma A.19; it does not condition on* $\mathbf{g}_\ell^N$.

A.5 Concentration of the stochastic residuals Γ_t^N and $\widetilde{\Gamma}_t^N$

We first record a bounded-difference inequality used to turn sensitivity bounds for empirical functionals into high-probability estimates.

A.5.1 Preliminary

McDiarmid's inequality will be applied conditionally on the population-level randomness, with the reference particles treated as independent inputs.

Lemma A.18 (McDiarmid's inequality (textbook: Boucheron et al. (2013)))**.** *Let $X_1, \dots, X_N$ be independent random variables taking values in measurable spaces $\mathcal{X}_1, \dots, \mathcal{X}_N$, and let $F : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ be measurable. Assume that there exist constants $c_1, \dots, c_N \ge 0$ such that for every $i \in \{1, \dots, N\}$,*

$$|F(x_1, \dots, x_N) - F(x_1, \dots, x_i', \dots, x_N)| \le c_i$$

for all x_j, x_i' for which the two arguments differ only in the i-th coordinate. If $\sum_i c_i^2 > 0$, then, for every $r > 0$,

$$\mathbb{P}(F(X_1, \dots, X_N) - \mathbb{E}[F(X_1, \dots, X_N)] \ge r) \le \exp\left(-\frac{2r^2}{\sum_{i=1}^N c_i^2}\right).$$

In particular,

$$\mathbb{P}(|F(X_1, \dots, X_N) - \mathbb{E}[F(X_1, \dots, X_N)]| \ge r) \le 2\exp\left(-\frac{2r^2}{\sum_{i=1}^N c_i^2}\right).$$

If every $c_i = 0$, changing coordinates one at a time gives $|F(x) - F(y)| \le \sum_i c_i = 0$ for any two inputs. Thus F is constant and both displayed tail probabilities are zero for $r > 0$. The parallel statement holds for the conditionally independent variables.

The jump covariance residual is controlled by a Hilbert–Schmidt second-moment estimate and McDiarmid's inequality; see lemma A.30.

A.5.2 CONCENTRATION OF Γ_t^N

We first control Γ_t^N, the weighted moment fluctuation of the reference empirical cloud. The key input is that, after conditioning on the common population-level jump randomness, the reference particles remain i.i.d. with law μ_t.

A formal proof of conditionally i.i.d. property.

We make the conditional i.i.d. structure explicit, since all concentration bounds below are applied after conditioning on the common jump randomness.

Definition A.5 (Conditional i.i.d. reference particles)**.** *For the population field histories $\mathcal{G}_\ell$ fixed in the notation section, conditional i.i.d. means*

$$\mathrm{Law}((Y_t^1,\dots,Y_t^N)\mid \mathcal{G}_\ell)=\mu_t^{\otimes N},\qquad t\in[t_\ell,t_{\ell+1}),$$

with the pre-jump trace at $t_{\ell+1}-$. A terminal post-jump value at T uses $\mathcal{G}_{L_{\mathrm{jump}}}$. Conditioning instead on $\mathcal{G}_{\mathrm{pop}}=\mathcal{G}_{L_{\mathrm{jump}}}$ gives the same identity for every time and each of its relevant traces.

Lemma A.19 (The population GP jump scheme preserves conditional i.i.d.)**.** *The population-first construction in Definition A.1 and Assumption A.2 has the conditional i.i.d. property in Definition A.5.*

Proof. Use the conditional product calculation (115) in lemma A.14: the common map $\mathcal{T}_t$ satisfies $Y_t^i=\mathcal{T}_t(U_i)$ and $\mu_t=(\mathcal{T}_t)_{\#}\mu_0$, also at each left trace. On $[t_\ell,t_{\ell+1})$ and its final left trace, these maps and laws are already $\mathcal{G}_\ell$-measurable. For bounded measurable f_i, the tower property gives

$$\begin{aligned}\mathbb{E}\left[\prod_i f_i(Y_t^i)\,\middle|\,\mathcal{G}_\ell\right]&=\mathbb{E}\left[\mathbb{E}\left[\prod_i f_i(Y_t^i)\,\middle|\,\mathcal{G}_{\mathrm{pop}}\right]\middle|\mathcal{G}_\ell\right]\\&=\mathbb{E}\left[\prod_i\int f_i(y)\mu_t(dy)\,\middle|\,\mathcal{G}_\ell\right]=\prod_i\int f_i(y)\mu_t(dy).\end{aligned}$$

A terminal post-jump trace uses $\mathcal{G}_{L_{\mathrm{jump}}}$. The finite Gaussian coordinate is included in neither conditioning field. □

Unconditional regularity and a conditional moment event. In the unconditional moment assumption, write K_0 for the deterministic one-particle moment bound. Conditional moment bounds will be consequences on the event below, rather than almost-sure assumptions with the same K_0. The population sigma-field $\mathcal{G}_{\mathrm{pop}}$ is generated by the population fields in the population-first coupling. In particular, $\mu_t=\mathrm{Law}(Y_t^1\mid\mathcal{G}_{\mathrm{pop}})$.

Lemma A.20 (Conditional moments on a population-measurable event)**.** *Let $0<t_1<\cdots<t_{L_{\mathrm{jump}}}\le T$ be deterministic jump times. Suppose that the paths X^1,Y^1 have right values and left limits at these times and are B-Lipschitz on every intervening open interval, almost surely, with the continuous branch extensions and endpoint traces of definition A.1, where $B\ge 0$ is deterministic. Assume that, for $S\in\{X,Y\}$,*

$$\sup_{0\le t\le T}\sup_{q\ge 1}q^{-1/\alpha}\big(\mathbb{E}\|S_t^1\|_{\mathbb{R}^d}^q\big)^{1/q}\le K_0,\qquad 0<\alpha\le 2. \tag{118}$$

The same bounds at left limits are understood by Fatou's lemma. For $p_0\in(0,1)$, define

$$Q=\lceil(1+B)T\rceil+2(L_{\mathrm{jump}}+1),\qquad b=\max\{1,K_0\{2e(1+\alpha)\}^{1/\alpha}\},\qquad A_0=\frac{4Q}{p_0}.$$

Here b is a deterministic exponential-moment scale, Q bounds the number of grid time-side pairs, and A_0 is the threshold allocating failure probability p_0 to their conditional moments. Set $c_\alpha=2^{(\alpha-1)_+}$ and

$$b_{\mathrm{flow}}=c_\alpha^{1/\alpha}b,\qquad M_{\mathrm{pop}}=A_0\exp\big(b^{-\alpha}\big)\le\frac{4eQ}{p_0}, \tag{119}$$

and define

$$k_{\mathrm{pop}}^\alpha=\frac{c_\alpha}{\log 2}\{b^\alpha\log A_0+1\},\qquad K_{\mathrm{pop}}=C_\alpha k_{\mathrm{pop}}, \tag{120}$$

where C_α depends only on α and can be chosen so that an exponential moment bound with scale k implies the moment bound $C_\alpha k q^{1/\alpha}$ for all $q \geq 1$. There is an event $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$ with $\mathbb{P}(E_{\mathrm{pop}}) \geq 1 - p_0$ such that, simultaneously for $S \in \{X, Y\}$, all $t \in [0, T]$, and both sides of every jump,

$$\mathbb{E}\left[\exp\left\{\frac{\|S_t^1\|_{\mathbb{R}^d}^\alpha}{b_{\mathrm{flow}}^\alpha}\right\}\middle|\, \mathcal{G}_{\mathrm{pop}}\right] \leq M_{\mathrm{pop}}, \tag{121}$$

as well as

$$\mathbb{E}\left[\exp\left\{\frac{\|S_t^1\|_{\mathbb{R}^d}^\alpha}{k_{\mathrm{pop}}^\alpha}\right\}\middle|\, \mathcal{G}_{\mathrm{pop}}\right] \leq 2, \tag{122}$$

and hence

$$\sup_{q \geq 1} q^{-1/\alpha} \left(\mathbb{E}[\|S_t^1\|_{\mathbb{R}^d}^q \mid \mathcal{G}_{\mathrm{pop}}]\right)^{1/q} \leq K_{\mathrm{pop}}. \tag{123}$$

In particular $\mu_t \in \mathcal{P}_{\alpha, K_{\mathrm{pop}}}$ on this event. No conditional independence of the X particles is asserted or needed.

Proof. **Step 1: an unconditional exponential moment.** For $q < 1$, Lyapunov's inequality and the $q = 1$ case of (118) give $\mathbb{E}\|S_t^1\|_{\mathbb{R}^d}^q \leq K_0^q$. Consequently, for each integer $j \geq 1$,

$$\mathbb{E}\|S_t^1\|_{\mathbb{R}^d}^{\alpha j} \leq K_0^{\alpha j} \max\{1, \alpha j\}^j.$$

Since $j! \geq (j/e)^j$, $\max\{1, \alpha j\} \leq (1+\alpha)j$, and $b^\alpha \geq 2e(1+\alpha)K_0^\alpha$, we have

$$\frac{\mathbb{E}\|S_t^1\|_{\mathbb{R}^d}^{\alpha j}}{j! b^{\alpha j}} \leq \left(\frac{eK_0^\alpha \max\{1, \alpha j\}}{j b^\alpha}\right)^j \leq \left(\frac{eK_0^\alpha (1+\alpha)}{b^\alpha}\right)^j \leq 2^{-j}.$$

Monotone convergence for the nonnegative exponential series therefore gives

$$\mathbb{E}\exp\{\|S_t^1\|_{\mathbb{R}^d}^\alpha / b^\alpha\} \leq 1 + \sum_{j \geq 1} 2^{-j} = 2.$$

The same inequality holds at a left trace by Fatou's lemma.

Step 2: one grid event and a common conditional path law. The deterministic branch speed B is used here to transfer moment bounds from a finite time grid to every time on each branch. It is fixed before the random sample-localization event $E_{T,D}$ is introduced. Set $t_0 = 0$ and $t_{L_{\mathrm{jump}}+1} = T$. On each nonempty ODE branch choose a deterministic grid of mesh at most $(1+B)^{-1}$, including the initial right value and final left trace. The number of entries on this branch is at most $\lceil (1+B)(t_{\ell+1} - t_\ell) \rceil + 1$. At a jump the two traces are separate entries. If $t_{L_{\mathrm{jump}}} = T$, the zero-length final branch contributes the single post-jump value at T. Thus the set $\mathcal{T}$ of time-side pairs satisfies

$$\#\mathcal{T} \leq \sum_{\ell=0}^{L_{\mathrm{jump}}} \left(\lceil (1+B)(t_{\ell+1} - t_\ell) \rceil + 1\right) \leq (1+B)T + 2(L_{\mathrm{jump}} + 1) \leq Q.$$

To make the branchwise Lipschitz inequalities hold simultaneously under the same conditional law at every grid point and time trace, use a regular conditional law of the pair of paths (X^1, Y^1) given $\mathcal{G}_{\mathrm{pop}}$. Equivalently, regard each path as its finitely many continuous branch extensions with the prescribed endpoint traces, together with any isolated terminal post-jump value. This is a standard Borel path space. All conditional expectations in this proof are evaluated using this one conditional path law. Define the countable intersection

$$\mathcal{L} := \bigcap_{S \in \{X,Y\}} \bigcap_{\ell=0}^{L_{\mathrm{jump}}} \bigcap_{s,t \in \mathbb{Q} \cap (t_\ell, t_{\ell+1})} \{\|S_t^1 - S_s^1\|_{\mathbb{R}^d} \leq B|t-s|\}.$$

The assumed branchwise Lipschitz property gives $\mathbb{P}(\mathcal{L}) = 1$. The tower property implies

$$\mathbb{E}[\mathbb{P}(\mathcal{L}^c \mid \mathcal{G}_{\mathrm{pop}})] = \mathbb{P}(\mathcal{L}^c) = 0.$$

Since the conditional probability is nonnegative, it is zero almost surely. Continuity on each branch extends the rational-time inequalities to all time pairs and to that branch's endpoint traces. Thus these inequalities hold simultaneously under the conditional path law outside one population-null set.

For $s \in \mathcal{T}$, put

$$Z_s^S := \mathbb{E}[\exp\{\|S_s^1\|_{\mathbb{R}^d}^\alpha / b^\alpha\} \mid \mathcal{G}_{\mathrm{pop}}], \qquad E_{\mathrm{pop}} := \{\mathbb{P}(\mathcal{L} \mid \mathcal{G}_{\mathrm{pop}}) = 1\} \cap \bigcap_{S \in \{X,Y\}} \bigcap_{s \in \mathcal{T}} \{Z_s^S \le A_0\}.$$

The event is $\mathcal{G}_{\mathrm{pop}}$-measurable. Step 1, the tower property, Markov's inequality, and the finite union bound give

$$\mathbb{P}(E_{\mathrm{pop}}^c) \le \sum_{S \in \{X,Y\}} \sum_{s \in \mathcal{T}} \frac{\mathbb{E} Z_s^S}{A_0} \le (2Q)\frac{2}{A_0} = p_0.$$

Step 3: extend both exponential bounds to all times. Fix a population realization in E_{pop}. For any time-side pair t on a nonempty branch, choose $s \in \mathcal{T}$ on that branch with $|t - s| \le (1 + B)^{-1}$. Under the conditional path law,

$$\|S_t^1\|_{\mathbb{R}^d} \le \|S_s^1\|_{\mathbb{R}^d} + B|t - s| \le \|S_s^1\|_{\mathbb{R}^d} + 1.$$

For an isolated terminal value, use $s = t$. The inequality $(u + v)^\alpha \le c_\alpha(u^\alpha + v^\alpha)$ and $b_{\mathrm{flow}}^\alpha = c_\alpha b^\alpha$ from (119) give

$$\mathbb{E}[\exp\{\|S_t^1\|_{\mathbb{R}^d}^\alpha / b_{\mathrm{flow}}^\alpha\} \mid \mathcal{G}_{\mathrm{pop}}] \le e^{b^{-\alpha}} Z_s^S \le e^{b^{-\alpha}} A_0 = M_{\mathrm{pop}}.$$

This proves (121); its fixed scale will be retained when estimating the clipping bias.

For the normalized exponential bound, put $\eta = \log 2 / \log M_{\mathrm{pop}} \in (0, 1)$. By eqs. (119) and (120), $k_{\mathrm{pop}}^\alpha = b_{\mathrm{flow}}^\alpha / \eta$. Apply conditional Jensen directly to the all-time estimate (121):

$$\begin{aligned}\mathbb{E}\left[e^{\|S_t^1\|_{\mathbb{R}^d}^\alpha / k_{\mathrm{pop}}^\alpha} \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right] &= \mathbb{E}\left[\left(e^{\|S_t^1\|_{\mathbb{R}^d}^\alpha / b_{\mathrm{flow}}^\alpha}\right)^\eta \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right] \\ &\le \left(\mathbb{E}\left[e^{\|S_t^1\|_{\mathbb{R}^d}^\alpha / b_{\mathrm{flow}}^\alpha} \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right]\right)^\eta \le M_{\mathrm{pop}}^\eta = 2.\end{aligned}$$

The same conditional path law is used at all time traces, proving (122) simultaneously.

Step 4: obtain every ordinary conditional moment. Fix $q \ge 1$. For $u > 0$,

$$\frac{d}{du} \log\{u^q e^{-(u/k_{\mathrm{pop}})^\alpha}\} = \frac{q}{u} - \frac{\alpha u^{\alpha-1}}{k_{\mathrm{pop}}^\alpha}.$$

This derivative is positive before $u = k_{\mathrm{pop}}(q/\alpha)^{1/\alpha}$ and negative after it. Evaluation at that maximum, also including $u = 0$ by continuity, gives

$$u^q \le k_{\mathrm{pop}}^q \left(\frac{q}{\alpha e}\right)^{q/\alpha} e^{(u/k_{\mathrm{pop}})^\alpha} \qquad (u \ge 0).$$

Integrate under the same conditional path law and use (122) to obtain

$$\left(\mathbb{E}[\|S_t^1\|_{\mathbb{R}^d}^q \mid \mathcal{G}_{\mathrm{pop}}]\right)^{1/q} \le 2^{1/q} k_{\mathrm{pop}} \left(\frac{q}{\alpha e}\right)^{1/\alpha} \le C_\alpha k_{\mathrm{pop}} q^{1/\alpha}, \qquad C_\alpha = 2\max\{1, (\alpha e)^{-1/\alpha}\}.$$

This proves (123) for every $q \ge 1$ and every time trace on E_{pop}. Finally, $\mu_t = \mathrm{Law}(Y_t^1 \mid \mathcal{G}_{\mathrm{pop}})$ implies $\mu_t \in \mathcal{P}_{\alpha, K_{\mathrm{pop}}}$ there. □

Probability bookkeeping. The event E_{pop} is measurable with respect to the sigma-field under which the reference samples are conditionally iid. Thus all conditional concentration arguments can be applied for each population realization in this event, with the deterministic scale K_{pop}. If their conditional failure probability is at most p_1, the unconditional failure probability, including $E_{\mathrm{pop}}^{\mathrm{c}}$, is at most $p_0 + p_1$. Conditioning on a sample-localization event is not used to claim independence. If only the reference moment bound is needed, the X constraints can be omitted and A_0 replaced by $2Q/p_0$.

Conditional concentration of Γ_t^N

We first prove a localized uniform-in-time bound: truncate the features, apply conditional McDiarmid at grid times, and then remove the truncation.

Throughout this subsection, the jump times are deterministic, with $0 = t_0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \le T$ and $t_{L_{\mathrm{jump}}+1} = T$. Every supremum and time grid includes both traces at each jump. More precisely, the closure of an ODE interval has its post-jump value at the left endpoint and its pre-jump value at the right endpoint. The last interval ends at T, and a jump at T has an additional terminal post-jump value. An interval of zero length contributes that terminal value. We condition only on the full population randomness $\mathcal{G}_{\mathrm{pop}}$. Conditional on it, the reference trajectories are independent copies of the same population trajectory and have marginal law μ_t.

Let $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$ be the event supplied by the population-moment localization lemma. Retain both of its conclusions: on E_{pop}, for deterministic $K_{\mathrm{pop}} > 0$, $K_0 > 0$, $c_0 = (K_0/b_{\mathrm{flow}})^\alpha > 0$, and $M_{\mathrm{pop}} \ge 1$,

$$\begin{aligned} &\sup_{t\le T}\sup_{r\ge 1} r^{-1/\alpha}\left(\int \|y\|_{\mathbb{R}^d}^r\,\mu_t(dy)\right)^{1/r} \le K_{\mathrm{pop}},\\ &\sup_{t\le T}\int \exp\!\left(c_0\left(\frac{\|y\|_{\mathbb{R}^d}}{K_0}\right)^\alpha\right)\mu_t(dy) \le M_{\mathrm{pop}}. \end{aligned} \tag{124}$$

Here K_0 is the unconditional moment scale; K_{pop} and M_{pop} are deterministic functions of the probability budget and the time grid. In particular, they are not asserted to be bounded uniformly in the realized population Gaussian fields.

Fix $D \ge 1$. The deterministic weights from eq. (44) are

$$\gamma_m = \gamma_m(K_{\mathrm{pop}}, D) = \mathrm{L}_D a_m.$$

Define the radial projection onto the radius-D ball by

$$\Pi_D(y) = y\min\{1, D/\|y\|_{\mathbb{R}^d}\} \quad (y \ne 0), \qquad \Pi_D(0) = 0.$$

As the metric projection onto a closed convex ball, $\|\Pi_D x - \Pi_D y\|_{\mathbb{R}^d} \le \|x - y\|_{\mathbb{R}^d}$. Let $\phi_m^{(D)} = \phi_m \circ \Pi_D$. The feature assumption assumption A.1(i), (20), gives $\|\phi_m^{(D)}(y)\|_{\mathsf{H}_m} \le L_m(1 + D^m)$. This amplitude bounds the effect of replacing one sample in McDiarmid's inequality. The spatial Lipschitz bound below is used separately for time interpolation. Define the respective envelopes

$$A(D, K_{\mathrm{pop}}) = 2\sum_{m\ge 1}\gamma_m L_m(1 + D^m), \qquad \Lambda(D, K_{\mathrm{pop}}) = \sum_{m\ge 1}\gamma_m L_m(1 + 2D^{m-1}). \tag{125}$$

For $\|u\|_{\mathbb{R}^d}, \|v\|_{\mathbb{R}^d} \le D$, the segment $v + s(u - v)$ remains in the ball. Integrating the first derivative in (20) gives

$$\|\phi_m(u) - \phi_m(v)\|_{\mathsf{H}_m} \le \int_0^1 L_m(1 + D^{m-1})\|u - v\|_{\mathbb{R}^d}\,ds = L_m(1 + D^{m-1})\|u - v\|_{\mathbb{R}^d}.$$

Apply this with $u = \Pi_D x$, $v = \Pi_D y$ and use the 1-Lipschitz bound above. Thus $\phi_m^{(D)}$ is globally Lipschitz with constant at most $L_m(1+D^{m-1}) \le L_m(1+2D^{m-1})$, as used in (125). Differentiability of Π_D on the sphere is not needed. Define also

$$\begin{aligned} S(D, K_{\mathrm{pop}}) &= \sum_{m\ge 1}\gamma_m L_m\left[1 + \{K_{\mathrm{pop}}(2m)^{1/\alpha}\}^m\right],\\ \overline{\beta}_T(D) &= 2S(D, K_{\mathrm{pop}})\sqrt{M_{\mathrm{pop}}}\exp\left[-\frac{c_0}{2}\left(\frac{D}{K_0}\right)^\alpha\right]. \end{aligned} \tag{126}$$

Lemma A.21 (Localized uniform high-probability bound under the GP jump scheme)**.** *Assume the conditional i.i.d. conclusion of lemma A.19, the feature bounds in assumption A.1(i), and assumption A.4 with deterministic constant B. Assume* (124) *on E_{pop}, and assume the sums in* (125) *and* (126) *are finite. For $N \ge 2$ and $p_\Gamma \in (0, 1)$, set $Q_N = (L_{\mathrm{jump}} + 1)(N + 1)$ and*

$$\delta_\Gamma(D) = \overline{\beta}_T(D) + \frac{A(D, K_{\mathrm{pop}})}{\sqrt{N}} + A(D, K_{\mathrm{pop}})\sqrt{\frac{\log(Q_N/p_\Gamma)}{2N}} + \frac{2BT\Lambda(D, K_{\mathrm{pop}})}{N}. \tag{127}$$

Then

$$\mathbb{P}\left(E_{\mathrm{pop}} \cap E_{T,D} \cap \left\{\sup_{t\le T}\Gamma_t^N > \delta_\Gamma(D)\right\}\right) \le p_\Gamma. \tag{128}$$

Consequently, if $\mathbb{P}(E_{\mathrm{pop}}^c) \le p_{\mathrm{pop}}$ *and* $\mathbb{P}(E_{T,D}^c) \le p_{\mathrm{rad}}$*, then*

$$\mathbb{P}\left(E_{\mathrm{pop}} \cap E_{T,D} \cap \left\{\sup_{t\le T} \Gamma_t^N \le \delta_\Gamma(D)\right\}\right) \ge 1 - p_{\mathrm{pop}} - p_{\mathrm{rad}} - p_\Gamma.$$

Proof. **Step 1: concentrate the clipped empirical moments.** Condition on $\mathcal{G}_{\mathrm{pop}}$, without conditioning on $E_{T,D}$, and write $\mathbb{E}_{\mathcal{G}}[Z] = \mathbb{E}[Z \mid \mathcal{G}_{\mathrm{pop}}]$. At a fixed time trace, let

$$F_t^D = \sum_{m\ge 1} \gamma_m \left\| \frac{1}{N}\sum_{i=1}^N \phi_m^{(D)}(Y_t^i) - \int \phi_m^{(D)}(y)\, \mu_t(dy) \right\|_{\mathsf{H}_m}.$$

If a single sample y is replaced by y', the reverse triangle inequality and $\|\phi_m^{(D)}(y)\|_{\mathsf{H}_m} \le L_m(1+D^m)$ give

$$|F_t^D(y, Y^2, \ldots) - F_t^D(y', Y^2, \ldots)| \le \frac{1}{N}\sum_m \gamma_m \|\phi_m^{(D)}(y) - \phi_m^{(D)}(y')\|_{\mathsf{H}_m} \le \frac{A(D, K_{\mathrm{pop}})}{N}.$$

For each fixed m, conditional independence makes every off-diagonal term in the squared centered sample mean zero: for $i \ne j$,

$$\begin{aligned}&\mathbb{E}_{\mathcal{G}}\left\langle \phi_m^{(D)}(Y_t^i) - \mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^i), \phi_m^{(D)}(Y_t^j) - \mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^j)\right\rangle \\ &\quad = \left\langle \mathbb{E}_{\mathcal{G}}(\phi_m^{(D)}(Y_t^i) - \mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^i)), \mathbb{E}_{\mathcal{G}}(\phi_m^{(D)}(Y_t^j) - \mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^j))\right\rangle = 0.\end{aligned}$$

The remaining diagonal terms give

$$\begin{aligned}&\mathbb{E}_{\mathcal{G}}\left\| \frac{1}{N}\sum_i \phi_m^{(D)}(Y_t^i) - \mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^1)\right\|_{\mathsf{H}_m}^2 \\ &\quad = \frac{1}{N}\left(\mathbb{E}_{\mathcal{G}}\|\phi_m^{(D)}(Y_t^1)\|_{\mathsf{H}_m}^2 - \|\mathbb{E}_{\mathcal{G}}\phi_m^{(D)}(Y_t^1)\|_{\mathsf{H}_m}^2\right) \le \frac{L_m^2(1+D^m)^2}{N}.\end{aligned}$$

Cauchy–Schwarz bounds the expected norm by the square root of this quantity. Summing with the nonnegative weights γ_m gives

$$\mathbb{E}[F_t^D \mid \mathcal{G}_{\mathrm{pop}}] \le \frac{1}{\sqrt{N}}\sum_m \gamma_m L_m(1+D^m) \le \frac{A(D, K_{\mathrm{pop}})}{\sqrt{N}}.$$

Apply lemma A.18 with squared-difference sum $N\{A(D, K_{\mathrm{pop}})/N\}^2 = A(D, K_{\mathrm{pop}})^2/N$. It gives, for every $u > 0$,

$$\mathbb{P}\left(F_t^D > \frac{A(D, K_{\mathrm{pop}})}{\sqrt{N}} + u \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right) \le \exp\left(-\frac{2Nu^2}{A(D, K_{\mathrm{pop}})^2}\right).$$

When $A = 0$, this conclusion is read as $F_t^D = 0$. These arguments apply first to finite partial sums and then to the infinite weighted sum by the finite envelope and monotone convergence.

Step 2: extend the estimate between grid points. Subdivide each closed ODE interval, with its correctly labelled endpoint traces, into N equal subintervals. The union has at most Q_N nodes. For s, t on the same ODE interval, use the clipped-feature Lipschitz bound from (125). The reverse triangle inequality separates the empirical and population changes as

$$\begin{aligned}|F_t^D - F_s^D| &\le \sum_m \gamma_m L_m(1+2D^{m-1})\left(\frac{1}{N}\sum_i \|Y_t^i - Y_s^i\|_{\mathbb{R}^d} + \mathbb{E}_{\mathcal{G}}\|Y_t^1 - Y_s^1\|_{\mathbb{R}^d}\right) \\ &\le 2B\Lambda(D, K_{\mathrm{pop}})|t-s|.\end{aligned}$$

At each grid node substitute $u = A(D, K_{\mathrm{pop}})\sqrt{\log(Q_N/p_\Gamma)/(2N)}$ in the preceding concentration inequality. When $A > 0$, the union over at most Q_N nodes costs

$$Q_N \exp\left(-\frac{2N}{A(D, K_{\mathrm{pop}})^2}\frac{A(D, K_{\mathrm{pop}})^2 \log(Q_N/p_\Gamma)}{2N}\right) = p_\Gamma.$$

When $A = 0$, $F_t^D = 0$ directly. Every time trace lies within T/N of a grid node on its own branch. The time Lipschitz estimate adds at most $2BT\Lambda(D, K_{\mathrm{pop}})/N$. Consequently, for each population realization in E_{pop},

$$\mathbb{P}\left(\sup_{t\le T} F_t^D > \delta_\Gamma(D) - \overline{\beta}_T(D) \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right) \le p_\Gamma. \tag{129}$$

Averaging this conditional bound preserves its failure budget.

Step 3: remove clipping on the sample event. On $E_{T,D}$, the clipped and original empirical feature averages agree. For $\|y\|_{\mathbb{R}^d} \le D$, the population integrand difference is zero; otherwise the feature growth bound and $D^m \le \|y\|_{\mathbb{R}^d}^m$ give

$$\|\phi_m(y) - \phi_m(\Pi_D y)\|_{\mathsf{H}_m} \le L_m(2 + \|y\|_{\mathbb{R}^d}^m + D^m) \le 2L_m(1 + \|y\|_{\mathbb{R}^d}^m).$$

Thus the triangle inequality gives

$$\Gamma_t^N \le F_t^D + \beta_T^{\mathcal{G}}(D), \qquad \beta_T^{\mathcal{G}}(D) = 2\sum_m \gamma_m L_m \sup_{s\le T} \int (1 + \|y\|_{\mathbb{R}^d}^m)\mathbf{1}_{\{\|y\|_{\mathbb{R}^d} > D\}}\, \mu_s(dy). \tag{130}$$

This is a conditional population integral, not the unconditional moment of a randomly selected population field. Lemma A.22, using only the population input (124), gives $\beta_T^{\mathcal{G}}(D) \le \overline{\beta}_T(D)$ on E_{pop}. Intersecting the event in (129) with E_{pop} and $E_{T,D}$ proves (128) and the final probability bound. □

Lemma A.22 (Conditional population truncation bias)**.** *On E_{pop}, the bias $\beta_T^{\mathcal{G}}(D)$ in* (130) *is bounded by $\overline{\beta}_T(D)$ in* (126)*. If only the first, moment-bound part of* (124) *is retained, the valid alternative, for $D \ge eK_{\mathrm{pop}}$, is*

$$\beta_T^{\mathcal{G}}(D) \le 2S(D, K_{\mathrm{pop}}) \exp\left[-\frac{1}{2}\left(\frac{D}{eK_{\mathrm{pop}}}\right)^\alpha\right]. \tag{131}$$

Proof. Write $u = c_0(D/K_0)^\alpha$. Conditional Markov, using the exponential-moment bound (121) of lemma A.20, and Cauchy–Schwarz give, uniformly over time traces,

$$\mu_t(\|y\|_{\mathbb{R}^d} > D) \le \min\{1, M_{\mathrm{pop}} e^{-u}\},$$

$$\int (1 + \|y\|_{\mathbb{R}^d}^m)\mathbf{1}_{\{\|y\|_{\mathbb{R}^d} > D\}}\, \mu_t(dy) \le \left[1 + \{K_{\mathrm{pop}}(2m)^{1/\alpha}\}^m\right] \sqrt{M_{\mathrm{pop}}} e^{-u/2}.$$

Indeed, Cauchy–Schwarz followed by the L^2 triangle inequality gives

$$\begin{aligned}\int (1 + \|y\|_{\mathbb{R}^d}^m)\mathbf{1}_{\{\|y\|_{\mathbb{R}^d} > D\}}\, \mu_t(dy) &\le \left(\int (1 + \|y\|_{\mathbb{R}^d}^m)^2\, \mu_t(dy)\right)^{1/2} \mu_t(\|y\|_{\mathbb{R}^d} > D)^{1/2} \\ &\le \left[1 + \left(\int \|y\|_{\mathbb{R}^d}^{2m}\, \mu_t(dy)\right)^{1/2}\right] \sqrt{\min\{1, M_{\mathrm{pop}} e^{-u}\}}.\end{aligned} \tag{132}$$

Now use (124) with $r = 2m$ and $\sqrt{\min\{1, z\}} \le \sqrt{z}$, also when $z > 1$. Summing with weights $2\gamma_m L_m$ and recalling S in (126) gives $\beta_T^{\mathcal{G}}(D) \le 2S(D, K_{\mathrm{pop}})\sqrt{M_{\mathrm{pop}}} e^{-u/2} = \overline{\beta}_T(D)$. For the moment-only version, set $r = (D/(eK_{\mathrm{pop}}))^\alpha \ge 1$. Markov's inequality and the moment input give

$$\mu_t(\|y\|_{\mathbb{R}^d} > D) \le D^{-r} \int \|y\|_{\mathbb{R}^d}^r \mu_t(dy) \le \left(\frac{K_{\mathrm{pop}} r^{1/\alpha}}{D}\right)^r = e^{-r}. \tag{133}$$

Substitute this probability into (132) and sum as in (130):

$$\beta_T^{\mathcal{G}}(D) \le 2\sum_m \gamma_m L_m [1 + \{K_{\mathrm{pop}}(2m)^{1/\alpha}\}^m] e^{-r/2} = 2S(D, K_{\mathrm{pop}}) e^{-r/2}.$$

This is (131). □

Lemma A.23 (Concrete bounds for the amplitude, Lipschitz, and tail sums)**.** *Let* $\kappa = \kappa_{\mathrm{dec}} > \max\{1, \kappa_{\mathrm{gr}}\}$ *and* $q = \kappa/(\kappa - 1)$*. Lemmas A.4 and A.6 give*

$$\gamma_m(K, D)L_m \le C \exp\left(C\{(\log(e + D))^q + (\log(e + K))^q\} - cm^\kappa\right), \tag{134}$$

with fixed $C, c > 0$*. In particular,*

$$A(D, K) \vee \Lambda(D, K) \vee S(D, K) \le C \exp\left(C\{(\log(e + D))^q + (\log(e + K))^q\}\right). \tag{135}$$

Here the constants do not depend on D, K, N, T, p*.*

Proof. By eqs. (38) and (44), $\gamma_m(K, D) = \mathrm{L}_{K,D} a_m$ and $\mathrm{L}_{K,D} \le Ce^{C\chi(K+D)}$. Use lemma A.1, (26), first without the braces (which are at least one). Since

$$\log(e + K + D) \le \log(e + K) + \log(e + D), \qquad \chi(K + D) \le 2^{q-1}\{\chi(K) + \chi(D)\},$$

this proves (134), with $c = c_a/4$ allowed. For the three sums, substitute in (27) as follows:

$$\begin{aligned}
A(D, K) &= 2\mathrm{L}_{K,D} \sum_m a_m L_m (1 + D^m) \le C\mathrm{L}_{K,D} e^{C\chi(K+D)}, \\
\Lambda(D, K) &\le 2\mathrm{L}_{K,D} \sum_m a_m L_m (2 + D^m) \le C\mathrm{L}_{K,D} e^{C\chi(K+D)}, \\
S(D, K) &= \mathrm{L}_{K,D} \sum_m a_m L_m [1 + \{2^{1/\alpha} K\}^m m^{m/\alpha}] \\
&\le C\mathrm{L}_{K,D} e^{C\chi(2^{1/\alpha} K+D)}.
\end{aligned}$$

Here $D \ge 1$ gives $1 + 2D^{m-1} \le 2(2 + D^m)$. Use (28) to absorb the fixed rescaling of $K + D$, then apply the displayed bound for $\chi(K + D)$. This proves (135). □

Lemma A.24 (Uniform radius bound for both particle clouds)**.** *Suppose each one-particle marginal in both clouds obeys the unconditional bound* $\sup_{t \le T, r \ge 1} r^{-1/\alpha} (\mathbb{E}\|Z_t^i\|_{\mathbb{R}^d}^r)^{1/r} \le K_0$*, where* $Z = X, Y$*, and both clouds have deterministic ODE-interval Lipschitz bound* B*. For* $D \ge 2eK_0$*,*

$$\mathbb{P}(E_{T,D}^c) \le 2N \left(2(L_{\mathrm{jump}} + 1) + \frac{2BT}{D}\right) \exp\left[-\left(\frac{D}{2eK_0}\right)^\alpha\right]. \tag{136}$$

In particular, $\mathbb{P}(E_{T,D}^c) \le p_{\mathrm{rad}}$ *when*

$$D \ge 2eK_0 \left[1 + \log \frac{2N\{2(L_{\mathrm{jump}} + 1) + BT/(eK_0)\}}{p_{\mathrm{rad}}}\right]^{1/\alpha}. \tag{137}$$

Proof. Apply the Markov calculation (133) of lemma A.22 to the unconditional law of Z_t^i, with K_{pop} replaced by K_0 and D by $v \ge eK_0$: $\mathbb{P}(\|Z_t^i\|_{\mathbb{R}^d} > v) \le \exp[-(v/(eK_0))^\alpha]$. For $B > 0$, choose a grid $\mathcal{T}_D$ of mesh at most $D/(2B)$ on each closed ODE branch, including both endpoint traces and any isolated terminal post-jump value. As in lemma A.20, $\#\mathcal{T}_D \le 2(L_{\mathrm{jump}} + 1) + 2BT/D$. If $\|Z_t^i\|_{\mathbb{R}^d} > D$, choose a grid node s on the same branch with $|t - s| \le D/(2B)$. Then

$$\begin{aligned}
\|Z_s^i\|_{\mathbb{R}^d} &\ge \|Z_t^i\|_{\mathbb{R}^d} - B|t - s| > D - D/2 = D/2, \\
E_{T,D}^c &\subseteq \bigcup_{Z \in \{X,Y\}} \bigcup_{i=1}^N \bigcup_{s \in \mathcal{T}_D} \{\|Z_s^i\|_{\mathbb{R}^d} > D/2\}.
\end{aligned}$$

For an isolated terminal value use $s = t$. Substituting $v = D/2 \ge eK_0$ in the preceding tail bound and summing over this union gives (136); independence is not needed. For $B = 0$, one node per constant segment suffices, and the displayed bound remains valid. Finally, $D \ge 2eK_0$ implies $2BT/D \le BT/(eK_0)$, and (137) implies that $(D/(2eK_0))^\alpha$ exceeds the logarithm of the union factor divided by p_{rad}. This proves the final assertion. □

Lemma A.25 (Convenient uniform high-probability bound)**.** *Assume assumptions A.1, A.3 and A.4 and the conditional product law of lemma A.19. Choose the population, radius, and concentration failure budgets each as $p/3$. Keep the deterministic schedule size explicit and set*

$$\Xi = \frac{3N(e+T)(L_{\mathrm{jump}}+1)(1+B)}{p}.$$

Suppose the population localization lemma gives

$$\mathbb{P}(E_{\mathrm{pop}}^c) \le p/3, \qquad M_{\mathrm{pop}} \le C\Xi^{c_M}, \qquad K_{\mathrm{pop}} \le C(1+\log\Xi)^{b_K},$$

where C, c_M, b_K and K_0, c_0, α are fixed. For every $\varepsilon > 0$, one can choose C_D, C_ε depending only on the displayed structural constants and on ε such that, with

$$D = C_D(1+\log\Xi)^{1/\alpha}, \tag{138}$$

the following holds for all sufficiently large N:

$$\mathbb{P}\left(E_{\mathrm{pop}} \cap E_{T,D} \cap \left\{\sup_{t\le T} \Gamma_t^N(D, K_{\mathrm{pop}}) \le C_\varepsilon \Xi^\varepsilon \left(N^{-1/2} + \frac{BT}{N}\right)\right\}\right) \ge 1-p.$$

For a uniformly bounded B, BT/N may be replaced by T/N after changing the constant. If L_{jump} is also uniformly bounded, the factor Ξ^ε is equivalent, up to constants, to $\{N(e+T)/p\}^\varepsilon$.

Proof. **Step 1: choose the radius and control the bias.** Increasing C_D first makes (137) hold with $p_{\mathrm{rad}} = p/3$. By (135),

$$A(D, K_{\mathrm{pop}}) \vee \Lambda(D, K_{\mathrm{pop}}) \vee S(D, K_{\mathrm{pop}}) \le C\exp\{C(1+\log\log(e+\Xi))^q\}.$$

Consequently each of these quantities is bounded by $C_\eta \Xi^\eta$ for every fixed $\eta > 0$. The bias formula (126) gives

$$\overline{\beta}_T(D) \le CS(D, K_{\mathrm{pop}})\Xi^{c_M/2}\exp\left[-\frac{c_0}{2}\left(\frac{C_D}{K_0}\right)^\alpha (1+\log\Xi)\right].$$

Choose C_D large enough that $(c_0/2)(C_D/K_0)^\alpha \ge c_M/2 + 3$. Use the subpolynomial bound with exponent 1, so $S(D, K_{\mathrm{pop}}) \le C\Xi$. It follows that

$$\overline{\beta}_T(D) \le C\Xi^{1+c_M/2-(c_M/2+3)} = C\Xi^{-2} \le N^{-1}$$

for sufficiently large N, since $\Xi \ge 3eN$.

Step 2: absorb the concentration logarithm and combine budgets. Finally, $Q_N/(p/3) \le C\Xi$ and $\sqrt{\log(C\Xi)} \le C_\varepsilon \Xi^{\varepsilon/2}$. Use (135) with exponent $\varepsilon/2$ for A and Λ. Substitution in lemma A.21, (127), gives

$$\begin{aligned}\delta_\Gamma(D) &\le N^{-1} + \frac{C_\varepsilon \Xi^{\varepsilon/2}}{\sqrt{N}}(1+\Xi^{\varepsilon/2}) + \frac{C_\varepsilon BT\Xi^{\varepsilon/2}}{N} \\ &\le C_\varepsilon \Xi^\varepsilon (N^{-1/2} + BT/N).\end{aligned}$$

The population event, the radius bound (137), and concentration (128) each cost $p/3$. Their sum is p. □

A.5.3 Weighted-factorization control of $\widetilde{\Gamma}_t^N$

We bound the full jump covariance residual directly in a Hilbert space of weighted features. The proof has three steps: factor the Hessian kernel, concentrate the feature covariance in Hilbert–Schmidt norm, and convert that estimate to the nuclear norm of the finite jump covariance matrix.

Throughout this subsection, fix a population realization in E_{pop} from lemma A.20, and write

$$\mathbb{E}_{\mathcal{G}}[Z] := \mathbb{E}[Z \mid \mathcal{G}_{\mathrm{pop}}], \qquad \mathbb{P}_{\mathcal{G}}(A) := \mathbb{P}(A \mid \mathcal{G}_{\mathrm{pop}}).$$

The conditional product law in lemma A.19 and the conditional moment bound (123) give, at every fixed time and its relevant trace,

$$\mathrm{Law}((Y_t^1, \dots, Y_t^N) \mid \mathcal{G}_{\mathrm{pop}}) = \mu_t^{\otimes N}, \qquad (\mathbb{E}_{\mathcal{G}}\|Y_t^1\|_{\mathbb{R}^d}^r)^{1/r} \le K_{\mathrm{pop}} r^{1/\alpha} \quad (r \ge 1). \tag{139}$$

Here K_{pop} is the deterministic envelope chosen for this population event; it may depend on the horizon and its probability budget. We never condition further on the sample localization event.

For the fixed deterministic radius $D \ge 1$, recall $\mathrm{L}_D = \mathrm{L}_{K_{\mathrm{pop}},D}$ from definition A.4. The population moment belongs to $\mathcal{V}_{K_{\mathrm{pop}},D}$ on E_{pop} by lemma A.3. Thus the moment-Hessian bound (39) is available here without a bound on the realized sample positions. We retain L_D^2 explicitly.

Weighted features and the integral Hessian kernel. Recall $a_m = e^{-c_a m^{\kappa_{\mathrm{dec}}}}$ and $\|h\|_{\mathsf{H}_a}^2 = \sum_m a_m^2 \|h_m\|_{\mathsf{H}_m}^2$ from assumption A.1. Also recall the feature map $J_0 : \mathbb{R}^d \to \mathsf{H}_a$, $J_0(x) = (\phi_m(x))_{m\geq 1}$, from lemma A.5, with $\mathrm{D}J_0(x)b = (\mathrm{D}\phi_m(x)b)_{m\geq 1}$. Define

$$\mathsf{E} := \bigoplus_{m\geq 1} \mathsf{H}_m, \qquad \mathsf{U}_a : \mathsf{H}_a \longrightarrow \mathsf{E}, \quad (\mathsf{U}_a h)_m = a_m h_m.$$

The direct sum has its unweighted Hilbert norm, so $\|\mathsf{U}_a h\|_{\mathsf{E}} = \|h\|_{\mathsf{H}_a}$, and $(\mathsf{U}_a^{-1}u)_m = a_m^{-1}u_m$ satisfies $\|\mathsf{U}_a^{-1}u\|_{\mathsf{H}_a}^2 = \sum_m \|u_m\|_{\mathsf{H}_m}^2$. Thus U_a is an isometric isomorphism. Express the derivative of J_0 in these unweighted coordinates by defining

$$\begin{aligned} \Psi_\infty(x) &:= \mathsf{U}_a \circ \mathrm{D}J_0(x) : \mathbb{R}^d \to \mathsf{E}, \\ \Psi_\infty(x)b &= \mathsf{U}_a((\mathrm{D}\phi_m(x)b)_m) = (a_m \mathrm{D}\phi_m(x)b)_m. \end{aligned} \tag{140}$$

This map is Hilbert–Schmidt:

$$\|\Psi_\infty(x)\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}^2 = \sum_{m\geq 1} a_m^2 \|\mathrm{D}\phi_m(x)\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{H}_m}^2 \leq d \sum_{m\geq 1} a_m^2 L_m^2 (1 + \|x\|_{\mathbb{R}^d}^{m-1})^2 < \infty. \tag{141}$$

The factor d follows by summing the squared operator bound over an orthonormal basis of $\mathbb{R}^d$. Finiteness follows by squaring lemma A.1, (26), with $K = 1$ and $D = 1 + \|x\|_{\mathbb{R}^d}$, and summing.

The following factorization uses the full feature space E.

Lemma A.26 (Weighted Hessian factorization)**.** *For $M(\mu) \in \mathcal{V}_{K_{\mathrm{pop}},D}$, there is a bounded self-adjoint operator $\overline{A}_\infty(\mu)$ on E such that*

$$\langle u, \overline{A}_\infty(\mu)v\rangle = \mathrm{D}^2 G(M(\mu))[\mathsf{U}_a^{-1}u, \mathsf{U}_a^{-1}v], \tag{142}$$

$$\nabla_W^2 F(\mu; x, z) = \Psi_\infty(x)^* \overline{A}_\infty(\mu)\Psi_\infty(z), \qquad \|\overline{A}_\infty(\mu)\|_{\mathsf{E}\to\mathsf{E}} \leq \mathrm{L}_D. \tag{143}$$

Proof. **Step 1: represent the bounded moment bilinear form.** By the isometry U_a and (39),

$$|\mathrm{D}^2 G(M(\mu))[\mathsf{U}_a^{-1}u, \mathsf{U}_a^{-1}v]| \leq \mathrm{L}_D \|u\|_{\mathsf{E}} \|v\|_{\mathsf{E}}.$$

For fixed $v \in \mathsf{E}$, the map $u \mapsto \mathrm{D}^2 G(M(\mu))[\mathsf{U}_a^{-1}u, \mathsf{U}_a^{-1}v]$ is a continuous linear functional of norm at most $\mathrm{L}_D\|v\|_{\mathsf{E}}$. Riesz representation gives its unique representing vector $\overline{A}_\infty(\mu)v$. Uniqueness and linearity in v make this a linear operator, with

$$\|\overline{A}_\infty(\mu)v\|_{\mathsf{E}} = \sup_{\|u\|_{\mathsf{E}}=1} |\mathrm{D}^2 G(M(\mu))[\mathsf{U}_a^{-1}u, \mathsf{U}_a^{-1}v]| \leq \mathrm{L}_D\|v\|_{\mathsf{E}}.$$

Symmetry of the second derivative gives $\langle u, \overline{A}_\infty(\mu)v\rangle = \langle \overline{A}_\infty(\mu)u, v\rangle$, so this bounded operator is self-adjoint.

Step 2: identify the spatial Hessian kernel. For $b, c \in \mathbb{R}^d$, eq. (140) gives $\mathsf{U}_a^{-1}\Psi_\infty(x)b = \mathrm{D}J_0(x)b$ and the corresponding identity at z. Substituting these identities into eq. (54) and using eq. (142) gives

$$\begin{aligned} \langle b, \nabla_W^2 F(\mu; x, z)c\rangle_{\mathbb{R}^d} &= \mathrm{D}^2 G(M(\mu))[\mathrm{D}J_0(x)b, \mathrm{D}J_0(z)c] \\ &= \mathrm{D}^2 G(M(\mu))[\mathsf{U}_a^{-1}\Psi_\infty(x)b, \mathsf{U}_a^{-1}\Psi_\infty(z)c] \\ &= \langle \Psi_\infty(x)b, \overline{A}_\infty(\mu)\Psi_\infty(z)c\rangle_{\mathsf{E}}. \end{aligned}$$

Both argument sequences belong to H_a by (141); finite partial sums converge there, so the displayed equality also identifies the infinite coordinate series. □

Lemma A.27 (Weighted feature bounds with explicit conditional moment dependence)**.** *Assume* (139). *Recall $\chi(u) = [\log(e+u)]^{\kappa_{\mathrm{dec}}/(\kappa_{\mathrm{dec}}-1)}$ from* (25)*, and define the single fourth-moment envelope*

$$C_{\Phi,4}(K) := \left[d \sum_{m\geq 1} a_m^2 L_m^2 \left\{ 1 + K^{m-1}(4(m-1)\vee 1)^{(m-1)/\alpha} \right\}^2 \right]^2. \tag{144}$$

All constants C below depend only on the fixed feature parameters, d, and α, not on $K_{\mathrm{pop}}, N, t, D$ or any probability budget. The following assertions hold.

1. *The full feature map satisfies*

$$\|\Psi_\infty(x)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le C\exp\{C\chi(\|x\|_{\mathbb{R}^d})\}. \tag{145}$$

The feature map is continuous in Hilbert–Schmidt norm and strongly measurable.

2. *Uniformly in t,*

$$\mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\|^4_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le C_{\Phi,4}(K_{\mathrm{pop}}) \le Ce^{C\chi(K_{\mathrm{pop}})}. \tag{146}$$

3. *The Bochner integral $\Sigma_t := \mathbb{E}_{\mathcal{G}}[\Psi_\infty(Y_t^1)\Psi_\infty(Y_t^1)^*]$ is a positive trace-class operator, and*

$$\Sigma_t \preceq C_{\Phi,4}(K_{\mathrm{pop}})^{1/2} I_{\mathsf{E}}. \tag{147}$$

Proof. **Step 1: continuity and the pointwise bound.** For $R \ge 1$ and $\|x\|_{\mathbb{R}^d}, \|y\|_{\mathbb{R}^d} \le R$, integrate $\mathrm{D}^2\phi_m$ on the line segment from y to x. The derivative-growth assumption gives

$$\|\mathrm{D}\phi_m(x) - \mathrm{D}\phi_m(y)\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{H}_m} \le \sqrt{d}\, L_m(1 + R^{(m-2)_+})\|x-y\|_{\mathbb{R}^d}.$$

After summing squares with weights a_m^2, this yields

$$\|\Psi_\infty(x) - \Psi_\infty(y)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le d\|x-y\|^2_{\mathbb{R}^d}\sum_{m\ge1} a_m^2 L_m^2(1+R^{(m-2)_+})^2.$$

The coefficient is finite by squaring (26) with $K=1$, $D=R$ and summing. This proves continuity on every ball. Its image lies in a separable closed subspace, since $\mathbb{R}^d$ is separable and the map is continuous. Strong measurability means being an almost-sure norm limit of measurable simple functions. A continuous map from the separable space $\mathbb{R}^d$ has separable image, so this property holds for $\Psi_\infty(Y_t^1)$ without assuming that the whole feature space is separable. For the pointwise bound, $1 + \|x\|_{\mathbb{R}^d}^{m-1} \le 2 + (1+\|x\|_{\mathbb{R}^d})^m$. Apply lemma A.1, (26), with $K=1$ and $D = 1 + \|x\|_{\mathbb{R}^d}$:

$$\|\Psi_\infty(x)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le d\sum_m [a_m L_m\{2+(1+\|x\|_{\mathbb{R}^d})^m\}]^2 \le Ce^{C\chi(\|x\|_{\mathbb{R}^d})}\sum_m e^{-(c_a/2)m^{\kappa_{\mathrm{dec}}}}. \tag{148}$$

Here (28) with $c=1$ absorbs $\chi(2+\|x\|_{\mathbb{R}^d})$. The finite sum proves (145).

Step 2: the fourth moment. For $m \ge 2$, (139) with $r = 4(m-1)$ gives

$$\|1 + \|Y_t^1\|_{\mathbb{R}^d}^{m-1}\|_{L^4(\mathbb{P}_{\mathcal{G}})} \le 1 + (\mathbb{E}_{\mathcal{G}}\|Y_t^1\|_{\mathbb{R}^d}^{4(m-1)})^{1/4} \le 1 + K_{\mathrm{pop}}^{m-1}(4(m-1))^{(m-1)/\alpha}.$$

For $m=1$ the left side is 2, matching the convention in (144). Minkowski's inequality for a finite sum in $L^2(\mathbb{P}_{\mathcal{G}})$ now gives

$$\begin{aligned}\left\|\sum_{m=1}^M a_m^2\|\mathrm{D}\phi_m(Y_t^1)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{H}_m}\right\|_{L^2(\mathbb{P}_{\mathcal{G}})} &\le d\sum_{m\le M} a_m^2 L_m^2\|1+\|Y_t^1\|_{\mathbb{R}^d}^{m-1}\|^2_{L^4(\mathbb{P}_{\mathcal{G}})}\\ &\le d\sum_{m\le M} a_m^2 L_m^2\{1 + K_{\mathrm{pop}}^{m-1}(4(m-1)\vee 1)^{(m-1)/\alpha}\}^2.\end{aligned}$$

Squaring and applying monotone convergence gives the explicit connection between the finite sums and the full fourth moment:

$$\begin{aligned}\mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\|^4_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} &= \lim_{M\to\infty}\mathbb{E}_{\mathcal{G}}\left[\sum_{m=1}^M a_m^2\|\mathrm{D}\phi_m(Y_t^1)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{H}_m}\right]^2\\ &\le C_{\Phi,4}(K_{\mathrm{pop}}).\end{aligned} \tag{149}$$

For every $m \ge 1$, including $m=1$, the braces in (144) satisfy

$$1 + K^{m-1}(4(m-1)\vee 1)^{(m-1)/\alpha} \le 1 + \{4^{1/\alpha}(1+K)\}^m m^{m/\alpha}.$$

Use (26) with K replaced by $4^{1/\alpha}(1+K_{\mathrm{pop}})$ and $D=1$. Squaring each summand, summing, and then squaring the sum in (144) yields

$$C_{\Phi,4}(K_{\mathrm{pop}}) \le \left[Ce^{C\chi(K_{\mathrm{pop}})} \sum_m e^{-(c_a/2)m^{\kappa_{\mathrm{dec}}}} \right]^2 \le Ce^{C\chi(K_{\mathrm{pop}})},$$

where (28) absorbs the fixed rescaling. This proves (146).

Step 3: the covariance. For Hilbert–Schmidt maps $A, B : \mathbb{R}^d \to \mathsf{E}$, its singular values s_j give

$$\mathrm{tr}(AA^*) = \sum_j s_j^2 = \|A\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}, \qquad \|AA^*\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} = (\sum_j s_j^4)^{1/2} \le \sum_j s_j^2.$$

Moreover, $AA^* - BB^* = (A-B)A^* + B(A-B)^*$ gives

$$\|AA^* - BB^*\|_{*;\,\mathsf{E}\to\mathsf{E}} \le (\|A\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} + \|B\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}})\|A-B\|_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}.$$

Thus $A \mapsto AA^*$ is continuous from Hilbert–Schmidt norm to trace norm, using (116). Composition with the strongly measurable feature map from Step 1 therefore gives strong measurability in trace norm. For integrability, positivity of $\Psi_\infty(Y_t^1)\Psi_\infty(Y_t^1)^*$ and Cauchy–Schwarz give

$$\begin{aligned}\mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\Psi_\infty(Y_t^1)^*\|_{*;\,\mathsf{E}\to\mathsf{E}} &= \mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \\ &\le \left(\mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\|^4_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}\right)^{1/2} \le C_{\Phi,4}(K_{\mathrm{pop}})^{1/2} < \infty.\end{aligned}$$

This establishes Bochner integrability in trace norm. For $u \in \mathsf{E}$, Cauchy–Schwarz gives

$$\langle u, \Sigma_t u\rangle = \mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)^* u\|^2_{\mathbb{R}^d} \le \|u\|^2_{\mathsf{E}}\mathbb{E}_{\mathcal{G}}\|\Psi_\infty(Y_t^1)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}\|u\|^2_{\mathsf{E}}.$$

□

The full residual as a feature covariance error. Recall the kernel covariance residual from lemma A.16:

$$U_t(z) := [\nabla^2_W F(\mu_t; Y_t^i, z)]_{i=1}^N, \qquad \widetilde{\Gamma}_t^N = \frac{1}{N}\left\| \int U_t(z)U_t(z)^\top(\mu_t - \rho_{Y_t})(dz) \right\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}. \tag{150}$$

The matrix $U_t(z)$ has N block rows and d columns. Define

$$\Psi_t^{Y,N}(v_1,\dots,v_N) := \sum_{i=1}^N \Psi_\infty(Y_t^i)v_i, \qquad \Psi_t^{Y,N} : \mathbb{R}^{dN} \longrightarrow \mathsf{E},$$

where the domain has the ordinary Euclidean inner product. Recall Σ_t from lemma A.27 and define

$$\widehat{\Sigma}_t^N := \frac{1}{N}\sum_{i=1}^N \Psi_\infty(Y_t^i)\Psi_\infty(Y_t^i)^* = \frac{1}{N}\Psi_t^{Y,N}(\Psi_t^{Y,N})^*.$$

Lemma A.28 (Full covariance representation and nuclear-norm bound)**.** *The full residual satisfies*

$$\widetilde{\Gamma}_t^N = \frac{1}{N}\left\|(\Psi_t^{Y,N})^*\overline{A}_\infty(\mu_t)(\Sigma_t - \widehat{\Sigma}_t^N)\overline{A}_\infty(\mu_t)\Psi_t^{Y,N}\right\|_{*;\,\mathbb{R}^{dN}\to\mathbb{R}^{dN}}, \tag{151}$$

$$\widetilde{\Gamma}_t^N \le \mathrm{L}_D^2\,\mathrm{tr}(\widehat{\Sigma}_t^N)\|\widehat{\Sigma}_t^N - \Sigma_t\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}}. \tag{152}$$

Consequently,

$$\mathbb{E}_{\mathcal{G}}\widetilde{\Gamma}_t^N \le \frac{\mathrm{L}_D^2 C_{\Phi,4}(K_{\mathrm{pop}})}{\sqrt{N}}. \tag{153}$$

Proof. **Step 1: factor and bound the covariance residual.** By (143),

$$U_t(z) = (\Psi_t^{Y,N})^*\overline{A}_\infty(\mu_t)\Psi_\infty(z).$$

Since $\overline{A}_\infty(\mu_t)$ is self-adjoint, the product is

$$U_t(z)U_t(z)^\top = (\Psi_t^{Y,N})^* \overline{A}_\infty(\mu_t)[\Psi_\infty(z)\Psi_\infty(z)^*]\overline{A}_\infty(\mu_t)\Psi_t^{Y,N}.$$

Only the bracketed factor depends on z. Integration of this factor $\Psi_\infty(z)\Psi_\infty(z)^*$ against μ_t gives Σ_t, and integration against $\rho_{Y_t} = N^{-1}\sum_i \delta_{Y_t^i}$ gives $\widehat{\Sigma}_t^N$. Subtracting proves (151). Apply the Hilbert–Schmidt product bound (116) to the two factors $(\Psi_t^{Y,N})^*\overline{A}_\infty(\mu_t)$ and $(\Sigma_t - \widehat{\Sigma}_t^N)\overline{A}_\infty(\mu_t)\Psi_t^{Y,N}$. Then use submultiplicativity of the Hilbert–Schmidt norm with the operator norm, also after taking adjoints:

$$\begin{aligned}
\widetilde{\Gamma}_t^N &\le \frac{1}{N}\|(\Psi_t^{Y,N})^*\overline{A}_\infty(\mu_t)\|_{\mathrm{HS};\,\mathsf{E}\to\mathbb{R}^{dN}}\|(\Sigma_t - \widehat{\Sigma}_t^N)\overline{A}_\infty(\mu_t)\Psi_t^{Y,N}\|_{\mathrm{HS};\,\mathbb{R}^{dN}\to\mathsf{E}} \\
&\le \frac{1}{N}\|\Psi_t^{Y,N}\|^2_{\mathrm{HS};\,\mathbb{R}^{dN}\to\mathsf{E}}\|\overline{A}_\infty(\mu_t)\|^2_{\mathsf{E}\to\mathsf{E}}\|\Sigma_t - \widehat{\Sigma}_t^N\|_{\mathsf{E}\to\mathsf{E}} \\
&\le \mathrm{L}_D^2\,\mathrm{tr}(\widehat{\Sigma}_t^N)\|\Sigma_t - \widehat{\Sigma}_t^N\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}}.
\end{aligned}$$

Here $N^{-1}\|\Psi_t^{Y,N}\|^2_{\mathrm{HS};\,\mathbb{R}^{dN}\to\mathsf{E}} = N^{-1}\sum_i \|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} = \mathrm{tr}(\widehat{\Sigma}_t^N)$.

Step 2: bound its conditional expectation. For the expectation bound put $W_i = \Psi_\infty(Y_t^i)\Psi_\infty(Y_t^i)^*$ and $Z_i = \|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}$. Jensen's inequality gives

$$\mathbb{E}_{\mathcal{G}}\left(\mathrm{tr}\,\widehat{\Sigma}_t^N\right)^2 = \mathbb{E}_{\mathcal{G}}(N^{-1}\sum_i Z_i)^2 \le N^{-1}\sum_i \mathbb{E}_{\mathcal{G}} Z_i^2 \le C_{\Phi,4}(K_{\mathrm{pop}}).$$

In the Hilbert space $\mathcal{S}_2(\mathsf{E})$, conditional independence and $\mathbb{E}_{\mathcal{G}}(W_i - \Sigma_t) = 0$ imply, for $i \ne j$,

$$\mathbb{E}_{\mathcal{G}}\langle W_i - \Sigma_t, W_j - \Sigma_t\rangle_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} = \langle \mathbb{E}_{\mathcal{G}}(W_i - \Sigma_t), \mathbb{E}_{\mathcal{G}}(W_j - \Sigma_t)\rangle_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} = 0.$$

In particular, the singular-value calculation in lemma A.27 gives $\|W_1\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \le (\mathrm{tr}\,W_1)^2 = Z_1^2$, with an inequality rather than equality in general. Expanding gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}}\|\widehat{\Sigma}_t^N - \Sigma_t\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} &= N^{-2}\sum_{i,j}\mathbb{E}_{\mathcal{G}}\langle W_i - \Sigma_t, W_j - \Sigma_t\rangle_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \\
&= N^{-2}\sum_i \mathbb{E}_{\mathcal{G}}\|W_i - \Sigma_t\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \\
&= N^{-1}(\mathbb{E}_{\mathcal{G}}\|W_1\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} - \|\Sigma_t\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}}) \\
&\le N^{-1}\mathbb{E}_{\mathcal{G}} Z_1^2 \le C_{\Phi,4}(K_{\mathrm{pop}})/N.
\end{aligned} \tag{154}$$

Conditional Cauchy–Schwarz in (152) now gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}}\widetilde{\Gamma}_t^N &\le \mathrm{L}_D^2[\mathbb{E}_{\mathcal{G}}(\mathrm{tr}\,\widehat{\Sigma}_t^N)^2]^{1/2}[\mathbb{E}_{\mathcal{G}}\|\widehat{\Sigma}_t^N - \Sigma_t\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}}]^{1/2} \\
&\le \mathrm{L}_D^2 C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}[C_{\Phi,4}(K_{\mathrm{pop}})/N]^{1/2}.
\end{aligned}$$

The two factors use the same sample; their independence is not needed. □

Lemma A.29 (A deterministic envelope at the sampled particles)**.** *Fix a time trace t, $N \ge 2$, and $\delta \in (0,1)$. Define*

$$R_{N,\delta} := K_{\mathrm{pop}}[e(\alpha \vee 1)\{1 + \log(8N/\delta)\}]^{1/\alpha},$$

$$\tau_{N,\delta} := 1 \vee \left[d\sum_{m\ge 1} a_m^2 L_m^2 (1 + R_{N,\delta}^{m-1})^2\right]. \tag{155}$$

Then

$$\mathbb{P}_{\mathcal{G}}\left(\max_{i\le N}\|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} > \tau_{N,\delta}\right) \le \delta/8, \tag{156}$$

and the individual-particle probability is at most $\delta/(8N)$. Moreover $\tau_{N,\delta} \le C\exp\{C\chi(R_{N,\delta})\}$. If $\Xi \ge N/\delta$ and $K_{\mathrm{pop}} \le C_K(\log\Xi)^{s_K}$ for fixed C_K, s_K, then, for every $\varepsilon > 0$,

$$\tau_{N,\delta}^2\sqrt{\log(8/\delta)} + C_{\Phi,4}(K_{\mathrm{pop}}) \le C_\varepsilon \Xi^\varepsilon. \tag{157}$$

Proof. Use the Markov calculation of lemma A.22, (133), with threshold $R_{N,\delta}$. In the moment input (139), take $r = (\alpha \vee 1)\{1 + \log(8N/\delta)\} \geq 1$. Then

$$\mathbb{P}_{\mathcal{G}}(\|Y_t^i\|_{\mathbb{R}^d} > R_{N,\delta}) \leq \left(\frac{K_{\mathrm{pop}} r^{1/\alpha}}{K_{\mathrm{pop}}(er)^{1/\alpha}}\right)^r = e^{-r/\alpha} \leq \frac{\delta}{8N}.$$

By (141) and (155),

$$\{\|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} > \tau_{N,\delta}\} \subseteq \{\|Y_t^i\|_{\mathbb{R}^d} > R_{N,\delta}\}.$$

The union over N particles costs $N\delta/(8N) = \delta/8$, proving (156). The calculation (148) in lemma A.27, with $\|x\|_{\mathbb{R}^d}$ replaced by $R_{N,\delta}$, proves the growth bound. Since $\Xi \geq N/\delta \geq 2$, $\log(8N/\delta) \leq \log 8 + \log \Xi \leq 4 \log \Xi$. Hence

$$R_{N,\delta} \leq C(\log \Xi)^{s_K + 1/\alpha}, \qquad \log(e + R_{N,\delta}) \leq C + C \log\log \Xi.$$

For the whole left-hand side of (157), these bounds and (146) give

$$\begin{aligned}&\log\{\tau_{N,\delta}^2 \sqrt{\log(8/\delta)} + C_{\Phi,4}(K_{\mathrm{pop}})\} \\ &\quad \leq C + C[1 + \log\log(e + \Xi)]^{\kappa_{\mathrm{dec}}/(\kappa_{\mathrm{dec}}-1)} \leq C_\varepsilon + \varepsilon \log \Xi.\end{aligned}$$

The last inequality holds because any fixed power of $\log\log(e + \Xi)$ is $o(\log \Xi)$; a fixed constant covers bounded $\Xi \geq 2$. Exponentiation proves the claim. The constant may depend on C_K, s_K and ε, but not on Ξ. □

Lemma A.30 (Full weighted covariance concentration in Hilbert–Schmidt norm)**.** *Fix a time trace t, $N \geq 2$, and $\delta \in (0,1)$. Use the threshold $\tau_{N,\delta}$ in* (155)*. There is an event of conditional probability at least $1 - \delta$ on which*

$$\max_{i \leq N} \|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \leq \tau_{N,\delta}, \tag{158}$$

$$\|\widehat{\Sigma}_t^N - \Sigma_t\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \leq \frac{C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}}{\sqrt{N}}(1 + \sqrt{\delta/8}) + \tau_{N,\delta}\sqrt{\frac{2\log(8/\delta)}{N}}. \tag{159}$$

Proof. The repeated random objects needed in this proof are

$$W_i := \Psi_\infty(Y_t^i)\Psi_\infty(Y_t^i)^*, \quad Z_i := \|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}}, \quad W_i^{(\tau)} := W_i \mathbf{1}_{\{Z_i \leq \tau_{N,\delta}\}}.$$

The conditional mean of $W_i^{(\tau)}$ is $\Sigma_t^{(\tau)} := \mathbb{E}_{\mathcal{G}} W_1^{(\tau)}$. The sample mean of $W_i^{(\tau)}$ is $\widehat{\Sigma}_t^{N,(\tau)} := N^{-1}\sum_i W_i^{(\tau)}$. Conditional on $\mathcal{G}_{\mathrm{pop}}$, these clipped variables remain independent. We apply concentration before intersecting with $\{\max_i Z_i \leq \tau_{N,\delta}\}$.

Step 1: mean square and bounded differences. The singular-value inequality in lemma A.27 gives $\|W_i^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \leq Z_i \mathbf{1}_{\{Z_i \leq \tau_{N,\delta}\}} \leq \tau_{N,\delta}$. These operators are conditionally i.i.d. and square integrable. Apply the centered-sum calculation (154) of lemma A.28, replacing W_i, Σ_t by $W_i^{(\tau)}, \Sigma_t^{(\tau)}$:

$$\mathbb{E}_{\mathcal{G}}\|\widehat{\Sigma}_t^{N,(\tau)} - \Sigma_t^{(\tau)}\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \leq N^{-1}\mathbb{E}_{\mathcal{G}}\|W_1^{(\tau)}\|^2_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \leq N^{-1}\mathbb{E}_{\mathcal{G}} Z_1^2 \leq C_{\Phi,4}(K_{\mathrm{pop}})/N.$$

Cauchy–Schwarz therefore bounds the mean norm by $\sqrt{C_{\Phi,4}(K_{\mathrm{pop}})/N}$. If sample i is replaced, denote the new sample mean by $\widehat{\Sigma}_t^{N,(\tau)\prime}$ and its changed summand by $W_i^{(\tau)\prime}$. The threshold and population mean are fixed, so

$$\begin{aligned}&\left|\|\widehat{\Sigma}_t^{N,(\tau)} - \Sigma_t^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} - \|\widehat{\Sigma}_t^{N,(\tau)\prime} - \Sigma_t^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}}\right| \\ &\quad \leq \|\widehat{\Sigma}_t^{N,(\tau)} - \widehat{\Sigma}_t^{N,(\tau)\prime}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} = N^{-1}\|W_i^{(\tau)} - W_i^{(\tau)\prime}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \leq 2\tau_{N,\delta}/N.\end{aligned}$$

In lemma A.18, the sum of squared differences is $N(2\tau_{N,\delta}/N)^2 = 4\tau_{N,\delta}^2/N$. Substituting $s = \tau_{N,\delta}\sqrt{2\log(8/\delta)/N}$ gives failure probability

$$\exp\left(-\frac{2s^2}{4\tau_{N,\delta}^2/N}\right) = \exp(-\log(8/\delta)) = \delta/8.$$

Thus, except on that event,

$$\|\widehat{\Sigma}_t^{N,(\tau)} - \Sigma_t^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \le \sqrt{C_{\Phi,4}(K_{\mathrm{pop}})/N} + \tau_{N,\delta}\sqrt{2\log(8/\delta)/N}.$$

Step 2: restore the conditional mean and the sample. The mean difference is $\Sigma_t - \Sigma_t^{(\tau)} = \mathbb{E}_{\mathcal{G}}[W_1 \mathbf{1}_{\{Z_1 > \tau_{N,\delta}\}}]$. The individual tail bound in lemma A.29 and Cauchy–Schwarz give

$$\begin{aligned}\|\Sigma_t - \Sigma_t^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} &\le \mathbb{E}_{\mathcal{G}}[\|W_1\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \mathbf{1}_{\{Z_1 > \tau_{N,\delta}\}}] \\ &\le (\mathbb{E}_{\mathcal{G}} Z_1^2)^{1/2} \mathbb{P}_{\mathcal{G}}(Z_1 > \tau_{N,\delta})^{1/2} \le C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}\sqrt{\delta/(8N)}.\end{aligned}$$

The bound (156) in lemma A.29 costs another $\delta/8$ to ensure $\max_i Z_i \le \tau_{N,\delta}$ and hence $\widehat{\Sigma}_t^N = \widehat{\Sigma}_t^{N,(\tau)}$. On the intersection,

$$\begin{aligned}\|\widehat{\Sigma}_t^N - \Sigma_t\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} &\le \|\widehat{\Sigma}_t^{N,(\tau)} - \Sigma_t^{(\tau)}\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} + \|\Sigma_t^{(\tau)} - \Sigma_t\|_{\mathrm{HS};\,\mathsf{E}\to\mathsf{E}} \\ &\le \frac{C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}}{\sqrt{N}}(1 + \sqrt{\delta/8}) + \tau_{N,\delta}\sqrt{\frac{2\log(8/\delta)}{N}}.\end{aligned}$$

The population bias bound is deterministic after conditioning and costs no further probability. Thus the failure budget is $\delta/8 + \delta/8 \le \delta$, proving (159). □

Corollary A.2 (The full jump residual at a fixed time)**.** *For every fixed time trace t, $N \ge 2$, and $\delta \in (0,1)$, with conditional probability at least $1-\delta$,*

$$\widetilde{\Gamma}_t^N \le \frac{\mathrm{L}_D^2 \tau_{N,\delta}}{\sqrt{N}}\left[2C_{\Phi,4}(K_{\mathrm{pop}})^{1/2} + \tau_{N,\delta}\sqrt{2\log(8/\delta)}\right]. \tag{160}$$

Proof. On the joint event of lemma A.30,

$$\mathrm{tr}(\widehat{\Sigma}_t^N) = N^{-1}\sum_i \|\Psi_\infty(Y_t^i)\|^2_{\mathrm{HS};\,\mathbb{R}^d\to\mathsf{E}} \le \tau_{N,\delta}.$$

Substitute this and (159) into lemma A.28, (152):

$$\widetilde{\Gamma}_t^N \le \mathrm{L}_D^2 \tau_{N,\delta}\left[\frac{C_{\Phi,4}(K_{\mathrm{pop}})^{1/2}}{\sqrt{N}}(1 + \sqrt{\delta/8}) + \tau_{N,\delta}\sqrt{\frac{2\log(8/\delta)}{N}}\right].$$

Use $1 + \sqrt{\delta/8} \le 2$ to obtain (160) on the same event. □

Corollary A.3 (Union bound over prescribed jump times)**.** *Let $0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \le T$ be deterministic, and set $q = p/(L_{\mathrm{jump}} + 1)$ for $p \in (0,1)$. Define*

$$\widetilde{\delta}^{\mathrm{HS}}_{N,T,p} := \frac{\mathrm{L}_D^2 \tau_{N,q}}{\sqrt{N}}\left[2C_{\Phi,4}(K_{\mathrm{pop}})^{1/2} + \tau_{N,q}\sqrt{2\log(8/q)}\right]. \tag{161}$$

Then, with conditional probability at least $1-p$,

$$\max_{1\le\ell\le L_{\mathrm{jump}}} \widetilde{\Gamma}^N_{t_\ell -} \le \widetilde{\delta}^{\mathrm{HS}}_{N,T,p}. \tag{162}$$

The maximum is zero when $L_{\mathrm{jump}} = 0$. The parameters K_{pop} and L_D retain their horizon and population-confidence dependence. If K_{pop}, D grow at most as fixed powers of $\log \Xi$, $\Xi \ge N(L_{\mathrm{jump}} + 1)/p$, and the structural parameters are fixed, then for every $\varepsilon > 0$,

$$\widetilde{\delta}^{\mathrm{HS}}_{N,T,p} \le C_\varepsilon N^{-1/2}\Xi^\varepsilon. \tag{163}$$

In particular the residual has rate $N^{-1/2+o(1)}$ when Ξ is polynomial in N; its square root in the jump comparison has rate $N^{-1/4+o(1)}$.

Proof. **Step 1: combine the prescribed jump times.** Apply corollary A.2 at each deterministic pre-jump trace with $\delta = q = p/(L_{\mathrm{jump}}+1)$ in (160). The union costs $L_{\mathrm{jump}}q = pL_{\mathrm{jump}}/(L_{\mathrm{jump}}+1) \le p$; different times need not be independent.

Step 2: simplify the deterministic coefficient. For the rate, (38) bounds $\mathrm{L}_D \le Ce^{C\chi(K_{\mathrm{pop}}+D)}$. For polylogarithmic K_{pop}, D, this is at most $C_\eta \Xi^\eta$ for every $\eta > 0$, by the logarithm calculation proving (157) in lemma A.29. Apply (157) with the same η. Since $\log(8/q) > 1$, it implies

$$\tau_{N,q} \le C_\eta \Xi^{\eta/2}, \quad C_{\Phi,4}(K_{\mathrm{pop}})^{1/2} \le C_\eta \Xi^{\eta/2}, \quad \tau_{N,q}^2 \sqrt{\log(8/q)} \le C_\eta \Xi^\eta.$$

The bracket after multiplication by $\tau_{N,q}$ is therefore at most $C_\eta \Xi^\eta$. Together with $\mathrm{L}_D^2 \le C_\eta \Xi^{2\eta}$, this gives (163) by taking $\eta = \varepsilon/3$.

Step 3: average the conditional failure probability. Finally, since $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$,

$$\begin{aligned}&\mathbb{P}\left(E_{\mathrm{pop}} \cap \{\max_\ell \widetilde{\Gamma}^N_{t_\ell -} > \widetilde{\delta}^{\mathrm{HS}}_{N,T,p}\}\right)\\ &\quad = \mathbb{E}\left[\mathbf{1}_{E_{\mathrm{pop}}} \mathbb{P}_{\mathcal{G}}\left(\max_\ell \widetilde{\Gamma}^N_{t_\ell -} > \widetilde{\delta}^{\mathrm{HS}}_{N,T,p}\right)\right] \le p.\end{aligned}$$

Adding $\mathbb{P}(E_{\mathrm{pop}}^c)$ gives the unconditional failure budget when the population event is also required. □

Remark A.2 (Why no feature cutoff or nuclear-norm rank loss is needed)**.** *The finite $dN \times dN$ residual factors through the full Hilbert feature space E. The independent random operators are clipped by their norm before scalar bounded differences is applied in $\mathcal{S}_2(\mathsf{E})$. Neither argument uses the dimension of a feature space. The nuclear norm is bounded by the empirical feature trace times the Hilbert–Schmidt covariance error on one event. This proves the bound for $\widetilde{\Gamma}^N_t$ directly.*

A.6 Fixed-schedule comparison with explicit high-probability rates

We assemble the deterministic ODE–jump comparison and the two stochastic concentration bounds into the final fixed-schedule tracking theorem.

This subsection packages the detailed estimates from lemma A.12, lemma A.17, lemma A.31, lemma A.25, corollary A.3 into a single fixed-schedule statement.

We keep only the quantities that enter the final comparison bound.

- the pathwise comparison variable is
$$r_t = \frac{1}{\sqrt{N}} \|\boldsymbol{X}_t - \boldsymbol{Y}_t\|_2,$$
- the continuous stochastic residual is
$$\Gamma^N_t = \sum_{m \ge 1} \mathrm{L}_D a_m \left\| \frac{1}{N} \sum_{i=1}^N \phi_m(Y^i_t) - \int \phi_m(y)\, \mu_t(dy) \right\|_{\mathsf{H}_m},$$
as in (82),
- the jump defect is controlled by $\widetilde{\Gamma}^N_{t_\ell -}$ from (150) only at the prescribed jump times,
- curvature is retained through the propagator $G^Y_\kappa(t,s) = \exp\Big(-\int_s^t \kappa^{Y,N}_u\, du\Big)$, using the sampled full Hessian curvature in (11).

The first lemma solves the deterministic comparison inequality across ODE intervals and multiplicative jump errors.

Lemma A.31 (Piecewise Grönwall estimate with recursive jump amplification)**.** *Fix $T > 0$ and jump times*

$$0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \le T, \qquad t_0 := 0, \qquad t_{L_{\mathrm{jump}}+1} := T.$$

Assume that $r_t := \frac{1}{\sqrt{N}}\|\boldsymbol{X}_t - \boldsymbol{Y}_t\|_2$ *is absolutely continuous on each interval* $(t_\ell, t_{\ell+1})$, $\ell = 0, \dots, L_{\mathrm{jump}}$. *Assume right-continuity and absolute continuity up to each branch endpoint. Here* c_D *is a deterministic upper bound on the ODE quadratic coefficient, and* $q(p) = 1 + \sqrt{2\log(2/p)}$. *For a given pathwise radius* $R > 0$, *define*

$$\tau_R := \inf\{t \in [0,T] : r_t \geq R\},$$

with the convention $\inf \emptyset = \infty$.

Let

$$\beta_t := \kappa_t^{Y,N} - c_D R - \Gamma_t^N.$$

Assume the following.

1. *For every* $\ell = 0, \dots, L_{\mathrm{jump}}$ *and for a.e.* $t \in (t_\ell, t_{\ell+1}) \cap [0, \tau_R)$,
$$\dot{r}_t \leq -\beta_t r_t + \Gamma_t^N. \tag{164}$$
(corresponding to the conclusion in lemma A.12)

2. *For each jump time* $t_\ell \leq \tau_R$, *including a candidate first exit at that jump,* $\ell = 1, \dots, L_{\mathrm{jump}}$, *the previous jump lemma yields*
$$r_{t_\ell} \leq \big(1 + a_\ell\big) r_{t_\ell -} + b_\ell, \tag{165}$$
where
$$a_\ell(\mathrm{D}_{\sup}, p) := c_J(D)\eta_\ell q(p), \qquad b_\ell := a_\ell\Big(\sqrt{\Gamma^N_{t_\ell -}} + \sqrt{\widetilde{\Gamma}^N_{t_\ell -}}\Big).$$
(corresponding to lemma A.17 on the event $E_{T,\mathrm{D}_{\sup}}$*).*

Then, for every $t \in [0,T] \cap [0, \tau_R)$,

$$r_t \leq \int_0^t \exp\Big(-\int_s^t \beta_u \, du\Big) \prod_{\ell:\, s < t_\ell \leq t} (1 + a_\ell)\, \Gamma_s^N \, ds \tag{166}$$

$$+ \sum_{\ell:\, t_\ell \leq t} \exp\Big(-\int_{t_\ell}^t \beta_u \, du\Big) \prod_{j:\, t_\ell < t_j \leq t} (1 + a_j)\, b_\ell \tag{167}$$

if $r_0 = 0$. *The same bound holds at a finite candidate exit* τ_R, *using the pre-jump or post-jump trace as appropriate. No adaptedness of* R *is required.*

Proof. **Step 1: integrate within one ODE interval.** For $0 \leq s \leq t \leq T$, define

$$G(t,s) := \exp\Big(-\int_s^t \beta_u \, du\Big).$$

By the definition of β, this is $G(t,s) = G_\kappa^Y(t,s)\exp\{c_D R(t-s) + \int_s^t \Gamma_u^N \, du\}$, with the curvature propagator of definition 3.1. On each continuity interval $(t_\ell, t_{\ell+1})$, the differential inequality (164) implies

$$r_t \leq G(t,s)\, r_s + \int_s^t G(t,u)\Gamma_u^N \, du, \qquad t_\ell \leq s \leq t < t_{\ell+1}, \quad t < \tau_R. \tag{168}$$

Indeed,

$$\dot{r}_t + \beta_t r_t \leq \Gamma_t^N,$$

and multiplying by the integrating factor $\exp\Big(\int_s^t \beta_u \, du\Big)$ and integrating from s to t gives (168).

Step 2: propagate the bound through each jump and interval. We now prove (166) by induction over the jump intervals.

First, if $t \in [0, t_1) \cap [0, \tau_R)$, then there are no jumps before time t, hence (166) reduces exactly to (168) with $s = 0$.

Next, let $m \in \{1, \dots, L_{\text{jump}}\}$ and assume that (166) has been proved for all times in $[0, t_m)$. Take $t \in [t_m, t_{m+1}) \cap [0, \tau_R)$. Applying (168) on $[t_m, t]$ yields

$$r_t \le G(t, t_m)\, r_{t_m} + \int_{t_m}^{t} G(t, s)\Gamma_s^N \, ds. \tag{169}$$

Using the jump recursion (165),

$$r_{t_m} \le (1 + a_m) r_{t_m -} + b_m,$$

we obtain

$$r_t \le G(t, t_m)(1 + a_m) r_{t_m -} + G(t, t_m) b_m + \int_{t_m}^{t} G(t, s)\Gamma_s^N \, ds. \tag{170}$$

Now apply the induction hypothesis at time t_m-:

$$r_{t_m -} \le G(t_m, 0) \prod_{\ell=1}^{m-1} (1 + a_\ell)\, r_0 \quad + \int_0^{t_m} G(t_m, s) \prod_{\ell:\, s < t_\ell < t_m} (1 + a_\ell)\Gamma_s^N \, ds \tag{171}$$

$$+ \sum_{\ell=1}^{m-1} G(t_m, t_\ell) \prod_{j:\, t_\ell < t_j < t_m} (1 + a_j)\, b_\ell. \tag{172}$$

Substituting (172) into (170), we get

$$\begin{aligned} r_t \le\ & G(t, t_m)(1 + a_m) G(t_m, 0) \prod_{\ell=1}^{m-1} (1 + a_\ell)\, r_0 \\ & + G(t, t_m)(1 + a_m) \int_0^{t_m} G(t_m, s) \prod_{\ell:\, s < t_\ell < t_m} (1 + a_\ell)\Gamma_s^N \, ds \\ & + G(t, t_m)(1 + a_m) \sum_{\ell=1}^{m-1} G(t_m, t_\ell) \prod_{j:\, t_\ell < t_j < t_m} (1 + a_j)\, b_\ell \\ & + G(t, t_m) b_m + \int_{t_m}^{t} G(t, s)\Gamma_s^N \, ds. \end{aligned}$$

Using the semigroup property of G,

$$G(t, t_m) G(t_m, s) = G(t, s), \qquad 0 \le s \le t_m \le t,$$

we rewrite this as

$$\begin{aligned} r_t \le\ & G(t, 0) \prod_{\ell=1}^{m} (1 + a_\ell)\, r_0 + \int_0^{t_m} G(t, s) \prod_{\ell:\, s < t_\ell \le t_m} (1 + a_\ell)\Gamma_s^N \, ds \\ & + \sum_{\ell=1}^{m-1} G(t, t_\ell) \prod_{j:\, t_\ell < t_j \le t_m} (1 + a_j)\, b_\ell + G(t, t_m) b_m + \int_{t_m}^{t} G(t, s)\Gamma_s^N \, ds. \end{aligned}$$

Since there is no jump in (t_m, t), for $s \in (t_m, t)$ the product $\prod_{\ell:\, s < t_\ell \le t} (1 + a_\ell)$ is empty and equals 1. Also, because $t \in [t_m, t_{m+1})$, the sets

$$\{\ell :\ s < t_\ell \le t\} \quad \text{and} \quad \{\ell :\ s < t_\ell \le t_m\}$$

coincide when $s < t_m$, and similarly

$$\{j :\ t_\ell < t_j \le t\} = \{j :\ t_\ell < t_j \le t_m\} \qquad (\ell \le m - 1),$$

while for $\ell = m$ the product over $j :\ t_m < t_j \le t$ is empty and equals 1. Therefore,

$$\begin{aligned} r_t \le\ & G(t, 0) \prod_{\ell:\, t_\ell \le t} (1 + a_\ell)\, r_0 + \int_0^{t} G(t, s) \prod_{\ell:\, s < t_\ell \le t} (1 + a_\ell)\Gamma_s^N \, ds \\ & + \sum_{\ell:\, t_\ell \le t} G(t, t_\ell) \prod_{j:\, t_\ell < t_j \le t} (1 + a_j)\, b_\ell, \end{aligned}$$

which is exactly (166). This completes the induction. Taking the left limit gives the bound at a continuous candidate exit; at a candidate exit that is a jump, apply the assumed recursion at that jump itself. Empty sums and products cover the case of no jumps. □

The next lemma packages the local coefficients without discarding their dependence on the population moment scale or anti-spikiness.

Lemma A.32 (Explicit local comparison coefficients)**.** *Fix the deterministic population scale $K = K_{\mathrm{pop}}$ and radius $D \geq 1$. Use $\mathrm{L}_D, \gamma_m(D) = \mathrm{L}_D a_m$ from lemma A.6 and put*

$$\Lambda(D) = \sum_{m\geq 1} \gamma_m(D) L_m (1 + 2D^{m-1}), \qquad c_D = \Lambda(D) + (2 + A_{\mathrm{sp}})\mathrm{L}_D,$$

$$c_J(D) = 1 + \sqrt{d}\{\Lambda(D) + 2\mathrm{L}_D\} + \sqrt{2d\mathrm{L}_D}.$$

On $E_{\mathrm{pop}} \cap E_{T,D} \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$, c_D and $c_J(D)$ are the coefficients in lemmas A.12 and A.17. With $p_0 = p/8$, write $\Xi = 8N(e+T)(L_{\mathrm{jump}}+1)(1+B)/p$ and choose

$$\mathrm{D}_{\mathrm{sup}} = D = C_D(1 + \log \Xi)^{1/\alpha}, \tag{173}$$

where the fixed C_D is sufficiently large for (137) *with budget p_0 and for the clipping-bias estimate in Lemma A.25. For fixed model parameters and uniformly bounded K_0, B, A_{sp}, for every $\varepsilon > 0$ there is C_ε such that*

$$1 + c_D + c_J(D) + \Lambda(D) \leq C_\varepsilon \Xi^\varepsilon. \tag{174}$$

The constants may depend on the fixed anchor derivatives and C_3.

Proof. The empirical moment displacement is $L_m(1 + 2D^{m-1})r_t$ by lemma A.10, (84). The sum of the two Hessian measure errors is bounded with weights $\gamma_m(D)$ by eq. (46), so its coefficient is $\Lambda(D)$. The two spatial kernel slots contribute $2\mathrm{L}_D$ by eq. (50); the diagonal term contributes $A_{\mathrm{sp}}\mathrm{L}_D$ by eq. (49) and eq. (24). Thus the ODE coefficient is exactly c_D. The jump coefficient is exactly the one in Lemma A.17. For the common growth bound, $D \geq 1$ and (37) give

$$\Lambda(D) \leq 2\mathrm{L}_D \sum_m a_m L_m (2 + D^m + K_{\mathrm{pop}}^m m^{m/\alpha}) \leq \mathrm{L}_D^2.$$

For the last inequality, use $2u \leq 4(1+u)^3$ for $u \geq 0$ and $1 + g_0 + h_0 + C_3 \geq 1$ in the definition of L_D. Using $\mathrm{L}_D \geq 1$ in the displayed definitions therefore yields

$$1 + c_D + c_J(D) + \Lambda(D) \leq (6 + A_{\mathrm{sp}} + 3\sqrt{d} + \sqrt{2d})\mathrm{L}_D^2. \tag{175}$$

By lemma A.20, eqs. (119) and (120), the grid count is at most $C(e+T)(L_{\mathrm{jump}}+1)(1+B)$, so $\log A_0 \leq C + \log \Xi$ and $K_{\mathrm{pop}} \leq C(1 + \log \Xi)^{1/\alpha}$ when K_0, B are bounded. Insert this and (173) into lemma A.4, (38):

$$\mathrm{L}_D^2 \leq C \exp\{C[1 + \log\log(e + \Xi)]^{\kappa_{\mathrm{dec}}/(\kappa_{\mathrm{dec}}-1)}\} \leq C_\varepsilon \Xi^\varepsilon.$$

The last step uses $[\log\log(e+\Xi)]^{\kappa_{\mathrm{dec}}/(\kappa_{\mathrm{dec}}-1)} = o(\log \Xi)$. This proves (174). □

Setup and curvature-response factors for the tracking theorem. Fix $N \geq 2$, $p \in (0,1)$, a deterministic schedule $0 < t_1 < \cdots < t_{L_{\mathrm{jump}}} \leq T$ with $\eta_\ell \geq 0$, and $A_{\mathrm{sp}} \geq 1$. Set $p_0 = p/8$ and use the population constants of lemma A.20 with budget p_0. Choose D satisfying (137) with budget p_0; (173) is one such choice. Use c_D, c_J from lemma A.32. Define the actual deterministic residual envelopes by

$$\delta_{N,T,p} := \delta_\Gamma(D) \quad \text{from (127), with } p_\Gamma = p_0, \tag{176}$$

$$\widetilde{\delta}_{N,T,p} := \widetilde{\delta}^{\mathrm{HS}}_{N,T,p_0} \quad \text{from (161)}, \tag{177}$$

For readability write $\delta = \delta_{N,T,p}$ and $\widetilde{\delta} = \widetilde{\delta}_{N,T,p}$ in the following statement and proof.

For the sampled full Hessian curvature $\kappa_t^{Y,N}$, on any path where $\int_0^T |\kappa_t^{Y,N}|\, dt < \infty$, recall the curvature-response factors of definition 3.1:

$$G_\kappa^Y(t,s) = \exp\Big(-\int_s^t \kappa_u^{Y,N}\, du\Big), \qquad \mathfrak{C}_t^Y = \int_0^t G_\kappa^Y(t,s)\, ds, \qquad \mathfrak{J}_t^Y = \sum_{\ell: t_\ell \leq t} \eta_\ell G_\kappa^Y(t, t_\ell). \tag{178}$$

Set

$$q_p = 1 + \sqrt{2\log(2(L_{\mathrm{jump}}+1)/p_0)}, \qquad a_\ell = c_J(D) q_p \eta_\ell,$$

$$J_{N,T,p} := \prod_{\ell=1}^{L_{\mathrm{jump}}} (1 + a_\ell), \tag{179}$$

$$\mathfrak{B}_{N,T,p}(t) := \delta \mathfrak{C}_t^Y + c_J(D) q_p (\sqrt{\delta} + \sqrt{\widetilde{\delta}}) \mathfrak{J}_t^Y, \qquad \mathfrak{B}_* := \sup_{t \le T} \mathfrak{B}_{N,T,p}(t). \tag{180}$$

Empty sums and products are zero and one, respectively. The envelope $\mathfrak{B}$ is generally random because the curvature-response factors are random. Let $R = 2J_{N,T,p}\mathfrak{B}_*$ and define the nonlinear stability event

$$\mathcal{S}_{N,T,p} := \{T(\delta + c_D R) \le 1/4\}. \tag{181}$$

Which quantities are fixed. The following convention applies to the asymptotic corollary; the explicit theorem retains the actual deterministic envelopes.

Fixed structural bounds	d, α, feature growth/decay parameters, g_0, h_0, C_3, and the initial population law.
Uniform regularity assumptions	K_0, B, A_{sp}; these are assumptions along the growing horizon, not consequences of localization.
Deterministic growing quantities	$D, K_{\mathrm{pop}}, \delta_{N,T,p}, \widetilde{\delta}_{N,T,p}, c_D, c_J, \Lambda, q_p, J_{N,T,p}$. The residual envelopes are defined in eqs. (176) and (177); their dependence on N, T, p is retained.
Random reference quantities	$\kappa^{Y,N}, \mathfrak{C}^Y, \mathfrak{J}^Y, \mathfrak{B}_*, R$ and the nonlinear stability event. Their bounds are pathwise conditions.

Theorem A.1 (Fixed-schedule comparison on the nonlinear stability event)**.** *Assume assumptions A.1 to A.4 and use the deterministic schedule, $N \ge 2$, $p \in (0,1)$, and radius specified above, with $r_0 = 0$ and integrable curvature on $[0,T]$. Use the residual envelopes eqs.* (176) *and* (177)*, the amplification* (179)*, the forcing* (180)*, and the nonlinear stability event* (181)*. There is an event Ω_p with $\mathbb{P}(\Omega_p) \ge 1 - p$ such that on $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{S}_{N,T,p}$, simultaneously at all times and left limits,*

$$r_t \le e^{1/4} J_{N,T,p} \mathfrak{B}_{N,T,p}(t) < \tfrac{3}{2} J_{N,T,p} \mathfrak{B}_{N,T,p}(t) \quad \text{when } \mathfrak{B}_{N,T,p}(t) > 0, \tag{182}$$

and $r_t = 0$ if that envelope is zero. In particular,

$$\mathbb{P}(E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{S}_{N,T,p} \cap \{\exists t \le T : r_t > 2J_{N,T,p}\mathfrak{B}_{N,T,p}(t)\}) \le p.$$

This is a bound on failure within the stated events, not a bound on the conditional failure probability given $\mathcal{S}_{N,T,p}$.

Remark (Relation to the main-text growing-horizon theorem). The main-text result in Theorem 3.1 applies the preceding theorem on paths satisfying the controlled-curvature path condition $\mathcal{R}_N$ and uses the simple sufficient horizon $T \le N^{1/4-\upsilon-\zeta}$. The integrable-curvature condition imposes no separate model assumption there: on the common good event used in the proof, Equation (183) gives $|\kappa_t^{Y,N}| \le 2\mathrm{L}_D$, and hence

$$\int_0^T |\kappa_t^{Y,N}| \, \mathrm{d}t \le \int_0^T 2\mathrm{L}_D \, \mathrm{d}t = 2T\mathrm{L}_D < \infty.$$

The explicit stability requirement remains $T(\delta_{N,T,p} + c_D R) \le 1/4$ by Equation (181); it can also hold on longer stable trajectories. The strict margin $\zeta > 0$ in the main-text sufficient regime absorbs the subpolynomial concentration and localization factors. Finally, the probability statement is a bound on failure within the controlled-curvature path condition and the nonlinear stability event. It neither lower-bounds $\mathbb{P}(\mathcal{R}_N)$ nor gives a conditional failure probability given that condition.

Proof. **Step 1: a common good event.** Recall $p_0 = p/8$ and the residual envelopes (176)–(177). Define

$$E_\Gamma := \{\sup_{t\le T} \Gamma_t^N \le \delta\}, \qquad E_{\widetilde{\Gamma}} := \{\max_{\ell\le L_{\mathrm{jump}}} \widetilde{\Gamma}_{t_\ell -}^N \le \widetilde{\delta}\}.$$

Every supremum includes the relevant left and right traces. Use the population and radius events with budgets p_0 specified in the statement. For each jump, let $\Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)}$ be the Gaussian event in lemma A.17, and set

$$\Omega_p := E_{\mathrm{pop}} \cap E_{T,D} \cap E_\Gamma \cap E_{\widetilde{\Gamma}} \cap \bigcap_{\ell=1}^{L_{\mathrm{jump}}} \Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)}.$$

The failure bounds and their sources are:

Failure event	Budget	Source
E_{pop}^c	p_0	lemma A.20
$E_{T,D}^c$	p_0	lemma A.24, (137)
$E_{\mathrm{pop}} \cap E_{T,D} \cap E_\Gamma^c$	p_0	lemma A.21, (128)
$E_{\mathrm{pop}} \cap E_{\widetilde{\Gamma}}^c$	p_0	corollary A.3, (162)
$(\Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)})^c$	$\frac{p_0}{L_{\mathrm{jump}}+1}$	lemma A.17

The covariance estimate conditions on $\mathcal{G}_{\mathrm{pop}}$, while the Gaussian jump estimate conditions on $\mathcal{H}_{\ell-}$. The tower property converts these bounds to unconditional ones:

$$\mathbb{P}(E_{\mathrm{pop}} \cap E_{\widetilde{\Gamma}}^c) = \mathbb{E}[\mathbf{1}_{E_{\mathrm{pop}}} \mathbb{P}(E_{\widetilde{\Gamma}}^c \mid \mathcal{G}_{\mathrm{pop}})] \le p_0,$$
$$\mathbb{P}((\Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)})^c) = \mathbb{E}[\mathbb{P}((\Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)})^c \mid \mathcal{H}_{\ell-})] \le p_0/(L_{\mathrm{jump}}+1).$$

Consequently,

$$\begin{aligned}\mathbb{P}(\Omega_p^c) &\le \mathbb{P}(E_{\mathrm{pop}}^c) + \mathbb{P}(E_{T,D}^c) + \mathbb{P}(E_{\mathrm{pop}} \cap E_{T,D} \cap E_\Gamma^c) \\ &\quad + \mathbb{P}(E_{\mathrm{pop}} \cap E_{\widetilde{\Gamma}}^c) + \sum_{\ell=1}^{L_{\mathrm{jump}}} \mathbb{P}((\Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)})^c) \\ &\le 4p_0 + \frac{L_{\mathrm{jump}}}{L_{\mathrm{jump}}+1} p_0 < 5p_0 < p.\end{aligned}$$

This union bound does not require independence. The sample event $E_{T,D}$ is intersected after concentration; it is not part of the conditioning used to establish the reference product law.

Step 2: deterministic comparison on the good event. Fix a realization in $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}})$. The two deterministic inputs can now be tied to explicit parts of this event. On $E_{\mathrm{pop}} \cap E_{T,D} \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap E_\Gamma$, lemma A.12 gives

$$\dot{r}_t \le -\kappa_t^{Y,N} r_t + c_D r_t^2 + \delta r_t + \delta.$$

On the other hand, on $E_{\mathrm{pop}} \cap E_{T,D} \cap E_\Gamma \cap E_{\widetilde{\Gamma}} \cap \Omega^{\mathrm{jump}}_{\ell,p_0/(L_{\mathrm{jump}}+1)}$, lemma A.17 gives

$$r_{t_\ell} \le (1 + a_\ell) r_{t_\ell -} + a_\ell(\sqrt{\delta} + \sqrt{\widetilde{\delta}}).$$

Before applying these two inputs, lemma A.6, (47), and the block estimate in lemma A.9 give

$$|\kappa_t^{Y,N}| \le \|\mathrm{H}_t^Y\|_{\mathbb{R}^{dN}\to\mathbb{R}^{dN}} \le \big\|(\mathrm{L}_D/N)\mathbf{1}\mathbf{1}^\top + \mathrm{L}_D I_N\big\|_{\mathbb{R}^N\to\mathbb{R}^N} = 2\mathrm{L}_D. \tag{183}$$

Here $\mathbf{1} = (1,\dots,1)^\top \in \mathbb{R}^N$ and $P = N^{-1}\mathbf{1}\mathbf{1}^\top$ satisfies $P^\top = P$, $P^2 = P$, and $P\mathbf{1} = \mathbf{1}$. Thus $\|P + I_N\|_{\mathbb{R}^N\to\mathbb{R}^N} = 2$. In particular $\int_0^T |\kappa_t^{Y,N}|\,dt \le 2T\mathrm{L}_D < \infty$ on this good event, so the curvature-response factors in (178) are well-defined. The absolute-continuity premise of lemma A.31

is also automatic on each ODE branch. Indeed, the branchwise B-Lipschitz bounds in assumption A.4 give

$$|r_t - r_s| \le \frac{1}{\sqrt{N}}\|\Delta_t - \Delta_s\|_{\mathbb{R}^{dN}}$$
$$\le \frac{1}{\sqrt{N}}\left[\sum_{i=1}^{N}\big(\|X_t^i - X_s^i\|_{\mathbb{R}^d} + \|Y_t^i - Y_s^i\|_{\mathbb{R}^d}\big)^2\right]^{1/2} \le 2B|t-s|.$$

Suppose first $\mathfrak{B}_* > 0$, and recall $R = 2J_{N,T,p}\mathfrak{B}_*$ from the statement. Before a first exit from $r < R$, replace $c_D r_t^2$ by $c_D R r_t$. The exponent in lemma A.31 satisfies

$$\exp\left(-\int_s^t \beta_u\,du\right) = G_\kappa^Y(t,s)\exp\left((t-s)c_D R + \int_s^t \Gamma_u^N\,du\right)$$
$$\le G_\kappa^Y(t,s)e^{T(c_D R+\delta)}.$$

Every indicated subproduct of $(1+a_\ell)$ is at most $J_{N,T,p} = \prod_{\ell=1}^{L_{\mathrm{jump}}}(1+a_\ell)$, since $a_\ell \ge 0$. Apply lemma A.31, (166), retaining the two forcing terms:

$$r_t \le J_{N,T,p}e^{T(c_D R+\delta)}\left[\int_0^t G_\kappa^Y(t,s)\Gamma_s^N\,ds + \sum_{\ell:t_\ell\le t} G_\kappa^Y(t,t_\ell)b_\ell\right]. \tag{184}$$

The continuous and jump forcing terms satisfy

$$\int_0^t G_\kappa^Y(t,s)\Gamma_s^N\,ds \le \delta\mathfrak{C}_t^Y,$$
$$\sum_{\ell:t_\ell\le t} G_\kappa^Y(t,t_\ell)b_\ell \le c_J(D)q_p(\sqrt{\delta}+\sqrt{\tilde{\delta}})\sum_{\ell:t_\ell\le t}\eta_\ell G_\kappa^Y(t,t_\ell)$$
$$= c_J(D)q_p(\sqrt{\delta}+\sqrt{\tilde{\delta}})\mathfrak{J}_t^Y.$$

Here b_ℓ is the additive forcing in (165). The sum of the two right-hand sides is $\mathfrak{B}_{N,T,p}(t)$ from (180). Substitution into (184) gives

$$r_t \le J_{N,T,p}e^{T(c_D R+\delta)}\mathfrak{B}_{N,T,p}(t). \tag{185}$$

Step 3: exclude all first exits. The candidate-exit conclusion of lemma A.31, (166), applies at a continuous exit, a pre-jump trace, or a post-jump exit, including a jump at T. At a post-jump exit its partial products already include that jump. Thus the same bounds on the propagator, products, and forcing give eqs. (184) and (185) at the candidate exit itself; no extra factor $1+a_\ell$ is introduced. On the nonlinear stability event (181), the candidate value is therefore at most

$$e^{1/4}J_{N,T,p}\mathfrak{B}_* = \frac{e^{1/4}}{2}R < R.$$

There are finitely many branches and jumps, so this strict margin excludes an exit on $[0,T]$. The radius may depend on future reference curvature-response factors: the argument is pathwise and uses no stopping-time theorem.

If $\mathfrak{B}_* = 0$ and $T > 0$, positivity of $G_\kappa^Y(t,s)$ implies $\delta = 0$, and every nonnegative jump forcing $a_\ell(\sqrt{\delta}+\sqrt{\tilde{\delta}})$ is zero. On a branch and below any fixed positive radius R', the ODE inequality becomes $\dot{r}_t \le (|\kappa_t^{Y,N}| + c_D R')r_t$. Multiplication by the integrating factor, starting from $r = 0$, gives $r_t \le 0$ on that branch. The jump recursion sends zero to zero. Induction proves $r \equiv 0$; no zero radius is used in the first-exit lemma. For $T = 0$, the conclusion is just $r_0 = 0$.

Step 4: the failure event. The preceding argument proves the pathwise bound on $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{S}_{N,T,p}$. Consequently any violation within the regularity event and the nonlinear stability event belongs to Ω_p^c, whose probability is at most p. No conditional failure probability given the nonlinear stability event is asserted. □

Corollary A.4 (A simple sufficient horizon and the curvature-dependent rate)**.** *Use the same model as N varies, with fixed structural and anchor bounds and uniform K_0, B, A_{sp}. Fix $\upsilon \geq 0$, $\zeta > 0$ with $\upsilon + \zeta < 1/4$, and $C_p, C_0, L_0 > 0$. Suppose*

$$T \leq N^{1/4-\upsilon-\zeta}, \qquad p \geq N^{-C_p}, \qquad L_{\mathrm{jump}} \leq L_0, \qquad 0 \leq \eta_\ell \leq C_0,$$

and choose D as in (173)*. Define the controlled-curvature path condition $\mathcal{R}_N$ by*

$$\mathcal{R}_N := \left\{ \int_0^T |\kappa_t^{Y,N}|\, dt < \infty, \quad \mathfrak{C}_*^Y := \sup_{t\leq T} \mathfrak{C}_t^Y \leq C_0 N^\upsilon (1+T), \quad \mathfrak{J}_*^Y := \sup_{t\leq T} \mathfrak{J}_t^Y \leq C_0 N^\upsilon \right\}. \tag{186}$$

For all sufficiently large N, the condition $\mathcal{R}_N$ implies the nonlinear stability event $\mathcal{S}_{N,T,p}$ of (181)*. For every fixed $\varepsilon > 0$, on $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{R}_N$,*

$$r_t \leq C_\varepsilon \Xi^\varepsilon \left[(N^{-1/2} + T/N)\mathfrak{C}_t^Y + (N^{-1/4} + \sqrt{T/N})\mathfrak{J}_t^Y \right], \qquad t \leq T. \tag{187}$$

Thus the same bound holds with failure probability at most p within $E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{R}_N$.

Proof. **Step 1: recall the deterministic envelopes.** In this proof, write $\delta = \delta_{N,T,p}$ and $\widetilde{\delta} = \widetilde{\delta}_{N,T,p}$ as defined in eqs. (176) and (177). Use lemma A.25, (163), and (174). Uniform K_0, B, A_{sp} imply that K_{pop}, D are bounded by fixed powers of $\log \Xi$. For lemma A.25, use total budget $3p_0 = 3p/8 < 1$. Its population, radius, and concentration budgets are then each p_0, and its scale agrees exactly with the present one:

$$\frac{3N(e+T)(L_{\mathrm{jump}}+1)(1+B)}{3p_0} = \frac{N(e+T)(L_{\mathrm{jump}}+1)(1+B)}{p_0} = \Xi.$$

The polynomial-amplitude premise of that lemma holds with $c_M = 1$. Indeed, eq. (119) and the grid bound in lemma A.20 give

$$M_{\mathrm{pop}} \leq \frac{4eQ}{p_0} \leq \frac{C(e+T)(L_{\mathrm{jump}}+1)(1+B)}{p_0} = \frac{C\Xi}{N} \leq C\Xi.$$

For (163) with budget $p_0 = p/8$, the present $\Xi \geq N(L_{\mathrm{jump}}+1)/p_0$ satisfies its scale condition directly. For every $\theta > 0$, the cited bounds give

$$\begin{aligned} \delta \leq C_\theta \Xi^\theta (N^{-1/2} + T/N), \quad \widetilde{\delta} \leq C_\theta \Xi^\theta N^{-1/2}, \\ c_D + c_J(D) q_p + \Lambda(D) \leq C_\theta \Xi^\theta. \end{aligned} \tag{188}$$

For $c_J q_p$, apply the individual bounds with exponent $\theta/2$ to each factor: $q_p = 1 + \sqrt{2\log(2(L_{\mathrm{jump}}+1)/p_0)}$ also satisfies $q_p \leq C_\theta \Xi^{\theta/2}$.

Step 2: count the powers in the jump and curvature-response bounds. The amplitude bound $0 \leq \eta_\ell \leq C_0$, (179), and $L_{\mathrm{jump}} \leq L_0$ imply

$$J_{N,T,p} \leq (1 + C_0 c_J(D) q_p)^{L_0} \leq C_\theta \Xi^{L_0 \theta}.$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ in (188),

$$\sqrt{\delta} + \sqrt{\widetilde{\delta}} \leq C_\theta \Xi^{\theta/2} (N^{-1/4} + \sqrt{T/N}).$$

By (180) and the controlled-curvature path condition (186),

$$\mathfrak{B}_* \leq \delta \mathfrak{C}_*^Y + c_J(D) q_p (\sqrt{\delta} + \sqrt{\widetilde{\delta}}) \mathfrak{J}_*^Y.$$

Substitute the preceding estimates and $T/N \leq N^{-1/2}$ to get

$$\mathfrak{B}_* \leq C_\theta \Xi^\theta N^\upsilon (1+T) N^{-1/2} + C_\theta \Xi^{3\theta/2} N^\upsilon N^{-1/4}.$$

Multiplying by $2J_{N,T,p}$ and using $\Xi \geq 1$ yields

$$R \leq C_\theta \Xi^{(L_0+3/2)\theta} N^\upsilon \{(1+T)N^{-1/2} + N^{-1/4}\}.$$

Step 3: verify the nonlinear stability event. Recall $\Xi = 8N(e+T)(L_{\text{jump}}+1)(1+B)/p$. The horizon and probability assumptions imply $\Xi \le CN^{5/4+C_p}$. For the desired $\varepsilon > 0$, choose a fixed $\theta > 0$ sufficiently small that

$$(5/4 + C_p)(L_0 + 5/2)\theta \le \zeta/2, \qquad (L_0 + 3/2)\theta < \varepsilon. \tag{189}$$

This choice depends only on the fixed parameters and ε, ζ, not on N, T, p. The preceding bounds and $N^{\upsilon} \ge 1$ give

$$\begin{aligned} T(\delta + c_D R) &\le C_\theta \Xi^{(L_0+5/2)\theta} N^{\upsilon} \left\{ \frac{T(1+T)}{\sqrt{N}} + \frac{T}{N^{1/4}} \right\} \\ &\le C_\theta N^{\zeta/2} (N^{-1/4-\zeta} + N^{-\upsilon-2\zeta} + N^{-\zeta}) \le 3C_\theta N^{-\zeta/2} \longrightarrow 0. \end{aligned}$$

The second line substitutes $T \le N^{1/4-\upsilon-\zeta}$ into each term:

$$N^{\upsilon} T/\sqrt{N} \le N^{-1/4-\zeta}, \qquad N^{\upsilon} T^2/\sqrt{N} \le N^{-\upsilon-2\zeta}, \qquad N^{\upsilon} T/N^{1/4} \le N^{-\zeta}.$$

The choice (189) ensures $(5/4 + C_p)(L_0 + 5/2)\theta \le \zeta/2$. For all sufficiently large N, the last bound is at most $1/4$, proving $\mathcal{R}_N \subset \mathcal{S}_{N,T,p}$.

Step 4: substitute into the tracking bound. The exact theorem and Steps 1–2 give

$$r_t \le C_\theta \Xi^{(L_0+3/2)\theta} \left[(N^{-1/2} + T/N)\mathfrak{C}_t^Y + (N^{-1/4} + \sqrt{T/N})\mathfrak{J}_t^Y \right].$$

Since $(L_0 + 3/2)\theta < \varepsilon$, this proves (187). The exact theorem continues to apply at longer horizons on its nonlinear stability event (181). □

A.7 Functional-value comparison

This subsection converts the particle tracking estimate into a functional-value estimate. The key point is that no new stochastic residual is needed.

We now turn the r_t-comparison into the final estimate for

$$\left|F(\rho_{\boldsymbol{X}_t}) - F(\mu_t)\right|.$$

The comparison uses Γ_t^N for the empirical moment fluctuation of $\rho_{\mathbf{Y}_t}$ and r_t for the discrepancy between the two particle systems.

No additional error term is introduced.

The next lemma is the deterministic conversion step: the functional gap is controlled by the pathwise particle discrepancy and the same moment residual Γ_t^N.

Lemma A.33 (Fixed-time functional discrepancy)**.** *In this subsection we use the convention*

$$\gamma_m(D) = \mathrm{L}_D a_m \quad \textit{from eq. (44).}$$

Fix the population scale and work on E_{pop} as in the tracking theorem. Then, on the localization event $E_{T,\mathrm{D}_{\sup}}$, for every $t \in [0,T]$,

$$\left|F(\rho_{\boldsymbol{X}_t}) - F(\mu_t)\right| \le \Lambda(\mathrm{D}_{\sup})\, r_t + \Gamma_t^N, \tag{190}$$

where

$$\Lambda(\mathrm{D}_{\sup}) \coloneqq \sum_{m \ge 1} \gamma_m(\mathrm{D}_{\sup}) L_m (1 + 2\mathrm{D}_{\sup}^{m-1}).$$

Proof. **Step 1: split the functional error through the reference sample.** By the value bound in lemma A.6 and eq. (45),

$$\left|F(\rho_{\boldsymbol{X}_t}) - F(\mu_t)\right| \le \sum_{m \ge 1} \gamma_m(\mathrm{D}_{\sup}) \left\| M_m(\rho_{\boldsymbol{X}_t}) - M_m(\mu_t) \right\|_{\mathsf{H}_m}.$$

Split the difference as

$$M_m(\rho_{\boldsymbol{X}_t}) - M_m(\mu_t) = \big(M_m(\rho_{\boldsymbol{X}_t}) - M_m(\rho_{\boldsymbol{Y}_t})\big) + \big(M_m(\rho_{\boldsymbol{Y}_t}) - M_m(\mu_t)\big).$$

Step 2: bound the paired empirical moments and recall the residual. On $E_{T,\mathrm{D}_{\sup}}$, both configurations lie in the ball of radius $D = \mathrm{D}_{\sup}$. Set $s = 1$ in the empirical-moment calculation eq. (84) in the proof of lemma A.10. Since $\mathbf{Z}_t^{(1)} = \mathbf{X}_t$, it gives

$$\|M_m(\rho_{\boldsymbol{X}_t}) - M_m(\rho_{\boldsymbol{Y}_t})\|_{\mathsf{H}_m} \le L_m(1 + 2D^{m-1})r_t.$$

For the population-to-reference difference, recall lemma A.10 and eq. (82):

$$\sum_m \gamma_m(D)\|M_m(\rho_{\mathbf{Y}_t}) - M_m(\mu_t)\|_{\mathsf{H}_m} = \Gamma_t^N.$$

Summing the two comparisons with weights $\gamma_m(D)$ yields

$$\big|F(\rho_{\boldsymbol{X}_t}) - F(\mu_t)\big| \le \Lambda(\mathrm{D}_{\sup})r_t + \Gamma_t^N.$$

This is (190). □

Combining this deterministic conversion with the fixed-schedule tracking theorem gives the final uniform functional-value comparison.

Theorem A.2 (Uniform functional comparison on the tracking event)**.** *Under Theorem A.1, on the same event $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{S}_{N,T,p}$,*

$$|F(\rho_{\mathbf{X}_t}) - F(\mu_t)| \le 2\Lambda(D)J_{N,T,p}\mathfrak{B}_{N,T,p}(t) + \delta_{N,T,p}, \qquad t \le T.$$

Under all the assumptions of corollary A.4, for sufficiently large N the following rate holds on $\Omega_p \cap E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{R}_N$, where $\mathcal{R}_N$ is the controlled-curvature path condition in eq. (186)*. The curvature-response factors are those of theorem A.1. On this event,*

$$|F(\rho_{\mathbf{X}_t}) - F(\mu_t)| \le C_\varepsilon \Xi^\varepsilon \left[(N^{-1/2} + T/N)(1 + \mathfrak{C}_t^Y) + (N^{-1/4} + \sqrt{T/N})\mathfrak{J}_t^Y\right]. \tag{191}$$

More precisely, with the same Ω_p and no additional event,

$$\mathbb{P}\left(E_{T,\mathrm{sp}}(A_{\mathrm{sp}}) \cap \mathcal{R}_N \cap \{(191) \text{ fails for some } t \le T\}\right) \le p.$$

This is a joint failure bound within the indicated sets, not a conditional probability bound given those sets.

Proof. **Step 1: apply the deterministic value comparison.** On the event in the first assertion, theorem A.1 and eq. (182) gives $r_t \le e^{1/4}J_{N,T,p}\mathfrak{B}_{N,T,p}(t) \le 2J_{N,T,p}\mathfrak{B}_{N,T,p}(t)$. Also $\Gamma_t^N \le \delta_{N,T,p}$ by the construction of Ω_p in that theorem. Substituting these two bounds into lemma A.33 and eq. (190) gives

$$|F(\rho_{\mathbf{X}_t}) - F(\mu_t)| \le \Lambda(D)r_t + \Gamma_t^N \le 2\Lambda(D)J_{N,T,p}\mathfrak{B}_{N,T,p}(t) + \delta_{N,T,p}.$$

Step 2: substitute the envelopes for the sufficient-horizon rate. Fix $\varepsilon > 0$. Use exponent $\varepsilon/2$ in the coefficient bound lemma A.32 and eq. (174) and the envelope bound corollary A.4 and eq. (188):

$$\Lambda(D) \le C_\varepsilon \Xi^{\varepsilon/2}, \qquad \delta_{N,T,p} \le C_\varepsilon \Xi^{\varepsilon/2}(N^{-1/2} + T/N).$$

Apply the tracking rate in corollary A.4 and eq. (187) with exponent $\varepsilon/2$ as well. Multiplying that rate by $\Lambda(D)$ and adding $\delta_{N,T,p}$ yields

$$\begin{aligned}\Lambda(D)r_t + \delta_{N,T,p} &\le C_\varepsilon \Xi^{\varepsilon/2}\Xi^{\varepsilon/2}\left[(N^{-1/2} + T/N)\mathfrak{C}_t^Y + (N^{-1/4} + \sqrt{T/N})\mathfrak{J}_t^Y\right] \\ &\quad + C_\varepsilon \Xi^{\varepsilon/2}(N^{-1/2} + T/N) \\ &\le C_\varepsilon \Xi^\varepsilon \left[(N^{-1/2} + T/N)(1 + \mathfrak{C}_t^Y) + (N^{-1/4} + \sqrt{T/N})\mathfrak{J}_t^Y\right],\end{aligned}$$

where C_ε is enlarged as needed. The last step uses $\Xi^{\varepsilon/2}\Xi^{\varepsilon/2} = \Xi^\varepsilon$ and $\Xi \ge 1$. Together with lemma A.33 and eq. (190) and $\Gamma_t^N \le \delta_{N,T,p}$ on the same Ω_p, this proves (191). □

A.8 A GROWING-HORIZON EXAMPLE: ESCAPE AND CURVATURE RECOVERY

This example follows a single Gaussian jump from a negative-curvature mean to a positive-curvature region. Two lemmas verify the comparison assumptions and bound the curvature-response factors; the proposition then applies the general comparison theorem on a growing horizon.

Model and fixed parameters. In dimension one, let

$$F(\mu) = \frac{a}{2}\,\mathrm{Var}(\mu) + \beta\cos(\pi m_\mu), \qquad m_\mu := \int x\,\mu(dx), \qquad a > 0, \quad \beta > 0. \tag{192}$$

Fix $R_0 \geq 0$, a jump time $\tau > 0$, and an amplitude $\eta > 0$, independently of N and T. Let the deterministic initial law μ_0 be supported in $[-R_0, R_0]$ with $m_{\mu_0} = 0$, and set $X_0^i = Y_0^i = U_i$ for an i.i.d. sample $(U_i)_{i=1}^N$ from μ_0. Use the population-first coupling of lemma A.14, with exactly one prescribed jump at τ of amplitude η, and no later jumps. Write $m_t = m_{\mu_t}$ and $b := \beta\pi^2$. Below, C depends only on $a, \beta, R_0, \tau, \eta$ and may increase between inequalities; C_ε may also depend on ε. The minimum of F over $\mathcal{P}_2(\mathbb{R})$ is $-\beta$: variance is nonnegative, $\cos \geq -1$, and a Dirac mass at an odd integer attains equality.

Proposition A.1 (Tracking after escape on a growing horizon)**.** *For every fixed $\varepsilon > 0$, there are $C_\varepsilon < \infty$ and N_0, depending only on the fixed model parameters and ε, such that for $N \geq N_0$ and $T = N^{1/8}$, the following hold on the same event of probability at least $1 - N^{-1/16}$. Here r_t is the discrepancy in* (2.4)*, and $\rho_{\mathbf{X}_t} = N^{-1}\sum_i \delta_{X_t^i}$.*

(i) Uniformly over the growing horizon,

$$\sup_{0\leq t\leq T} [r_t + |F(\rho_{\mathbf{X}_t}) - F(\mu_t)|] \leq C_\varepsilon N^{-3/16+\varepsilon}. \tag{193}$$

(ii) After a time $C\log N < T$,

$$\sup_{C\log N\leq t\leq T} [r_t + |F(\rho_{\mathbf{X}_t}) - F(\mu_t)| + F(\rho_{\mathbf{X}_t}) + \beta] \leq C_\varepsilon N^{-1/2+\varepsilon}. \tag{194}$$

Lemma A.34 (Verification of the comparison assumptions)**.** *The model* (192) *with the initial law and single jump specified above satisfies:*

(i) assumption A.1, with structural parameters independent of N, T;

(ii) assumption A.2, with a spatially constant Gaussian field;

(iii) assumption A.3, with $\alpha = 2$ and $K_0 \leq C$;

(iv) assumption A.4, with $B \leq C$;

(v) assumption A.5, with $A_{\mathrm{sp}} = 1$ and $E_{T,\mathrm{sp}}(1)$ holding almost surely for every T.

Proof. **Step 1: moment structure and weighted smoothness, for (i).** Use scalar feature spaces with $\phi_1(x) = x$, $\phi_2(x) = x^2$, and $\phi_j = 0$ for $j \geq 3$, and set

$$G(\theta) = \frac{a}{2}(\theta_2 - \theta_1^2) + \beta\cos(\pi\theta_1).$$

The nonzero features have derivative triples $(\phi_1', \phi_1'', \phi_1''') = (1, 0, 0)$ and $(\phi_2', \phi_2'', \phi_2''') = (2x, 2, 0)$. They satisfy (20) with $C_\phi = c_{\mathrm{gr}} = 1$, $\kappa_{\mathrm{gr}} = 0$, and $L_j = ej$. Choose $a_j = e^{-j^2}$, so $c_a = 1$ and $\kappa_{\mathrm{dec}} = 2$. The weighted norm in (16) gives $|h_j| \leq a_j^{-1}\|h\|_{\mathsf{H}_a}$; hence each coordinate projection is continuous and the finite-coordinate composition G is C^3 on $\mathcal{U} = \mathsf{H}_a$. Its derivatives are

$$\begin{aligned} \mathrm{D}G(\theta)[h] &= (-a\theta_1 - \beta\pi\sin(\pi\theta_1))h_1 + \tfrac{a}{2}h_2, \\ \mathrm{D}^2G(\theta)[h,k] &= (-a - b\cos(\pi\theta_1))h_1k_1, \\ \mathrm{D}^3G(\theta)[h,k,r] &= \beta\pi^3\sin(\pi\theta_1)h_1k_1r_1. \end{aligned} \tag{195}$$

In particular,

$$|\mathrm{D}^3 G(\theta)[h,k,r]| \le \beta\pi^3 |h_1 k_1 r_1| \le \beta\pi^3 a_1^{-3} \|h\|_{\mathsf{H}_a} \|k\|_{\mathsf{H}_a} \|r\|_{\mathsf{H}_a}.$$

Thus (21) holds with $C_3 = \beta\pi^3 a_1^{-3}$; the anchor $\theta_0 = M(\delta_0) = 0$ in (30) gives $g_0 \le a/(2a_2)$ and $h_0 \le (a+b)/a_1^2$. For $\nu \in \mathcal{P}_2(\mathbb{R})$, the moment sequence has at most two nonzero coordinates and

$$\|M(\nu)\|_{\mathsf{H}_a}^2 = a_1^2 \left| \int x\, \nu(dx) \right|^2 + a_2^2 \left| \int x^2\, \nu(dx) \right|^2 < \infty,$$
$$G(M(\nu)) = \tfrac{a}{2} \operatorname{Var}(\nu) + \beta \cos(\pi m_\nu).$$

This verifies the objective identity of definition A.3.

Step 2: Gaussian perturbations, for (ii). In the chain-rule formulas of lemma A.6 and eqs. (52) to (54), substitute (195) and the feature derivatives:

$$\nabla_W F(\mu, x) = (-a m_\mu - \beta\pi \sin(\pi m_\mu)) \cdot 1 + \tfrac{a}{2} \cdot 2x,$$
$$\nabla_x \nabla_W F(\mu, x) = \tfrac{a}{2} \cdot 2, \qquad \nabla_W^2 F(\mu; x, y) = (-a - b\cos(\pi m_\mu)) \cdot 1 \cdot 1.$$

Consequently,

$$\begin{aligned} &\nabla_W F(\mu, x) = a(x - m_\mu) - \beta\pi \sin(\pi m_\mu), \\ &\nabla_x \nabla_W F(\mu, x) = a, \qquad \nabla_W^2 F(\mu; x, y) = c(m_\mu), \qquad c(u) := -a - b\cos(\pi u). \end{aligned} \tag{196}$$

The Hessian kernel is symmetric and $|c(u)| \le a + b$. Its covariance kernel is $K_\nu(x,y) = \int c(m_\nu)^2\, \nu(dz) = c(m_\nu)^2$. The corresponding operator sends f to $c(m_\nu)^2 \int f\, d\nu$; it has eigenvalue $c(m_\nu)^2$ on the constant unit function and zero on its orthogonal complement, so is positive and trace class. Take a fresh $Z \sim N(0,1)$ independent of the full pre-jump history $\mathcal{H}_{1-}$ of lemma A.14. The spatially constant field $\xi_{1,\mu_{\tau-}}(x) = c(m_{\tau-})Z$ is jointly measurable and has the required conditional covariance.

For $e_N = N^{-1/2}\mathbf{1} \in \mathbb{R}^N$ and $\overline{x} = N^{-1}\sum_i x_i$, the finite factor is

$$\mathbf{M}^N(\rho_{\mathbf{x}}, \mathbf{x}) = \sqrt{N} c(\overline{x}) e_N e_N^\top, \qquad \mathbf{M}^N(\rho_{\mathbf{x}}, \mathbf{x}) g = c(\overline{x}) \mathbf{1} (e_N^\top g). \tag{197}$$

Here $\mathbf{1}$ has all entries equal to one. Use the coordinate $\mathbf{g}_1^N$ constructed in lemma A.14 and eq. (114): conditional on $\mathcal{H}_{1-}$ it has law $N(0, I_N)$. Thus $\zeta := e_N^\top \mathbf{g}_1^N$ has conditional law $N(0,1)$, since $e_N^\top e_N = 1$. Writing $\overline{X}_t = N^{-1}\sum_i X_t^i$, the jumps are

$$Y_\tau^i = Y_{\tau-}^i + \eta c(m_{\tau-}) Z, \qquad X_\tau^i = X_{\tau-}^i + \eta c(\overline{X}_{\tau-}) \zeta. \tag{198}$$

The population pushforward and conditional-i.i.d. property follow from lemma A.19 and lemma A.14 and eq. (115).

Step 3: trajectories and unconditional moments, for (iii). Averaging the negative gradient in (196) gives, on each ODE interval,

$$\begin{aligned} \dot{m}_t &= -a \int (x - m_t)\, \mu_t(dx) + \beta\pi \sin(\pi m_t) = \beta\pi \sin(\pi m_t), \\ \dot{\overline{X}}_t &= -\frac{a}{N} \sum_i (X_t^i - \overline{X}_t) + \beta\pi \sin\big(\pi \overline{X}_t\big) = \beta\pi \sin\big(\pi \overline{X}_t\big). \end{aligned} \tag{199}$$

These scalar ODEs are globally Lipschitz. Since $m_0 = 0$, uniqueness gives $m_t = 0$ before the jump; hence (198) yields

$$m_\tau = -\eta(a+b) Z. \tag{200}$$

Subtracting the mean equations from the particle equations gives

$$\frac{d}{dt}(Y_t^i - m_t) = -a(Y_t^i - m_t), \qquad \frac{d}{dt}(X_t^i - \overline{X}_t) = -a(X_t^i - \overline{X}_t).$$

Each translation in (198) also translates its mean, so $Y_\tau^i - m_\tau = Y_{\tau-}^i - m_{\tau-}$ and $X_\tau^i - \overline{X}_\tau = X_{\tau-}^i - \overline{X}_{\tau-}$. Solving the centered equations therefore gives global trajectories

$$\begin{aligned} Y_t^i &= m_t + e^{-at} U_i, \qquad X_t^i = \overline{X}_t + e^{-at}(U_i - \overline{X}_0), \\ X_t^i - Y_t^i &= \overline{X}_t - m_t - e^{-at} \overline{X}_0. \end{aligned} \tag{201}$$

They hold also at both sides of the jump. For $\dot{u} = \beta\pi \sin(\pi u)$, the drift is positive on $(2k, 2k+1)$ and negative on $(2k-1, 2k)$. Uniqueness prevents crossing an integer equilibrium, so a mean moves at most one unit on each ODE interval. There are two intervals and one jump. Using $|c| \le a + b$ and $|U_i|, |\overline{X}_0| \le R_0$ gives

$$\sup_{t\ge 0} |Y_t^i| \le 3R_0 + 2 + \eta(a+b)|Z|, \qquad \sup_{t\ge 0} |X_t^i| \le 3R_0 + 2 + \eta(a+b)|\zeta|.$$

For a standard normal Z, integration by parts, with vanishing boundary term, gives

$$\mathbb{E}Z^{2k} = -\frac{1}{\sqrt{2\pi}} \int z^{2k-1} \frac{d}{dz} e^{-z^2/2}\, dz = (2k-1)\mathbb{E}Z^{2k-2} = \prod_{j=1}^{k} (2j-1) \le (2k)^k.$$

Taking $k = \lceil q/2 \rceil$ for $q \ge 1$ yields $\|Z\|_{L^q} \le \|Z\|_{L^{2k}} \le \sqrt{2k} \le 2\sqrt{q}$; ζ has the same marginal bound. Minkowski's inequality now gives

$$\left\| \sup_{t\ge 0} |Y_t^i| \right\|_{L^q} \vee \left\| \sup_{t\ge 0} |X_t^i| \right\|_{L^q} \le 3R_0 + 2 + 2\eta(a+b)\sqrt{q} \le C\sqrt{q}.$$

This proves (iii) with $\alpha = 2$ and a fixed $K_0 \le C$.

Step 4: piecewise Lipschitz regularity, for (iv). By (201), the centered coordinates have absolute value at most $2R_0$. The gradient formula (196) gives $|\dot{X}_t^i| \vee |\dot{Y}_t^i| \le 2aR_0 + \beta\pi$. Integrating on an ODE interval gives $|X_t^i - X_s^i| \vee |Y_t^i - Y_s^i| \le C(t-s)$. The continuous ODE branches and the translations supply the required traces.

Step 5: anti-spikiness, for (v). The difference in (201) is independent of i. Thus (2.4) gives $r_t = |\overline{X}_t - m_t - e^{-at}\overline{X}_0|$, and

$$\frac{1}{N}\sum_i |X_t^i - Y_t^i|^3 = |\overline{X}_t - m_t - e^{-at}\overline{X}_0|^3 = r_t^3.$$

This includes $r_t = 0$ and both jump traces, proving (24) with $A_{\mathrm{sp}} = 1$ almost surely. □

Remark A.3 (Local intuition for the coupled jump)**.** *Recall $c(u) = -a - b\cos(\pi u)$ from lemma A.34 and eq.* (196) *and write $\overline{Y}_t = N^{-1}\sum_i Y_t^i$. On the constant direction e_N in* (197)*, the alignment in lemma A.14 gives*

$$\zeta = -\operatorname{sign}_+(c(\overline{Y}_{\tau-}))Z, \qquad c(u) = -(a+b) + O(u^2) \quad (u \to 0),$$

where $\operatorname{sign}_+(v) = -1$ for $v < 0$ and $+1$ otherwise. When both empirical means are sufficiently near zero, $c(\overline{Y}_{\tau-}) < 0$, so $\zeta = Z$ and the jumps in (198) *satisfy*

$$X_\tau^i - X_{\tau-}^i \simeq -\eta(a+b)Z = Y_\tau^i - Y_{\tau-}^i.$$

This explains locally why the coupled translations agree to leading order; the comparison proof below uses the general theorem, not this approximation.

Lemma A.35 (Curvature-response and population-objective bounds after escape)**.** *For $N \ge 2$ sufficiently large, there is a population-measurable event E_N^{esc} with $\mathbb{P}(E_N^{\mathrm{esc}}) \ge 1 - \frac{1}{2}N^{-1/16}$ on which the curvature-response factors in definition 3.1 and eq.* (178) *satisfy*

$$\sup_{t\ge 0}(\mathfrak{C}_t^Y + \mathfrak{J}_t^Y) \le CN^{1/16},$$
$$\mathfrak{C}_t^Y \le C, \qquad \mathfrak{J}_t^Y \le CN^{-1/4}, \qquad 0 \le F(\mu_t) + \beta \le CN^{-1/2} \quad (t \ge C\log N). \tag{202}$$

The constants depend only on the fixed model parameters.

Proof. **Step 1: exclude landings close to unstable equilibria.** By lemma A.34 and eq. (200), $m_\tau = -\sigma Z$ with $\sigma = \eta(a+b) > 0$. Its density is $f_\sigma(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/(2\sigma^2)}$. For $0 < \rho < 1/2$, symmetry and monotonicity on $[0, \infty)$ give

$$\mathbb{P}\{\operatorname{dist}(m_\tau, 2\mathbb{Z}) < \rho\} = \sum_{k\in\mathbb{Z}} \int_{2k-\rho}^{2k+\rho} f_\sigma(x)\, dx$$
$$\le 2\rho f_\sigma(0) + 4\rho \sum_{k\ge 1} f_\sigma(2k-1) \le 2\rho\{f_\sigma(0) + 1\}.$$

For the last inequality, $\sum_{k\ge1} f_\sigma(2k-1) \le \sum_{k\ge1}\int_{2k-2}^{2k-1} f_\sigma(x)\,dx \le 1/2$. Here $\mathrm{dist}(u,2\mathbb{Z}) = \inf_{k\in\mathbb{Z}}|u-2k|$. Choose ρ_N to be a sufficiently small fixed positive multiple of $N^{-1/16}$. Then

$$E_N^{\mathrm{esc}} := \{\mathrm{dist}(m_\tau, 2\mathbb{Z}) \ge \rho_N\}, \qquad \mathbb{P}((E_N^{\mathrm{esc}})^c) \le \tfrac{1}{2}N^{-1/16}. \tag{203}$$

This event depends only on the population jump. Work on it below.

Step 2: compute the curvature and its recovery time. Insert lemma A.34 and eq. (196) into the sampled full Hessian of definition 3.1 and eq. (11). With $e_N = N^{-1/2}\mathbf{1}$ as in (197),

$$\begin{aligned} \mathrm{H}_t^Y &= aI_N + c(m_t)e_N e_N^\top, \\ \mathrm{H}_t^Y e_N &= -b\cos(\pi m_t)e_N, \qquad \mathrm{H}_t^Y v = av \quad (e_N^\top v = 0), \\ \kappa_t^{Y,N} &= \min\{a, -b\cos(\pi m_t)\} \ge -b. \end{aligned} \tag{204}$$

The constant direction and its orthogonal complement are nontrivial since $N \ge 2$. Also $|\kappa_t^{Y,N}| \le \max\{a,b\}$, so it is integrable on every finite horizon. Let $d_t = \mathrm{dist}(m_t, 2\mathbb{Z}) \in [0,1]$ for $t \ge \tau$. On either branch $m_t = 2k \pm d_t$, lemma A.34 and eq. (199) gives $\dot d_t = \beta\pi\sin(\pi d_t) \ge 0$. If $d_\tau = 1$, the mean is an odd-integer equilibrium and remains there. Otherwise $0 < d_\tau < 1$ on E_N^{esc}. We use the half-angle coordinate $\tan(\pi d_t/2)$ to turn the sine drift into a linear equation and obtain an explicit curvature-recovery time. Direct differentiation gives

$$\begin{aligned} \frac{d}{dt}\tan\frac{\pi d_t}{2} &= \frac{\pi}{2\cos^2(\pi d_t/2)}\beta\pi\sin(\pi d_t) = b\tan\frac{\pi d_t}{2}, \\ \tan\frac{\pi d_t}{2} &= e^{b(t-\tau)}\tan\frac{\pi d_\tau}{2}. \end{aligned} \tag{205}$$

Set $t_N = \tau + b^{-1}\log(2/\rho_N)$ and $\lambda = \min\{a, b/2\} > 0$. Since $\tan(\pi\rho_N/2) \ge \rho_N$, $\tan(\pi d_{t_N}/2) \ge (2/\rho_N)\rho_N = 2 > \sqrt{3} = \tan(\pi/3)$. Hence $d_t \ge 2/3$ for $t \ge t_N$, including the equilibrium case, and $\cos(\pi m_t) = \cos(\pi d_t) \le -1/2$. Consequently (204) gives

$$\kappa_t^{Y,N} \ge \lambda \quad (t \ge t_N), \qquad e^{bt_N} = 2e^{b\tau}/\rho_N \le CN^{1/16}. \tag{206}$$

Step 3: split the curvature-response factors at the recovery time. Recall from (178) that $G_\kappa^Y(t,s) = e^{-\int_s^t \kappa_u^{Y,N}\,du}$, $\mathfrak{C}_t^Y = \int_0^t G_\kappa^Y(t,s)\,ds$, and, for this one jump, $\mathfrak{J}_t^Y = 0$ before τ and $\mathfrak{J}_t^Y = \eta G_\kappa^Y(t,\tau)$ after it. For $t \le t_N$, eqs. (204) and (206) give

$$\mathfrak{C}_t^Y \le \int_0^t e^{b(t-s)}\,ds = (e^{bt}-1)/b \le C/\rho_N, \qquad \mathfrak{J}_t^Y \le \eta e^{b(t-\tau)} \le C/\rho_N \quad (t \ge \tau).$$

For $t \ge t_N$, split the integral and use $G_\kappa^Y(t,s) = G_\kappa^Y(t,t_N)G_\kappa^Y(t_N,s)$ when $s \le t_N$. By (206),

$$\begin{aligned} \mathfrak{C}_t^Y &= G_\kappa^Y(t,t_N)\mathfrak{C}_{t_N}^Y + \int_{t_N}^t G_\kappa^Y(t,s)\,ds \le (C/\rho_N)e^{-\lambda(t-t_N)} + \lambda^{-1}, \\ \mathfrak{J}_t^Y &= G_\kappa^Y(t,t_N)\mathfrak{J}_{t_N}^Y \le (C/\rho_N)e^{-\lambda(t-t_N)}. \end{aligned} \tag{207}$$

The integral over $[t_N, t]$ satisfies

$$\int_{t_N}^t G_\kappa^Y(t,s)\,ds \le \int_{t_N}^t e^{-\lambda(t-s)}\,ds = \frac{1-e^{-\lambda(t-t_N)}}{\lambda} \le \lambda^{-1}.$$

These bounds imply the all-time estimate in (202). We choose the waiting time so that the exponential decay cancels the factor ρ_N^{-1} in the bound for $\mathfrak{J}_t^Y$ in eq. (207), leaving a bound of order $N^{-1/4}$:

$$t \ge t_N + \lambda^{-1}\log\left(\rho_N^{-1}N^{1/4}\right) \quad \Longrightarrow \quad e^{-\lambda(t-t_N)} \le \rho_N N^{-1/4}. \tag{208}$$

The threshold on the left is at most $C\log N$, because ρ_N is a fixed multiple of $N^{-1/16}$. Substitution into (207) gives $\mathfrak{C}_t^Y \le CN^{-1/4} + \lambda^{-1} \le C$ and $\mathfrak{J}_t^Y \le CN^{-1/4}$ in this range.

Step 4: decay of the population objective. The population trajectory in lemma A.34 and eq. (201) is the pushforward $u \mapsto m_t + e^{-at}u$, so $\mathrm{Var}(\mu_t) = e^{-2at}\,\mathrm{Var}(\mu_0)$. For $t \ge t_N$, we have $\cos(\pi m_t) \le$

$-1/2$, whence $\sin^2(\pi m_t) = (1-\cos(\pi m_t))(1+\cos(\pi m_t)) \geq 1+\cos(\pi m_t)$. Differentiate (192) using (199), and recall $b = \beta\pi^2$:

$$\begin{aligned}\frac{d}{dt}[F(\mu_t)+\beta] &= \tfrac{a}{2}[-2a\,\mathrm{Var}(\mu_t)] - \beta\pi\sin(\pi m_t)[\beta\pi\sin(\pi m_t)] \\ &= -a^2\,\mathrm{Var}(\mu_t) - b\beta\sin^2(\pi m_t) \\ &\leq -\lambda a\,\mathrm{Var}(\mu_t) - 2\lambda\beta(1+\cos(\pi m_t)) = -2\lambda[F(\mu_t)+\beta],\end{aligned}$$

where $\lambda \leq a$ and $2\lambda \leq b$. Multiplying by $e^{2\lambda(t-t_N)}$ and integrating, with $F(\mu_{t_N})+\beta \leq aR_0^2/2 + 2\beta \leq C$, yields

$$0 \leq F(\mu_t)+\beta \leq Ce^{-2\lambda(t-t_N)} \leq C\rho_N^2 N^{-1/2} \leq CN^{-1/2} \qquad (t \geq C\log N),$$

where the second inequality is (208) squared. □

Proof of proposition A.1. **Step 1: verify stability on the escape event.** Lemma A.34 verifies all comparison assumptions with uniform constants. Take $T = N^{1/8}$ and $p = \frac{1}{2}N^{-1/16}$, with N large enough that $T \geq \tau$. For one jump, the scale in (173) satisfies

$$\Xi = 16(1+B)N(e+T)/p \leq CN^{1+1/8+1/16} = CN^{19/16}.$$

The three powers used here have separate roles. The choice $\rho_N \asymp N^{-1/16}$ pays the escape failure probability in eq. (203); the recovery estimate eq. (206) then bounds the transient curvature-response factor growth by $\rho_N^{-1} = O(N^{1/16})$; and the horizon $T = N^{1/8}$ leaves a strictly negative power in the nonlinear stability check below. Use the residual and coefficient bounds in the proof of corollary A.4, specifically (188). Since $T/N = N^{-7/8} \leq N^{-1/2}$ and those bounds hold for every positive exponent, applying them with exponent $16\theta/19$ gives, for any $\theta > 0$,

$$\begin{aligned}&\delta + \widetilde{\delta} \leq C_\theta N^{-1/2+\theta}, \qquad \sqrt{\delta} + \sqrt{\widetilde{\delta}} \leq C_\theta N^{-1/4+\theta}, \\ &1 + c_D + c_J(D)q_p + \Lambda(D) + J_{N,T,p} \leq C_\theta N^\theta.\end{aligned} \tag{209}$$

Here the square-root bound uses $\sqrt{\delta}+\sqrt{\widetilde{\delta}} \leq \sqrt{2(\delta+\widetilde{\delta})}$ and $N^{\theta/2} \leq N^\theta$. The one-jump product in (179) is $J_{N,T,p} = 1 + \eta c_J(D)q_p$, which proves its bound with the same exponent. The other quantities retain the definitions in eqs. (173), (176) and (177).

Work on E_N^{esc} from lemma A.35 and eq. (203). Insert its all-time curvature-response bound (202) and (209) into the forcing (180) and radius $R = 2J_{N,T,p}\mathfrak{B}_*$:

$$\begin{aligned}\mathfrak{B}_* &\leq \delta \sup_{t\leq T}\mathfrak{C}_t^Y + c_J(D)q_p(\sqrt{\delta}+\sqrt{\widetilde{\delta}})\sup_{t\leq T}\mathfrak{J}_t^Y \\ &\leq C_\theta\left[N^{-1/2+\theta+1/16} + N^{-1/4+2\theta+1/16}\right] \leq C_\theta N^{-3/16+2\theta}, \\ R &\leq C_\theta N^\theta N^{-3/16+2\theta} = C_\theta N^{-3/16+3\theta}, \\ T(\delta + c_D R) &\leq C_\theta\left[N^{-3/8+\theta} + N^{-1/16+4\theta}\right] \leq C_\theta N^{-1/16+4\theta}.\end{aligned}$$

Fix $4\theta < \min\{1/16, \varepsilon\}$. For sufficiently large N, the last bound is at most $1/4$, exactly as required by the nonlinear stability event (181). Thus $E_N^{\mathrm{esc}} \subset \mathcal{S}_{N,T,p}$.

Step 2: apply comparison and prove (i). By lemma A.34(v), anti-spikiness holds almost surely. Curvature integrability follows from lemma A.35 and eq. (204). Apply theorem A.1 with the above p. Its proof constructs Ω_p so that $\sup_{t\leq T}\Gamma_t^N \leq \delta$ and the population and localization events hold there. On $\Omega_p \cap E_N^{\mathrm{esc}}$, eqs. (180), (182) and (209) give

$$\begin{aligned}r_t &\leq 2J_{N,T,p}[\delta\mathfrak{C}_t^Y + c_J(D)q_p(\sqrt{\delta}+\sqrt{\widetilde{\delta}})\mathfrak{J}_t^Y] \\ &\leq C_\theta N^{3\theta}[N^{-1/2}\mathfrak{C}_t^Y + N^{-1/4}\mathfrak{J}_t^Y].\end{aligned}$$

Then lemma A.33 and eq. (190) implies

$$\begin{aligned}r_t + |F(\rho_{\mathbf{X}_t}) - F(\mu_t)| &\leq (1+\Lambda(D))r_t + \delta \\ &\leq C_\theta N^{4\theta}[N^{-1/2}(1+\mathfrak{C}_t^Y) + N^{-1/4}\mathfrak{J}_t^Y].\end{aligned} \tag{210}$$

The all-time bound in lemma A.35 and eq. (202) makes the last line at most $C_\theta N^{4\theta}[N^{-1/2}(1+N^{1/16})+N^{-1/4+1/16}] \le C_\varepsilon N^{-3/16+\varepsilon}$, proving (i). A union bound gives $\mathbb{P}(\Omega_p \cap E_N^{\mathrm{esc}}) \ge 1-p-\frac{1}{2}N^{-1/16} = 1-N^{-1/16}$.

Step 3: substitute the late-time bounds to prove (ii). For $t \ge C\log N$, substitute $\mathfrak{C}_t^Y \le C$ and $\mathfrak{J}_t^Y \le CN^{-1/4}$ from lemma A.35 and eq. (202) into (210). Its right-hand side is at most $C_\theta N^{4\theta}[N^{-1/2}+N^{-1/4}N^{-1/4}] \le C_\varepsilon N^{-1/2+\varepsilon}$. Finally, on the same event,

$$0 \le F(\rho_{\mathbf{X}_t})+\beta \le |F(\rho_{\mathbf{X}_t})-F(\mu_t)|+F(\mu_t)+\beta.$$

The population bound in (202) is $CN^{-1/2}$, so adding this inequality proves (ii), after increasing C_ε. Increase N_0 if needed so that $C\log N < N^{1/8}$. □

A.9 A FIXED-TIME COMPARISON OF SQUARED GRADIENTS

We compare the squared-gradient statistics on paths already coupled under the prescribed common schedule of theorem A.1. The observation time is deterministic and lies inside an ODE interval. Define

$$\begin{aligned} \mathcal{G}_t &:= \int \|\nabla_W F(\mu_t,x)\|^2\,\mu_t(dx), \\ \mathcal{G}_t^N &:= \frac{1}{N}\sum_{i=1}^N \|\nabla_W F(\rho_{\mathbf{X}_t},X_t^i)\|^2. \end{aligned} \tag{211}$$

Recall of the comparison quantities. Let $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$ and K_{pop} be the population moment event and its deterministic moment scale from lemma A.20. Conditional on the population fields $\mathcal{G}_{\mathrm{pop}}$, the variables Y_t^i are i.i.d. with law μ_t; see lemma A.19. For $D \ge 1$, recall the sample localization event

$$E_{T,D} = \left\{ \sup_{s\le T}\max_{i\le N}(\|X_s^i\| \vee \|Y_s^i\|) \le D \right\}$$

from eq. (36). The common local coefficients in eqs. (37) and (44) and the drift comparison are

$$\mathrm{L}_D = \mathrm{L}_{K_{\mathrm{pop}},D}, \qquad \gamma_m(D) = \mathrm{L}_D a_m, \qquad \Lambda(D) = \sum_{m\ge1}\gamma_m(D)L_m(1+2D^{m-1}).$$

Here a_m are the moment weights and L_m are the feature derivative bounds in assumption A.1. The existing discrepancy and moment residual are

$$r_t = \left(\frac{1}{N}\sum_i \|X_t^i-Y_t^i\|^2\right)^{1/2}, \qquad \Gamma_t^N = \sum_{m\ge1}\gamma_m(D)\left\|\frac{1}{N}\sum_i \phi_m(Y_t^i) - \int \phi_m\,d\mu_t\right\|;$$

see section 2.4 and lemma A.10. Let B be the deterministic ODE velocity bound in assumption A.4.

Lemma A.36 (Fixed-time discrepancy of the squared-gradient statistic)**.** *Assume assumptions A.1 to A.4, with the prescribed-schedule coupling in definition A.1 and lemma A.14. The reference conditional product law is* $\mathrm{Law}((Y_t^i)_i \mid \mathcal{G}_{\mathrm{pop}}) = \mu_t^{\otimes N}$ *by lemma A.19. Conditional moments are bounded on* E_{pop} *by* (123)*;* $E_{T,D}$ *is the sample event* (36)*. The existing* B*-Lipschitz ODE condition holds almost surely. Fix a deterministic* $t \in (0,T)\setminus\{t_1,\dots,t_{L_{\mathrm{jump}}}\}$ *and* $\delta \in (0,1)$*. Except on an event whose intersection with* $E_{\mathrm{pop}} \cap E_{T,D}$ *has probability at most* δ*, one has*

$$|\mathcal{G}_t^N - \mathcal{G}_t| \le 2B\{(\Lambda(D)+\mathrm{L}_D)r_t+\Gamma_t^N\} + B^2\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{212}$$

Equivalently,

$$\mathbb{P}\left(E_{\mathrm{pop}}\cap E_{T,D}\cap\left\{|\mathcal{G}_t^N-\mathcal{G}_t| > 2B\{(\Lambda(D)+\mathrm{L}_D)r_t+\Gamma_t^N\} + B^2\sqrt{\frac{\log(2/\delta)}{2N}}\right\}\right) \le \delta.$$

Proof. **Step 1: the velocity bound controls the actual gradients.** We first work on E_{pop}, without imposing $E_{T,D}$. Recall the ODE in definition A.1 and the population moment continuity calculation in the proof of lemma A.12. For two nearby times s, t in the same ODE interval, integrate $D\phi_m$ along the segment joining the two particle positions. The first-derivative and conditional moment bounds give

$$\sum_m a_m \|M_m(\mu_s) - M_m(\mu_t)\| \le B|s-t| \sum_m a_m L_m \{1 + 2(K_{\mathrm{pop}} m^{1/\alpha})^{m-1}\} \longrightarrow 0.$$

The series is finite by the decay estimate lemma A.4. Choose a finite $R \ge 1$ containing the finitely many continuous particle paths in a neighborhood of t. On E_{pop}, $\mu_s \in \mathcal{P}_{\alpha,K_{\mathrm{pop}}}$ and $\rho_{\mathbf{X}_s} \in \mathcal{P}_R$, so their moment vectors belong to $\mathcal{V}_{K_{\mathrm{pop}},R}$ by eq. (18). The local moment and spatial estimates eqs. (45) and (48) then imply

$$\begin{aligned}&\|\nabla_W F(\mu_s, Y_s^i) - \nabla_W F(\mu_t, Y_t^i)\| \\ &\quad \le \mathrm{L}_{K_{\mathrm{pop}},R} \left\{ \sum_m a_m \|M_m(\mu_s) - M_m(\mu_t)\| + \|Y_s^i - Y_t^i\| \right\} \longrightarrow 0.\end{aligned}$$

For the empirical moments the corresponding estimate is

$$\sum_m a_m \|M_m(\rho_{\mathbf{X}_s}) - M_m(\rho_{\mathbf{X}_t})\| \le B|s-t| \sum_m a_m L_m (1 + 2R^{m-1}) \longrightarrow 0.$$

The same local estimates therefore give continuity of $\nabla_W F(\rho_{\mathbf{X}_s}, X_s^i)$ at t. The radius R is used only to establish this local continuity; it imposes no additional sample event.

Using the integral form of the reference ODE, for $h \downarrow 0$,

$$\frac{Y_{t+h}^i - Y_t^i}{h} = -\frac{1}{h} \int_t^{t+h} \nabla_W F(\mu_s, Y_s^i)\, ds \longrightarrow -\nabla_W F(\mu_t, Y_t^i).$$

Every difference quotient has norm at most B, so its limit does too. Applying the identical argument to X gives, on E_{pop},

$$\|\nabla_W F(\mu_t, Y_t^i)\| \le B, \qquad \|\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i)\| \le B.$$

The population bound extends to μ_t-almost every point. Indeed, conditional-i.i.d. sampling and $E_{\mathrm{pop}} \in \mathcal{G}_{\mathrm{pop}}$ give

$$\begin{aligned}0 &= \mathbb{E}\Big[\mathbf{1}_{E_{\mathrm{pop}}} \mathbf{1}_{\{\|\nabla_W F(\mu_t, Y_t^1)\| > B\}}\Big] \\ &= \mathbb{E}\left[\mathbf{1}_{E_{\mathrm{pop}}} \int \mathbf{1}_{\{\|\nabla_W F(\mu_t, x)\| > B\}}\, \mu_t(dx)\right].\end{aligned}$$

The integrand is nonnegative. Its integral is therefore zero for almost every population realization in E_{pop}, proving $\|\nabla_W F(\mu_t, x)\| \le B$ for μ_t-almost every x.

Step 2: compare the gradients at paired particles. On $E_{\mathrm{pop}} \cap E_{T,D}$, insert the reference empirical moments and use eqs. (45) and (48):

$$\begin{aligned}&\|\nabla_W F(\rho_{\mathbf{X}_t}, X_t^i) - \nabla_W F(\mu_t, Y_t^i)\| \\ &\le \sum_m \gamma_m(D) \left\| \frac{1}{N} \sum_j \phi_m(X_t^j) - \int \phi_m\, d\mu_t \right\| + \mathrm{L}_D \|X_t^i - Y_t^i\| \\ &\le \sum_m \gamma_m(D) \frac{1}{N} \sum_j \|\phi_m(X_t^j) - \phi_m(Y_t^j)\| + \Gamma_t^N + \mathrm{L}_D \|X_t^i - Y_t^i\| \\ &\le \Lambda(D) r_t + \Gamma_t^N + \mathrm{L}_D \|X_t^i - Y_t^i\|.\end{aligned}$$

For the last step, the line segment lies in the radius-D ball, so the first-derivative bound gives

$$\begin{aligned}\|\phi_m(X_t^j) - \phi_m(Y_t^j)\| &\le \int_0^1 \|D\phi_m(Y_t^j + s(X_t^j - Y_t^j))\|\, \|X_t^j - Y_t^j\|\, ds \\ &\le L_m(1 + 2D^{m-1}) \|X_t^j - Y_t^j\|.\end{aligned}$$

Here $1 + D^{m-1} \le 1 + 2D^{m-1}$ allows reuse of $\Lambda(D)$. Cauchy–Schwarz gives

$$\frac{1}{N}\sum_j \|X_t^j - Y_t^j\| \le \left(\frac{1}{N}\sum_j \|X_t^j - Y_t^j\|^2\right)^{1/2} = r_t.$$

For two vectors of norm at most B, $|\|u\|^2 - \|v\|^2| = |\langle u - v, u + v\rangle| \le 2B\|u - v\|$. Apply this inequality to the two actual gradients bounded in Step 1 and average the preceding estimate:

$$\left|\mathcal{G}_t^N - \frac{1}{N}\sum_i \|\nabla_W F(\mu_t, Y_t^i)\|^2\right| \le 2B\{(\Lambda(D) + \mathrm{L}_D)r_t + \Gamma_t^N\}. \tag{213}$$

No bound on a gradient evaluated at an intermediate measure–point pair is needed.

Step 3: concentration of the reference average. For almost every population realization in E_{pop}, the scalars $\|\nabla_W F(\mu_t, Y_t^i)\|^2$ are conditionally i.i.d., lie in $[0, B^2]$, and have conditional mean

$$\mathbb{E}[\|\nabla_W F(\mu_t, Y_t^i)\|^2 \mid \mathcal{G}_{\mathrm{pop}}] = \int \|\nabla_W F(\mu_t, x)\|^2 \,\mu_t(dx) = \mathcal{G}_t.$$

For $B > 0$, the bounded-differences inequality lemma A.18 applied to their average has coordinate sensitivities B^2/N. Consequently

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_i \|\nabla_W F(\mu_t, Y_t^i)\|^2 - \mathcal{G}_t\right| > B^2\sqrt{\frac{\log(2/\delta)}{2N}} \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right)$$
$$\le 2\exp\left\{-\frac{2N}{B^4}\left(B^2\sqrt{\frac{\log(2/\delta)}{2N}}\right)^2\right\} = \delta.$$

For $B = 0$, both statistics are zero on E_{pop} and the result holds directly. For $B > 0$, the triangle inequality and (213) show that failure of (212) inside $E_{\mathrm{pop}} \cap E_{T,D}$ requires failure of the reference-average bound. Thus its probability is at most

$$\mathbb{E}\left[\mathbf{1}_{E_{\mathrm{pop}}}\,\mathbb{P}\left(\left|\frac{1}{N}\sum_i \|\nabla_W F(\mu_t, Y_t^i)\|^2 - \mathcal{G}_t\right| > B^2\sqrt{\frac{\log(2/\delta)}{2N}} \,\middle|\, \mathcal{G}_{\mathrm{pop}}\right)\right] \le \delta\,\mathbb{P}(E_{\mathrm{pop}}) \le \delta.$$

The conditioning throughout is on the population fields, not on $E_{T,D}$. □

Remark A.4 (A threshold decision at the same fixed time)**.** *Fix a deterministic threshold $\theta \ge 0$. Suppose that* (212) *holds and*

$$|\mathcal{G}_t - \theta| > 2B\{(\Lambda(D) + \mathrm{L}_D)r_t + \Gamma_t^N\} + B^2\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{214}$$

If $\mathcal{G}_t > \theta$, then $\mathcal{G}_t^N \ge \mathcal{G}_t - |\mathcal{G}_t^N - \mathcal{G}_t| > \theta$. If $\mathcal{G}_t < \theta$, then $\mathcal{G}_t^N \le \mathcal{G}_t + |\mathcal{G}_t^N - \mathcal{G}_t| < \theta$. The decisions agree at this prescribed observation time. This comparison does not identify or synchronize adaptive jump times.

A.10 Finite-feature transfer of negative curvature at a fixed time

This fixed-time supplement is not an input to the tracking theorem, objective comparison, or fixed-time gradient comparison. Its finite-feature conclusion does not control whole-path curvature-response factors or adaptive jump times.

We give a sufficient sample size for a negative population Hessian direction to remain negative in the reference sample. The statement uses an exact finite-feature representation and a spatial-Hessian condition. We first fix a deterministic law; the conditional interpretation for a fixed time of PWGF follows after the proof.

Finite features and the two Hessian terms. Consider the finite-feature case of assumption A.1: for admissible laws ν,

$$F(\nu) = G\left(\int \phi \, d\nu\right), \qquad \phi : \mathbb{R}^d \longrightarrow \mathbb{R}^q, \qquad q \in \mathbb{N}. \tag{215}$$

The dimension q is fixed independently of N; no feature cutoff or approximation of F is taken. At the law μ under consideration, define

$$\Phi(x) := \mathrm{D}\phi(x) \in \mathbb{R}^{q\times d}, \qquad A := \mathrm{D}^2 G\left(\int \phi \, d\mu\right) = A^\top \in \mathbb{R}^{q\times q}. \tag{216}$$

The moment chain rule in eq. (54) reduces to the finite sum

$$\nabla_W^2 F(\mu; x, y) = \sum_{j,k=1}^{q} A_{jk} \nabla\phi_j(x) \nabla\phi_k(y)^\top = \Phi(x)^\top A \Phi(y).$$

Recall the integral and spatial Hessian operators from section 2.1. On $L^2(\mu; \mathbb{R}^d)$ they are

$$\begin{aligned} (H_\mu v)(x) &:= \int \Phi(x)^\top A \Phi(y) v(y) \, \mu(dy), \\ (H'_\mu v)(x) &:= \nabla_x \nabla_W F(\mu, x) v(x). \end{aligned} \tag{217}$$

Their sum is the full operator $\mathcal{H}_\mu = H_\mu + H'_\mu$. For independent $Y_1, \ldots, Y_N \sim \mu$, its sampled full Hessian is

$$\begin{aligned} \mathcal{H}^{Y,N}(i,j) &:= \frac{1}{N} \Phi(Y_i)^\top A \Phi(Y_j) + \delta_{ij} \nabla_x \nabla_W F(\mu, Y_i), \\ \kappa^{Y,N} &:= \lambda_{\min}(\mathcal{H}^{Y,N}). \end{aligned} \tag{218}$$

The matrix acts on $(\mathbb{R}^d)^N$ with its Euclidean inner product. At $\mu = \mu_t$, this is precisely the reference-sample curvature $\kappa_t^{Y,N}$ of (11): derivatives are evaluated at μ_t, not at the empirical measure.

Proposition A.2 (Transfer of negative curvature to the reference sample)**.** *Fix μ as above, and suppose that the following hold for deterministic constants $K \geq 1$, $C_s \geq 0$, $\varepsilon \geq 0$, and $\gamma > 0$.*

1. *The coefficient and feature moments satisfy*

$$\|A\|_{\mathbb{R}^q \to \mathbb{R}^q} \leq K, \qquad \mathbb{E}\|\Phi(Y_1)\|_F^4 \leq K^4. \tag{219}$$

   Here $\|\cdot\|_F$ is the Frobenius norm, the square root of the sum of squared entries.

2. *The spatial Hessian is controlled by the population gradient:*

$$\operatorname*{ess\,sup}_{x\sim\mu} \|\nabla_x \nabla_W F(\mu, x)\|_{\mathbb{R}^d \to \mathbb{R}^d} \leq C_s \|\nabla_W F(\mu, \cdot)\|_{L^2(\mu; \mathbb{R}^d)}. \tag{220}$$

3. *Near-stationarity and negative full curvature hold separately:*

$$\begin{aligned} \|\nabla_W F(\mu, \cdot)\|_{L^2(\mu;\mathbb{R}^d)} &\leq \varepsilon, \qquad C_s \varepsilon \leq \gamma/8, \\ \inf \mathrm{Spec}(H_\mu + H'_\mu) &\leq -\gamma. \end{aligned} \tag{221}$$

Then, for every $\delta \in (0, 1)$,

$$N \geq \frac{1}{\delta} \max\left\{1, \frac{12^4 K^{12}}{\gamma^4}\right\} \quad \Longrightarrow \quad \mathbb{P}(\kappa^{Y,N} \leq -\gamma/2) \geq 1 - \delta. \tag{222}$$

Condition (220) has the form of Assumption 2, equation (8), in Yamamoto et al. (2025). That work imposes the condition for every law in its setting; here it is required only at the law in the proposition. It does not follow merely from finite features or from a small gradient.

Proof. **Step 1: a single covariance event.** Define the population and sample feature covariances

$$\Sigma := \mathbb{E}[\Phi(Y_1)\Phi(Y_1)^\top], \qquad \widehat{\Sigma} := \frac{1}{N}\sum_{i=1}^{N} \Phi(Y_i)\Phi(Y_i)^\top. \tag{223}$$

They are positive semidefinite $q \times q$ matrices. If s_j are the singular values of a matrix Φ, then

$$\|\Phi\Phi^\top\|_F^2 = \sum_j s_j^4 \le \left(\sum_j s_j^2\right)^2 = \|\Phi\|_F^4.$$

The centered matrices $Z_i := \Phi(Y_i)\Phi(Y_i)^\top - \Sigma$ are independent and have mean zero. Expanding the squared Frobenius norm gives

$$\mathbb{E}\|Z_i\|_F^2 = \mathbb{E}\|\Phi(Y_i)\Phi(Y_i)^\top\|_F^2 - \|\Sigma\|_F^2 \le K^4.$$

For $i \neq j$, independence gives $\mathbb{E}\langle Z_i, Z_j\rangle_F = \langle \mathbb{E}Z_i, \mathbb{E}Z_j\rangle_F = 0$. Consequently

$$\begin{aligned}\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 &= \frac{1}{N^2}\sum_{i,j} \mathbb{E}\langle Z_i, Z_j\rangle_F \\ &= \frac{1}{N^2}\sum_i \mathbb{E}\|Z_i\|_F^2 \le \frac{K^4}{N}.\end{aligned}$$

Markov's inequality, applied to this squared norm, yields

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_F > \frac{K^2}{\sqrt{N\delta}}\right) \le \frac{K^4/N}{K^4/(N\delta)} = \delta.$$

Thus, with probability at least $1 - \delta$,

$$\|\widehat{\Sigma} - \Sigma\|_{\mathbb{R}^q \to \mathbb{R}^q} \le \|\widehat{\Sigma} - \Sigma\|_F \le \frac{K^2}{\sqrt{N\delta}}. \tag{224}$$

All subsequent sample calculations use (224). Since $\|\Phi\Phi^\top\|_{\mathbb{R}^q \to \mathbb{R}^q} \le \|\Phi\|_F^2$,

$$\|\Sigma\|_{\mathbb{R}^q \to \mathbb{R}^q} \le \mathbb{E}\|\Phi(Y_1)\|_F^2 \le (\mathbb{E}\|\Phi(Y_1)\|_F^4)^{1/2} \le K^2.$$

The sample-size condition gives $N\delta \ge 1$, and hence, on (224),

$$\|\widehat{\Sigma}\|_{\mathbb{R}^q \to \mathbb{R}^q} \le \|\Sigma\|_{\mathbb{R}^q \to \mathbb{R}^q} + \|\widehat{\Sigma} - \Sigma\|_{\mathbb{R}^q \to \mathbb{R}^q} \le K^2 + \frac{K^2}{\sqrt{N\delta}} \le 2K^2.$$

Step 2: compare the square roots and Hessian matrices. Continue on the covariance bound (224). Take a unit eigenvector v of the symmetric matrix $\widehat{\Sigma}^{1/2} - \Sigma^{1/2}$ with eigenvalue λ such that $|\lambda| = \|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\|_{\mathbb{R}^q \to \mathbb{R}^q}$. Here each square root is the positive semidefinite square root. The identity

$$\widehat{\Sigma} - \Sigma = \widehat{\Sigma}^{1/2}(\widehat{\Sigma}^{1/2} - \Sigma^{1/2}) + (\widehat{\Sigma}^{1/2} - \Sigma^{1/2})\Sigma^{1/2}$$

holds even when the two matrices do not commute. Positive semidefiniteness gives

$$v^\top(\widehat{\Sigma}^{1/2} + \Sigma^{1/2})v \ge |v^\top(\widehat{\Sigma}^{1/2} - \Sigma^{1/2})v| = |\lambda|.$$

Multiplying the preceding identity by $v^\top$ and v, and using the eigenvector equation on its two sides, gives

$$\begin{aligned}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{R}^q \to \mathbb{R}^q} &\ge |v^\top(\widehat{\Sigma} - \Sigma)v| \\ &= |\lambda|\, v^\top(\widehat{\Sigma}^{1/2} + \Sigma^{1/2})v \\ &\ge |\lambda|^2 = \|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\|_{\mathbb{R}^q \to \mathbb{R}^q}^2.\end{aligned}$$

Thus Step 1 implies $\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\|_{\mathbb{R}^q \to \mathbb{R}^q} \le K(N\delta)^{-1/4}$. Define the symmetric matrices reused in the eigenvalue comparison by

$$\mathcal{T} := \Sigma^{1/2} A \Sigma^{1/2}, \qquad \widehat{\mathcal{T}} := \widehat{\Sigma}^{1/2} A \widehat{\Sigma}^{1/2}.$$

Inserting and subtracting $\Sigma^{1/2}A\widehat{\Sigma}^{1/2}$ gives

$$\begin{aligned}\widehat{\mathcal{T}} - \mathcal{T} &= (\widehat{\Sigma}^{1/2} - \Sigma^{1/2})A\widehat{\Sigma}^{1/2} + \Sigma^{1/2}A(\widehat{\Sigma}^{1/2} - \Sigma^{1/2}),\\ \|\widehat{\mathcal{T}} - \mathcal{T}\|_{\mathbb{R}^q\to\mathbb{R}^q} &\le K(N\delta)^{-1/4}\,K\,(\sqrt{2}K + K)\\ &\le 3K^3(N\delta)^{-1/4} \le \gamma/4.\end{aligned}$$

The last inequality follows directly from $(N\delta)^{1/4} \ge 12K^3/\gamma$ in (222).

Step 3: identify the negative eigenvalues. Define the bounded feature operator and its adjoint by

$$\begin{aligned}&\mathcal{B} : L^2(\mu;\mathbb{R}^d) \to \mathbb{R}^q, \qquad \mathcal{B}v = \int \Phi(x)v(x)\,\mu(dx),\\ &\mathcal{B}^* : \mathbb{R}^q \to L^2(\mu;\mathbb{R}^d), \quad (\mathcal{B}^*u)(x) = \Phi(x)^\top u.\end{aligned}$$

Indeed, Cauchy–Schwarz gives

$$\begin{aligned}\|\mathcal{B}v\|_{\mathbb{R}^q} &\le \int \|\Phi(x)\|_F\|v(x)\|_{\mathbb{R}^d}\,\mu(dx)\\ &\le (\mathbb{E}\|\Phi(Y_1)\|_F^2)^{1/2}\|v\|_{L^2(\mu;\mathbb{R}^d)} \le K\|v\|_{L^2(\mu;\mathbb{R}^d)}.\end{aligned}$$

Moreover $\langle \mathcal{B}v, u\rangle_{\mathbb{R}^q} = \int \langle v(x), \Phi(x)^\top u\rangle_{\mathbb{R}^d}\,\mu(dx)$, which verifies the adjoint formula. By eqs. (217) and (223), for $u \in \mathbb{R}^q$ and $v \in L^2(\mu;\mathbb{R}^d)$, substitution gives

$$\begin{aligned}\mathcal{B}\mathcal{B}^*u &= \int \Phi(x)\Phi(x)^\top u\,\mu(dx) = \Sigma u,\\ (\mathcal{B}^*A\mathcal{B}v)(x) &= \Phi(x)^\top A\int \Phi(y)v(y)\,\mu(dy) = (H_\mu v)(x).\end{aligned}$$

Thus $\mathcal{B}\mathcal{B}^* = \Sigma$ and $H_\mu = \mathcal{B}^*A\mathcal{B}$. In particular H_μ is self-adjoint and has finite rank.

We use the following elementary eigenvector calculation for products of two bounded maps U, V between two spaces. For $\lambda \ne 0$,

$$\begin{aligned}&UVx = \lambda x, \quad x \ne 0 \implies Vx \ne 0, \quad VU(Vx) = \lambda Vx,\\ &VUy = \lambda y, \quad y \ne 0 \implies Uy \ne 0, \quad UV(Uy) = \lambda Uy.\end{aligned}$$

For the first implication, $Vx = 0$ would give $\lambda x = UVx = 0$, a contradiction; the second is proved in the same way. Apply these identities first with $U = \mathcal{B}^*$, $V = A\mathcal{B}$, then with $U = A\Sigma^{1/2}$, $V = \Sigma^{1/2}$:

$$\mathcal{B}^*A\mathcal{B} \longleftrightarrow A\Sigma \longleftrightarrow \Sigma^{1/2}A\Sigma^{1/2} = \mathcal{T} \quad \text{for nonzero eigenvalues.}$$

To check why only eigenvalues matter here, take $v \perp \operatorname{ran} H_\mu$. Self-adjointness implies, for every w,

$$\langle H_\mu v, w\rangle_{L^2(\mu;\mathbb{R}^d)} = \langle v, H_\mu w\rangle_{L^2(\mu;\mathbb{R}^d)} = 0, \qquad \text{hence } H_\mu v = 0.$$

The range is finite-dimensional and thus closed, so

$$L^2(\mu;\mathbb{R}^d) = \operatorname{ran} H_\mu \oplus (\operatorname{ran} H_\mu)^\perp, \qquad H_\mu = (H_\mu|_{\operatorname{ran} H_\mu}) \oplus 0.$$

The first restriction is a finite-dimensional self-adjoint matrix. Consequently any negative spectral bottom is one of its eigenvalues. This justifies using the displayed eigenvector correspondence for its spectral bottom whenever that bottom is negative.

For the sample, set $P = N^{-1/2}[\Phi(Y_1)\ \cdots\ \Phi(Y_N)] \in \mathbb{R}^{q\times dN}$. Matrix multiplication gives

$$PP^\top = \widehat{\Sigma}, \qquad (P^\top AP)(i,j) = N^{-1}\Phi(Y_i)^\top A\Phi(Y_j).$$

The same two product calculations, replacing $\mathcal{B}$ by P, show that the negative eigenvalues of $P^\top AP$ and $\widehat{\mathcal{T}}$ coincide. No assertion about the multiplicity of zero eigenvalues is needed.

Step 4: restore both spatial-Hessian terms. By eqs. (220) and (221),

$$\begin{aligned}\|H'_\mu v\|^2_{L^2(\mu;\mathbb{R}^d)} &\le (C_s\varepsilon)^2\|v\|^2_{L^2(\mu;\mathbb{R}^d)},\\ \|H'_\mu\|_{L^2(\mu;\mathbb{R}^d)\to L^2(\mu;\mathbb{R}^d)} &\le C_s\varepsilon \le \gamma/8.\end{aligned}$$

For every unit $v \in L^2(\mu;\mathbb{R}^d)$, subtracting the spatial term gives

$$\begin{aligned}\langle v, H_\mu v\rangle_{L^2(\mu;\mathbb{R}^d)} &= \langle v, (H_\mu + H'_\mu)v\rangle_{L^2(\mu;\mathbb{R}^d)} - \langle v, H'_\mu v\rangle_{L^2(\mu;\mathbb{R}^d)} \\ &\le \langle v, (H_\mu + H'_\mu)v\rangle_{L^2(\mu;\mathbb{R}^d)} + \gamma/8.\end{aligned}$$

Taking infima of the Rayleigh quotients yields

$$\inf \mathrm{Spec}(H_\mu) \le \inf \mathrm{Spec}(H_\mu + H'_\mu) + \gamma/8 \le -7\gamma/8.$$

The negative spectral bottom of H_μ is an eigenvalue. Step 3 therefore gives $\lambda_{\min}(\mathcal{T}) \le -7\gamma/8$. For a unit minimum-eigenvalue vector u of $\mathcal{T}$,

$$\lambda_{\min}(\widehat{\mathcal{T}}) \le u^\top \widehat{\mathcal{T}} u \le u^\top \mathcal{T} u + \|\widehat{\mathcal{T}} - \mathcal{T}\|_{\mathbb{R}^q \to \mathbb{R}^q} \le -7\gamma/8 + \gamma/4 = -5\gamma/8.$$

Its negative eigenvalue corresponds to one of $P^\top AP$ by Step 3.

The essential bound in (220) holds at every Y_i almost surely, since a finite union of μ-null sets still has sampling probability zero. The block-diagonal spatial matrix has norm at most $\max_i \|\nabla_x \nabla_W F(\mu, Y_i)\|_{\mathbb{R}^d \to \mathbb{R}^d} \le \gamma/8$. Evaluating the sampled full Hessian on a minimum-eigenvalue vector of $P^\top AP$ now gives

$$\kappa^{Y,N} \le \lambda_{\min}(P^\top AP) + \max_i \|\nabla_x \nabla_W F(\mu, Y_i)\|_{\mathbb{R}^d \to \mathbb{R}^d} \le -5\gamma/8 + \gamma/8 = -\gamma/2.$$

The only event with nonzero failure probability is the failure of (224), with probability at most δ. □

Fixed-time conditional interpretation. At a deterministic time t, condition on the population fields $\mathcal{G}_{\mathrm{pop}}$ from lemma A.19 and put $\mu = \mu_t$. Let $E \in \mathcal{G}_{\mathrm{pop}}$ be any event on which eqs. (219) to (221) hold with the same deterministic $q, K, C_s, \varepsilon, \gamma$. Here $\Phi = \mathrm{D}\phi$ is measure-independent, while $A = \mathrm{D}^2 G(\int \phi \, d\mu_t)$ and all Hessian terms are evaluated at the population law μ_t. For N as in (222), the conditional-i.i.d. property and the tower rule give

$$\mathbb{P}(E \cap \{\kappa_t^{Y,N} > -\gamma/2\}) = \mathbb{E}\Big[\mathbf{1}_E \mathbb{P}(\kappa_t^{Y,N} > -\gamma/2 \mid \mathcal{G}_{\mathrm{pop}})\Big] \le \delta\, \mathbb{P}(E) \le \delta.$$

This statement does not estimate $\mathbb{P}(E)$ or a whole-path curvature-response factor, and it concerns the reference sample Y, not the empirical dynamics X.

A simple instance of the assumptions. In (192), take $a = 0$ and a law of mean zero. The one-feature representation is $\phi(x) = x$ and $G(z) = \beta \cos(\pi z)$. Thus $\Phi = 1$, $A = -\beta\pi^2$, $\nabla_W F(\mu, x) = 0$, and $\nabla_x \nabla_W F(\mu, x) = 0$. The constant unit function is an eigenvector of H_μ with eigenvalue $-\beta\pi^2$. Hence the conditions hold with $K = \max\{1, \beta\pi^2\}$, $C_s = \varepsilon = 0$, and $\gamma = \beta\pi^2$. This observation does not impose the additional spatial-Hessian condition on the variance-plus-cosine example with general a.

B Product-measure WGF and Matrix Factorization

B.1 Tensor notation and Hermite tensors

We collect the tensor conventions used below. They are standard; see, for example, Hackbusch (2012) for tensor spaces. For Hermite polynomials, Gaussian chaos, and the Wiener chaos product formula, see Peccati & Taqqu (2011, Chapter 1-7).

Tensor notation. For an integer $r \ge 0$, write $(\mathbb{R}^d)^{\otimes r}$ for order-r tensors, with $(\mathbb{R}^d)^{\otimes 0} = \mathbb{R}$. For vectors $u^{(1)}, \dots, u^{(r)} \in \mathbb{R}^d$,

$$\big(u^{(1)} \otimes \cdots \otimes u^{(r)}\big)_{i_1 \dots i_r} := \prod_{s=1}^{r} u_{i_s}^{(s)}, \qquad x^{\otimes r} := x \otimes \cdots \otimes x.$$

For $A, B \in (\mathbb{R}^d)^{\otimes r}$, the Frobenius inner product and norm are

$$\langle A, B\rangle_{\mathrm{HS}} := \sum_{i_1, \dots, i_r = 1}^{d} A_{i_1 \dots i_r} B_{i_1 \dots i_r}, \qquad \|A\|_{\mathrm{HS}} := \sqrt{\langle A, A\rangle}.$$

If $A \in (\mathbb{R}^d)^{\otimes p}$, $B \in (\mathbb{R}^d)^{\otimes q}$, and $0 \le a \le \min(p,q)$, we write $A \otimes_a B \in (\mathbb{R}^d)^{\otimes(p+q-2a)}$ for contraction over a matched indices:

$$\big(A \otimes_a B\big)_{i_1\dots i_{p-a} j_1 \dots j_{q-a}} := \sum_{\ell_1,\dots,\ell_a=1}^{d} A_{i_1\dots i_{p-a}\ell_1\dots\ell_a} B_{\ell_1\dots\ell_a j_1\dots j_{q-a}}.$$

The case $a = 0$ is the tensor product $A \otimes B$, and a full contraction gives the Frobenius inner product. When A and B are symmetric, the particular choice of contracted modes is immaterial after symmetrization.

Symmetrization. For $T \in (\mathbb{R}^d)^{\otimes r}$, define

$$\mathrm{Sym}(T) := \frac{1}{r!} \sum_{\pi \in \mathfrak{S}_r} T^\pi, \qquad (T^\pi)_{i_1\dots i_r} := T_{i_{\pi(1)}\dots i_{\pi(r)}}.$$

The symmetric subspace is

$$\mathrm{Sym}^r(\mathbb{R}^d) := \{T \in (\mathbb{R}^d)^{\otimes r} : T^\pi = T \text{ for all } \pi \in \mathfrak{S}_r\}.$$

The map Sym is the orthogonal projection onto $\mathrm{Sym}^r(\mathbb{R}^d)$ under the Frobenius inner product. Since $x^{\otimes r}$ is symmetric,

$$\left\langle T, x^{\otimes r} \right\rangle = \left\langle \mathrm{Sym}(T), x^{\otimes r} \right\rangle.$$

Normalized Hermite basis and moment tensors. Let He_n be the probabilists' Hermite polynomial and set

$$\psi_n(t) := \frac{\mathrm{He}_n(t)}{\sqrt{n!}}, \qquad n \ge 0.$$

For $\sigma \in L^2(\mathcal{N}(0,1); \mathbb{R}^{d_h})$, we use the normalized Hermite expansion

$$\sigma(t) = \sum_{n \ge 0} a_n \psi_n(t), \qquad a_n := \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\sigma(g)\psi_n(g)] \in \mathbb{R}^{d_h}.$$

When the weights are supported on the unit sphere, define

$$T_{i,n} := \int w^{\otimes n} \, \mathrm{d}\mu_i(w) \in \mathrm{Sym}^n(\mathbb{R}^{d_w}), \qquad T^o_{i,n} := \int w^{\otimes n} \, \mathrm{d}\mu^o_i(w), \qquad i = 1, 2.$$

The unit-norm convention is important: $w^\top z$ is standard normal only when $z \sim \mathcal{N}(0, I_{d_w})$ and $\|w\| = 1$.

Hermite tensors. Let $\gamma_d(x) := (2\pi)^{-d/2} \exp\Big(-\|x\|^2/2\Big)$ be the standard Gaussian density on $\mathbb{R}^d$. We use the normalized order-r Hermite tensor

$$\mathcal{H}_r(x) := \frac{1}{\sqrt{r!}}(-1)^r \frac{\nabla^{\otimes r}\gamma_d(x)}{\gamma_d(x)} = \frac{1}{\sqrt{r!}} \exp\Big(\|x\|^2/2\Big)(-\nabla)^{\otimes r} \exp\Big(-\|x\|^2/2\Big) \in \mathrm{Sym}^r(\mathbb{R}^d).$$

Thus

$$\mathcal{H}_0(x) = 1, \qquad \mathcal{H}_1(x) = x, \qquad \mathcal{H}_2(x) = \frac{x^{\otimes 2} - I_d}{\sqrt{2}}.$$

For $u \neq 0$,

$$\left\langle \mathcal{H}_r(x), u^{\otimes r} \right\rangle = \|u\|^r \psi_r\left(\frac{u^\top x}{\|u\|}\right) = \frac{\|u\|^r}{\sqrt{r!}} \mathrm{He}_r\left(\frac{u^\top x}{\|u\|}\right). \tag{225}$$

In particular, if $\|u\| = 1$, then

$$\left\langle \mathcal{H}_r(x), u^{\otimes r} \right\rangle = \psi_r(u^\top x).$$

Thus, under the unit-sphere convention,

$$\int \sigma(w^\top z) \, \mathrm{d}\mu_i(w) = \sum_{n \ge 0} a_n \left\langle \mathcal{H}_n(z), T_{i,n} \right\rangle. \tag{226}$$

If the weights are not constrained to the sphere, write $w = \rho\bar{w}$ with $\rho = \|w\|$ and $\|\bar{w}\| = 1$. Then

$$\sigma(w^\top z) = \sum_{n\geq 0} a_n(\rho)\left\langle \mathcal{H}_n(z), \bar{w}^{\otimes n}\right\rangle, \qquad a_n(\rho) := \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\sigma(\rho g)\psi_n(g)].$$

Equivalently, using $w^{\otimes n}$ directly introduces the factor $\rho^{-n}a_n(\rho)$.

If $G \sim \mathcal{N}(0, I_d)$, then for $A \in (\mathbb{R}^d)^{\otimes r}$ and $B \in (\mathbb{R}^d)^{\otimes s}$,

$$\mathbb{E}\big[\langle A, \mathcal{H}_r(G)\rangle \langle B, \mathcal{H}_s(G)\rangle\big] = \mathbf{1}_{\{r=s\}} \langle \mathrm{Sym}(A), \mathrm{Sym}(B)\rangle . \tag{227}$$

In particular, for symmetric $A \in \mathrm{Sym}^r(\mathbb{R}^d)$ and $B \in \mathrm{Sym}^s(\mathbb{R}^d)$, the right-hand side is $\mathbf{1}_{\{r=s\}}\langle A, B\rangle$.

Wiener chaos product formula. For symmetric tensors $A \in \mathrm{Sym}^n(\mathbb{R}^{d_w})$ and $B \in \mathrm{Sym}^m(\mathbb{R}^{d_w})$,

$$\langle \mathcal{H}_n(z), A\rangle \langle \mathcal{H}_m(z), B\rangle = \sum_{r=0}^{\min(n,m)} c_{n,m,r} \langle \mathcal{H}_{n+m-2r}(z), \mathrm{sym}\,(A \otimes_r B)\rangle , \tag{228}$$

where

$$c_{n,m,r} := r!\binom{n}{r}\binom{m}{r}\sqrt{\frac{(n+m-2r)!}{n!\,m!}}. \tag{229}$$

The square-root factor in eq. (229) is exactly the conversion from the unnormalized Hermite tensors to the normalized ones. If $\widetilde{\mathcal{H}}_q := \sqrt{q!}\,\mathcal{H}_q$ is used for every order q, then the coefficient is $r!\binom{n}{r}\binom{m}{r}$.

Gaussian integration by parts gives the normalized tensor Stein identity

$$\mathbb{E}\big[f(G)\mathcal{H}_r(G)\big] = \frac{1}{\sqrt{r!}}\mathbb{E}\big[\nabla^{\otimes r} f(G)\big],$$

whenever the expectations are finite. Equivalently, the degree-r normalized Hermite coefficient tensor of $f \in L^2(\gamma_d)$ is

$$\widehat{f}_r := \mathbb{E}\big[f(G)\mathcal{H}_r(G)\big] \in \mathrm{Sym}^r(\mathbb{R}^d).$$

A useful special case for a one-dimensional activation σ is

$$\mathbb{E}\big[\sigma(w^\top G)\mathcal{H}_r(G)\big] = \frac{1}{\sqrt{r!}}\mathbb{E}\big[\sigma^{(r)}(w^\top G)\big]w^{\otimes r}, \qquad w \in \mathbb{R}^d,$$

under the same integrability assumptions.

B.2 Wasserstein gradient flow on product measures

Before analyzing the matrix-decomposition functional, we record a simple form of Wasserstein gradient flow restricted to product measures. This is useful when the optimization variable is constrained to remain factorized across parameter blocks, so that admissible perturbations move each marginal separately. Related product-space Wasserstein gradient-flow structures appear in multi-species PDEs and multi-distribution optimization; see, e.g., Chen et al. (2016); Giunta et al. (2024); Conger et al. (2025).

Product family and product metric. Let $\Omega_i \subset \mathbb{R}^{d_i}$ and $\Omega = \prod_{i=1}^m \Omega_i$. We consider the product family

$$\mathcal{M} := \left\{ \mu = \bigotimes_{i=1}^m \mu_i \;\middle|\; \mu_i \in \mathcal{P}_2(\Omega_i) \right\}. \tag{230}$$

For $\mu = \bigotimes_i \mu_i$ and $\nu = \bigotimes_i \nu_i$, we equip $\mathcal{M}$ with the product Wasserstein metric

$$W_{2,\mathcal{M}}(\mu,\nu)^2 := \sum_{i=1}^m W_2(\mu_i,\nu_i)^2. \tag{231}$$

An admissible tangent vector on $\mathcal{M}$ is a coordinate-wise vector field $v = (v_1, \dots, v_m)$ with $v_i \in L^2(\mu_i; \mathbb{R}^{d_i})$, identified with $v(x) = (v_1(x_1), \dots, v_m(x_m))$ for $x = (x_1, \dots, x_m)$. The product inner product is

$$\langle u, v \rangle_{\mu,\mathcal{M}} \coloneqq \sum_{i=1}^{m} \int_{\Omega_i} \langle u_i(x_i), v_i(x_i) \rangle \, \mu_i(\mathrm{d}x_i) = \int_{\Omega} \sum_{i=1}^{m} \langle u_i(x_i), v_i(x_i) \rangle \, \mu(\mathrm{d}x). \tag{232}$$

The corresponding product perturbation keeps the measure in $\mathcal{M}$:

$$(\mathrm{Id} + hv)_\sharp \mu = \bigotimes_{i=1}^{m} (\mathrm{Id}_i + hv_i)_\sharp \mu_i. \tag{233}$$

Restricted Wasserstein gradient. Let $\overline{F}(\mu_1, \dots, \mu_m) \coloneqq F(\bigotimes_i \mu_i)$. For $\mu = \bigotimes_i \mu_i$, write $\mu_{-i} = \bigotimes_{j \neq i} \mu_j$ and $x_{-i} = (x_j)_{j \neq i}$. If F admits a first variation, then the partial first variation of $\overline{F}$ with respect to μ_i is, up to an additive constant,

$$\frac{\delta \overline{F}}{\delta \mu_i}(\mu_1, \dots, \mu_m; x_i) = \int_{\Omega_{-i}} \frac{\delta F}{\delta \mu}(\mu; x_i, x_{-i}) \, \mu_{-i}(\mathrm{d}x_{-i}). \tag{234}$$

Consequently, the product Wasserstein gradient is

$$g_i^\mu(x_i) \coloneqq \nabla_{x_i} \frac{\delta \overline{F}}{\delta \mu_i}(\mu_1, \dots, \mu_m; x_i) = \int_{\Omega_{-i}} \left(\nabla_\mu F(\mu; x_i, x_{-i}) \right)_i \, \mu_{-i}(\mathrm{d}x_{-i}), \qquad i = 1, \dots, m, \tag{235}$$

whenever differentiation and integration can be interchanged. Equivalently, $g^\mu = (g_1^\mu, \dots, g_m^\mu)$ is the orthogonal projection of the full Wasserstein gradient $\nabla_\mu F(\mu)$ onto coordinate-wise product tangent directions. Indeed, for every admissible $v = (v_1, \dots, v_m)$,

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} F\big((\mathrm{Id} + hv)_\sharp \mu\big) = \sum_{i=1}^{m} \int_{\Omega_i} \langle g_i^\mu(x_i), v_i(x_i) \rangle \, \mu_i(\mathrm{d}x_i). \tag{236}$$

Restricted WGF. The Wasserstein gradient flow of F restricted to $\mathcal{M}$ is the product curve $\mu_t = \bigotimes_i \mu_{i,t}$ satisfying

$$\partial_t \mu_{i,t} + \nabla_{x_i} \cdot (v_{i,t} \mu_{i,t}) = 0, \qquad v_{i,t}(x_i) = -g_i^{\mu_t}(x_i), \qquad i = 1, \dots, m. \tag{237}$$

Equivalently, as a continuity equation on Ω, this reads

$$\partial_t \mu_t + \sum_{i=1}^{m} \nabla_{x_i} \cdot (v_{i,t}(x_i) \mu_t) = 0. \tag{238}$$

Formally, the characteristic chain rule gives the dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\sum_{i=1}^{m} \int_{\Omega_i} \| g_i^{\mu_t}(x_i) \|^2 \, \mu_{i,t}(\mathrm{d}x_i) \leq 0. \tag{239}$$

In the local matrix-factorization result where this identity is used, its regularity premises are imposed on the characteristic solution and the identity is justified for almost every time in proposition B.5 and eq. (349).

B.3 Overview of the nonlinear-moment landscape

Why matrix factorization is the testbed. The example in section 4 should be read as a compact matrix-factorization instance of a broader tensor-decomposition theme. Classical low-rank matrix models represent an object by a small number of latent factors, while tensor decompositions replace a two-way array by multiway rank-one components and their variants (Kolda & Bader, 2009; Sidiropoulos et al., 2017). Such tensor decompositions are not only algebraic objects: in latent-variable learning, observable second- and third-order moments often reduce parameter recovery to a structured symmetric tensor decomposition problem (Anandkumar et al., 2014). This makes matrix

and tensor factorization a natural source of non-convex objectives whose geometry is simple enough to analyze but rich enough to contain saddles, symmetries, and multiple factor branches.

The same family of examples has also played a central role in non-convex optimization theory. Low-rank matrix recovery and matrix factorization provide canonical settings where, under appropriate measurement and rank assumptions, local minima are global or nearly global and saddle points carry negative curvature (Bhojanapalli et al., 2016; Ge et al., 2017). Orthogonal tensor decomposition gives a parallel strict-saddle model with many symmetry related optima and saddles (Ge et al., 2015). These works motivate looking not only at first-order stationarity, but also at the second-variation geometry that decides whether a stationary point is stable or escapable.

Our population model transfers this matrix/tensor-factorization geometry to a mean-field product-measure setting. The two finite factor vectors are replaced by two mean-field feature maps

$$h_{\mu_i}(z) = \int \sigma(w^\top z)\,\mathrm{d}\mu_i(w), \qquad i = 1, 2,$$

and the fitting term compares the matrix-valued products

$$h_{\mu_1}(z)h_{\mu_2}(z)^\top \quad \text{and} \quad h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top.$$

Thus the non-convexity is not introduced by a finite-dimensional factor matrix, but by a nonlinear product map on probability measures. In the Gaussian-input spherical specialization, the Hermite expansion in lemma B.3 rewrites this product map in terms of moment tensors $T_{i,n} = \int w^{\otimes n}\,\mathrm{d}\mu_i(w)$. The fitting loss becomes a squared residual system for the tensor blocks $\Delta\mathcal{M}_p$, while the metric regularizer controls the Wasserstein distance to the antipodal anchor. This is the precise sense in which the appendix studies a nonlinear moment landscape: the active coordinates are tensor moments, but the actual dynamics remain Wasserstein dynamics of the underlying measures.

This perspective is close to the mean-field two-layer neural-network viewpoint, where wide-network training is described by the evolution of a parameter measure, often through Wasserstein gradient flow (Mei et al., 2018; Chizat & Bach, 2018; Nitanda et al., 2022a). It is also close to the matrix-decomposition example in Yamamoto et al. (2025), where a single mean-field feature map is used through the symmetric self-product $h_\mu(z)h_\mu(z)^\top$. The present product version separates the two factor measures and studies the product WGF on $\mu_1 \otimes \mu_2$. This separation is useful because the two natural branches

$$\mathcal{M}_- = \{\mu_2 = (-\mathrm{Id})_\# \mu_1\}, \qquad \mathcal{M}_+ = \{\mu_1 = \mu_2\}$$

become visible in the active odd moments. For $n \in \mathcal{O}$, where $\mathcal{O}$ is the finite set of active odd Hermite degrees, recall

$$T_{i,n} = \int w^{\otimes n}\,\mathrm{d}\mu_i(w), \qquad S_n = \frac{T_{1,n} + T_{2,n}}{2}, \qquad A_n = \frac{T_{1,n} - T_{2,n}}{2}$$

from assumption B.2 and eq. (306). Thus membership in $\mathcal{M}_-$ implies $\mathbf{S} = 0$, while membership in $\mathcal{M}_+$ implies $\mathbf{A} = 0$. These finite moment implications are not converses; see (308). The appendix establishes four local statements: attraction, transverse descent, and positive second variation in two separate regions. Each statement has its own hypotheses.

B.3.1 Summary of the section

This subsection summarizes the separate population-level statements behind the diagram in section 4. The local incoming branch, transverse descent direction, and positive-curvature regions are proved below. A trajectory connecting these regions, or a finite-sample escape guarantee, is not established by these geometric statements. Two conventions apply throughout. First, throughout this appendix the dynamics are the product Wasserstein gradient flow, namely the WGF restricted to product measures and product tangent fields, as recorded in section B.2. This should not be read as a claim that the unrestricted WGF on all joint measures preserves product structure. Second, $D^2_{W_2}F(\mu)[v, v]$ is the spherical product pushforward second variation, defined as the full EXP-pushforward derivative. It equals the quadratic form of the full spherical product Hessian $\mathcal{H}_{\mathbb{S},\mu}$ in lemma B.6 and eq. (301). This includes both the measure derivative and the spatial Hessian, also at stationary states. When the prescribed pushforward path is a product Wasserstein geodesic, this quantity is its geodesic second derivative; general admissible fields are not assumed to generate such a geodesic.

The model is fixed by the spherical product objective eq. (244). Under the cap and summability assumptions in assumption B.1, and the finite odd-block symmetric-teacher assumption in assumption B.2, the fitting term admits the Hermite tensor expansion of lemma B.3. Our Hermite-basis analysis follows a now-standard Gaussian-input toolkit in which activation coefficients directly parameterize optimization(Ben Arous et al., 2021; Abbe et al., 2023; Bietti et al., 2022) and kernel phenomena (Daniely et al., 2016; Murray et al., 2023), rather than merely serving as a formal expansion. The metric regularizer is

$$F_{\mathrm{reg}}(\mu_1,\mu_2)=\frac{\lambda_{\mathrm{reg}}}{p_W}Z_{\pm}(\mu_1,\mu_2)^{p_W/2},\qquad p_W=2+\delta_W>2,$$

where $Z_{\pm}$ is the squared product distance to the antipodal Dirac pair $\nu_\star^{\pm}=\frac{1}{2}\delta_{a_\star}+\frac{1}{2}\delta_{-a_\star}$. The fixed-assignment geometry in lemma B.1 and definition B.2 is used locally on the two caps. The choice $p_W>2$ is deliberate: the regularizer gives positive curvature on metric annuli with $Z_{\pm}>0$, but it is Hessian-flat at the antipodal Dirac target

$$\mu^{\dagger}\coloneqq\mu_\star^{\pm}=\nu_\star^{\pm}\otimes\nu_\star^{\pm}.$$

Thus the quadratic saddle structure at $\mu^{\dagger}$ is governed by the fitting term.

The finite odd Hermite block gives the active coordinates used to describe the landscape; their full definition is in section B.6 and eq. (306). The measure-level anti-symmetric branch $\mathcal{M}_-$ implies $\mathbf{S}=0$ on the active odd block, whereas the same-density branch $\mathcal{M}_+$ implies $\mathbf{A}=0$. These coordinates are diagnostic coordinates for the finite active fitting term, not a complete parametrization of the full metric-regularized objective. The antipodal target belongs to the active-origin intersection and is first-order stationary, as stated in proposition B.1. The same subsection also identifies the reference zero-fit teacher point on the same-density branch in proposition B.2.

Local attraction on the anti-symmetric branch. Reflection covariance of the full first variation, proved in lemma B.7, yields branch invariance up to cap exit in proposition B.3. The attraction proof in section B.7 then has three steps. Recall the assigned-pole cost $c_\star(w)=d_{\mathbb{S}}(w,\pi_\star(w))^2/2$ from eq. (252) and its outward gradient $q_\star(w)=\nabla_{\mathbb{S}}c_\star(w)=-\mathrm{LOG}_w\pi_\star(w)$ from eq. (256). The constants $C_\star$ and L_{fit} below depend only on the fixed activation and teacher; they are defined immediately before the pole contraction lemma in eqs. (328) and (329).

1. **An admissible pole contraction.** Lemma B.8 uses the spherical transport eq. (332) to show
$$DF_{\mathrm{fit}}(\mu)[(q_\star,q_\star)]\geq -C_\star Z_{\pm}(\mu)^2,\qquad \sup_w\|g_{i,\mathrm{fit}}(\mu,w)\|\leq L_{\mathrm{fit}}\sqrt{Z_{\pm}(\mu)}.$$
Here $g_{i,\mathrm{fit}}=\nabla_{\mathbb{S}}(\delta F_{\mathrm{fit}}/\delta\mu_i)$, as defined in eq. (326).
2. **A local radial inequality.** Adding the regularizer derivative gives
$$DF(\mu)[(q_\star,q_\star)]\geq\lambda_{\mathrm{reg}}Z_{\pm}(\mu)^{p_W/2}-C_\star Z_{\pm}(\mu)^2.$$
For $2<p_W<4$, if $C_\star Z_{\pm}(\mu)^{(4-p_W)/2}\leq\lambda_{\mathrm{reg}}/2$, then proposition B.4 and eq. (344) gives
$$DF(\mu)[(q_\star,q_\star)]\geq\frac{\lambda_{\mathrm{reg}}}{2}Z_{\pm}(\mu)^{p_W/2}.$$
The identity
$$\frac{\mathrm{d}}{\mathrm{d}t}Z_{\pm}(\mu_t)=-2DF(\mu_t)[(q_\star,q_\star)]$$
then controls distance while $\mu_t\in\mathcal{M}_-\cap\mathcal{X}_{\pm}(r_{\mathrm{cap}})$.
3. **Confinement and convergence from initial data.** For $2<p_W<3$, choose $z_\star$ by eq. (345). For every globally well-posed characteristic product WGF starting from $\mu_0\in\mathcal{M}_-\cap\mathcal{X}_{\pm}(r_{\mathrm{cap}})$ with $Z_{\pm}(\mu_0)\leq z_\star$, the mean-distance barrier and the pointwise fitting-field estimate give cap confinement, dissipation while away from the target, and $\mu_t\to\mu^{\dagger}$, by proposition B.5. Here globally well posed means existence for all times and uniqueness in the characteristic-solution class. The distance decay and support barrier are eqs. (347) and (348).

The stationary-state conclusion uses the same direct smallness condition in eq. (341). The distance and support comparisons do not require a density and apply equally to atomic initial measures.

Spherical product pushforward second variation at anti-symmetric states. The positive-curvature estimate is proved in section B.8. The fitting term may have negative curvature, but on an anti-symmetric active ball $\|\mathbf{A}\|_{\mathcal{H}_\mathcal{O}} \le R_{\mathrm{act}}$ its negative second variation is uniformly bounded below by lemma B.9. The metric part gives the compensating positive curvature: lemma B.10 lower bounds the spherical Hessian of the fixed squared-distance cost on the caps, and lemma B.11 lifts this to the regularizer. Combining the fitting lower bound with the metric lower bound gives the ambient second-variation bound proposition B.6. Its branchwise specialization is corollary B.2. This is the appendix-level estimate underlying the annular label in the main-text diagram: at fixed-assignment states with $Z_\pm \ge \zeta_{\min}^2$, every admissible product tangent field has positive spherical product pushforward second variation with an explicit domination condition

$$\lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}}\zeta_{\min}^{\delta_W} > C_{\mathrm{fit}}^{-}(R_{\mathrm{act}}).$$

Here C_{fit}^{-} is defined in (363). The field may point outside $\mathcal{M}_-$; global geodesic convexity of the annulus is not asserted.

The escape direction at the saddle. The escape direction at the antipodal saddle is constructed in section B.9. The teacher imbalance condition is formulated as a two-sided transverse moment condition in assumption B.3; the associated transverse moment measure is defined in definition B.1. The tangential visibility operator at an antipodal Dirac pair is introduced in definition B.5, and lemma B.13 turns the transverse teacher signal into a strictly positive escape coefficient. The resulting same-density perturbation is summarized in proposition B.7: for a suitable spherical pushforward path starting at $\mu^\dagger = \nu_\star^\pm \otimes \nu_\star^\pm$ at the fixed anchor $a = a_\star$,

$$\frac{\mathrm{d}}{\mathrm{d}h}\bigg|_{h=0} F(\mu_h) = 0, \qquad F(\mu_h) = F(\mu^\dagger) - h^2\gamma_{\mathrm{esc}} + o(h^2), \qquad \gamma_{\mathrm{esc}} > 0.$$

At the same time, because $Z_\pm(\mu_h) = O(h^2)$ and $p_W = 2 + \delta_W$, the metric regularizer contributes only $O(h^{2+\delta_W}) = o(h^2)$. Hence the quadratic escape curvature is not masked by the regularizer. This verifies an available same-density escape direction, not that every perturbation necessarily follows it.

Our escape argument adapts the zero-output saddle mechanism of Yamamoto et al. (2025) to the product-spherical setting by replacing affine interpolation with a teacher-visible same-density tangent perturbation.

Ambient curvature near the same-density teacher. Finally, section B.10 proves the local teacher-side curvature estimate. The branch $\mathcal{M}_+$ is invariant by exchange symmetry, as shown in proposition B.8. At the same-density teacher point $\mu_+^o = \nu^o \otimes \nu^o$, the fitting residual vanishes and its second variation is a nonnegative square. If $Z_o := Z_\pm(\mu_+^o) > 0$, the metric regularizer supplies a strictly positive margin. The proof of proposition B.9 keeps this margin positive in a sufficiently small product Wasserstein neighborhood relative to the fixed-assignment cap class. Its key nearby-state estimate is

$$D_{W_2}^2 F(\mu)[v,v] \ge \left[\lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}} Z_\pm(\mu)^{(p_W-2)/2} - C\|h_{\mu_1}h_{\mu_2}^\top - h_{\nu^o}h_{\nu^o}^\top\|_{L_z^2(\mathrm{F})}\right]\|v\|_{L^2(\mu;\mathbb{R}^{d_w})}^2,$$

where the proof-local $C < \infty$ depends only on the fixed activation. The result does not imply that the teacher is stationary for the regularized objective or that a trajectory escaping the saddle reaches this neighborhood.

B.4 Product spherical model and Hermite coordinates

Role of this subsection. This subsection fixes the product-measure model and the coordinate system used in the geometry argument below. We restrict the dynamics to a product ansatz $\mu_t = \mu_{1,t} \otimes \mu_{2,t}$. Thus the flow considered here is the product Wasserstein gradient flow on the two marginals, rather than the unrestricted WGF on all couplings. This convention is natural for two interacting mean-field networks and keeps each marginal on the sphere.

B.4.1 Product-measure objective and product WGF

We first define the product state space, the fitting loss, and the metric regularizer used in the curvature example.

We consider a product mean-field neural network whose two marginals live on the sphere. The fitting term compares the matrix-valued product of two mean-field neural networks, while a metric regularizer softly anchors each marginal to an antipodal Dirac pair.

The cap constraint fixes the mass split and makes the antipodal assignment unambiguous.

Define the cap class

$$0 < r_{\mathrm{cap}} < \frac{\pi}{2}, \qquad K_{\pm}(r_{\mathrm{cap}}) := \{w \in \mathbb{S}^{d_w-1} : d_{\mathbb{S}}(w, \pm a_\star) < r_{\mathrm{cap}}\}, \tag{240}$$

$$\mathcal{P}_{\pm}(r_{\mathrm{cap}}) := \left\{\mu \in \mathcal{P}(\mathbb{S}^{d_w-1}) : \operatorname{supp}\mu \subset K_+(r_{\mathrm{cap}}) \cup K_-(r_{\mathrm{cap}}),\ \mu(K_+) = \mu(K_-) = \frac{1}{2}\right\}. \tag{241}$$

The product ansatz treats the two factor distributions as separate Wasserstein variables. Define the product state space:

$$\mu = \mu_1 \otimes \mu_2 \in \mathcal{X}_{\pm}(r_{\mathrm{cap}}) \coloneqq \mathcal{P}_{\pm}(r_{\mathrm{cap}}) \times \mathcal{P}_{\pm}(r_{\mathrm{cap}}). \tag{242}$$

We will use the above product measure space to define the objective.

Let $\sigma : \mathbb{R} \to \mathbb{R}^{d_h}$ be the activation, applied componentwise in the mean-field neural network

$$h_\mu(z) \ = \ \int \sigma(w^\top z)\,\mathrm{d}\mu(w) \in \mathbb{R}^{d_h}. \tag{243}$$

The regularizer is built from spherical Wasserstein distance to a fixed antipodal Dirac pair. Define W_2 metric on the sphere as

$$W^2_{2,\mathbb{S}}(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int d_{\mathbb{S}}(w,y)^2\,\mathrm{d}\gamma(w,y).$$

Fix a unit vector $a_\star \in \mathbb{S}^{d_w-1}$, and define the antipodal Dirac target

$$\nu^{\pm}_\star \coloneqq \frac{1}{2}\delta_{a_\star} + \frac{1}{2}\delta_{-a_\star}.$$

Let $p_W \coloneqq 2 + \delta_W,\ \delta_W > 0$. For product measures $\mu = \mu_1 \otimes \mu_2,\ \mu^o = \mu_1^o \otimes \mu_2^o$, we set

$$Z_{\pm,i}(\mu_i) \coloneqq W^2_{2,\mathbb{S}}(\mu_i, \nu^{\pm}_\star), \qquad Z_{\pm}(\mu_1,\mu_2) \coloneqq Z_{\pm,1}(\mu_1) + Z_{\pm,2}(\mu_2).$$

The objective combines the matrix-factorization fitting loss with this metric regularizer:

$$F(\mu_1 \otimes \mu_2) := F_{\mathrm{fit}}(\mu_1,\mu_2) + F_{\mathrm{reg}}(\mu_1,\mu_2), \tag{244}$$

$$F_{\mathrm{fit}} := \frac{1}{2}\mathbb{E}_z\left[\|h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top - h_{\mu_1}(z)h_{\mu_2}(z)^\top\|^2_{\mathrm{F}}\right], \tag{245}$$

$$F_{\mathrm{reg}} := \frac{\lambda_{\mathrm{reg}}}{p_W} Z_{\pm}(\mu_1,\mu_2)^{p_W/2}. \tag{246}$$

Because the regularizer is metric rather than density-based, it is well-defined at singular targets. In particular, at $\mu^{\pm}_\star := \nu^{\pm}_\star \otimes \nu^{\pm}_\star$, one has $Z_{\pm} = 0$. Since $p_W > 2$, this regularizer is Hessian-flat at the antipodal Dirac target, so the local saddle curvature there comes from the fitting term.

Product WGF dynamics. When discussing dynamics, we use the product WGF associated with this objective. Namely, for a curve $\mu_t = \mu_{1,t} \otimes \mu_{2,t}$, each marginal evolves on the sphere by

$$\partial_t \mu_{i,t} + \nabla_{\mathbb{S}} \cdot (v_{i,t}\mu_{i,t}) = 0, \qquad v_{i,t}(w) = -\boldsymbol{\nabla}_{\mathbb{S}} \frac{\delta F}{\delta \mu_i}(\mu_{1,t},\mu_{2,t})(w), \qquad i = 1,2. \tag{247}$$

Here $\boldsymbol{\nabla}_{\mathbb{S}}$ denotes the spherical gradient, equivalently the Euclidean gradient projected by $I - ww^\top$.Equivalently, infinitesimal steps are spherical pushforwards $\mu_{i,t+h} \approx (w \mapsto \mathrm{EXP}_w(hv_{i,t}(w)))_{\#}\mu_{i,t}$. Thus the dynamics remain on the sphere by construction. This should be viewed as WGF on the product ansatz, not as a claim that the unrestricted WGF on all measures over the product space preserves product form. In discrete or perturbed steps we use spherical pushforwards $w \mapsto \mathrm{EXP}_w(\eta v_i(w))$, so each marginal remains supported on $\mathbb{S}^{d_w-1}$. For a product tangent vector $v = (v_1, v_2)$, we use

$$\|v\|^2_{L^2(\mu)} := \|v_1\|^2_{L^2(\mu_1)} + \|v_2\|^2_{L^2(\mu_2)}.$$

We use product WGF, so each marginal follows its own spherical Wasserstein gradient flow.

B.4.2 HERMITE AND MOMENT COORDINATES

The Gaussian input allows us to rewrite the fitting term in orthogonal Hermite coordinates. Throughout, we use the normalized Hermite basis

$$\psi_n(t) := \frac{\mathrm{He}_n(t)}{\sqrt{n!}}, \qquad \sigma(t) = \sum_{n\geq 0} a_n \psi_n(t), \qquad a_n := \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\sigma(g)\psi_n(g)] \in \mathbb{R}^{d_h}. \tag{248}$$

We also define the moment tensors

$$T_{i,n} := \int w^{\otimes n}\, \mathrm{d}\mu_i(w) \in \mathrm{Sym}^n(\mathbb{R}^{d_w}), \qquad T^o_{i,n} := \int w^{\otimes n}\, \mathrm{d}\mu^o_i(w), \qquad i = 1, 2. \tag{249}$$

For each $n \geq 0$, we let $\mathcal{H}_n(z) \in \mathrm{Sym}^n(\mathbb{R}^{d_w})$ be the multivariate Hermite tensor.

The Hermite coordinates diagonalize the Gaussian-input L^2 geometry of the fitting term.

B.4.3 STANDING ASSUMPTIONS

We now list the model-specific assumptions used only for this geometric verification.

The first assumption keeps the Hermite expansion controlled and all measures in the fixed-cap regime.

Assumption B.1 (Product spherical cap class and Hermite summability)**.** *We assume:*

1. *The activation has fourth-order weighted square-summable Hermite coefficients:*

$$\sum_{n\geq 0} n^4 \|a_n\|_2^2 < \infty.$$

2. *For $i = 1, 2$, the student and teacher marginals satisfy*

$$\mu_i, \mu^o_i \in \mathcal{P}_{\pm}(r_{\mathrm{cap}}).$$

The cap condition places the model in a fixed-assignment regime where spherical logarithm/exponential maps and the metric regularizer have explicit second variations. In the calculations below we also use the finite odd-block assumption, stated next, to justify differentiating the polynomial Gaussian-input outputs and their products.

The second assumption isolates a finite odd Hermite block and imposes symmetry between the two teacher factors.

Assumption B.2 (Finite odd-block activation and symmetric teacher setting)**.** *There exists a finite set $\mathcal{O} \subset 2\mathbb{Z}_{\geq 0} + 1$ such that*

$$a_n \neq 0 \quad (n \in \mathcal{O}), \qquad a_n = 0 \quad (n \notin \mathcal{O}).$$

Moreover, the teacher is symmetric across the two factors:

$$\mu^o_1 = \mu^o_2 =: \nu^o, \qquad T^o_n := T^o_{1,n} = T^o_{2,n} \quad (n \in \mathcal{O}).$$

The finite odd block is a simplifying active-coordinate assumption. Odd degrees separate the antisymmetric and same-density branches because antipodal reflection sends $T_n(\nu)$ to $(-1)^n T_n(\nu)$, while the symmetric teacher assumption isolates the product symmetry responsible for the saddle mechanism.

The escape direction at the antipodal point is generated by a transverse first-order imbalance of the teacher. To express this imbalance without choosing one of the two poles $+a$ or $-a$, we slice the sphere by the two-sided latitude variable

$$r_a(w) = |a^\top w| \in [0, 1].$$

On each slice $\{w : |a^\top w| = r\}$, we measure the average transverse component $\Pi_{a^\perp} w$. The following vector-valued measure records these sliced transverse first moments.

Definition B.1 (Two-sided transverse teacher moment measure)**.** *Fix $a_\star \in \mathbb{S}^{d_w-1}$, and let*

$$r_{a_\star}(w) \coloneqq |a_\star^\top w|, \qquad \Pi_{a_\star^\perp} \coloneqq I - a_\star a_\star^\top.$$

For every Borel set $E \subset [0,1]$, define the $a_\star^\perp$-valued finite measure

$$\mathfrak{m}^o_{a_\star}(E) \coloneqq \int_{\{w:\, r_{a_\star}(w)\in E\}} \Pi_{a_\star{}^\perp} w \,\mathrm{d}\nu^o(w) \in a_\star{}^\perp.$$

We call $\mathfrak{m}^o_{a_\star}$ the two-sided transverse teacher moment measure around the antipodal axis $\{\pm a_\star\}$.

Equivalently, if

$$\bar\nu^o_{a_\star} \coloneqq (r_{a_\star})_\# \nu^o$$

and

$$\nu^o = \int_0^1 \nu^o_{a_\star,r} \,\mathrm{d}\bar\nu^o_{a_\star}(r)$$

is a disintegration over the two-sided latitude sets

$$\Sigma^\pm_{a_\star}(r) \coloneqq \{w \in \mathbb{S}^{d_w-1} : |a_\star{}^\top w| = r\},$$

then

$$\mathfrak{m}^o_{a_\star}(E) = \int_E m^o_{a_\star}(r) \,\mathrm{d}\bar\nu^o_{a_\star}(r),$$

where

$$m^o_{a_\star}(r) \coloneqq \int_{\Sigma^\pm_{a_\star}(r)} \Pi_{a_\star{}^\perp} w \,\mathrm{d}\nu^o_{a_\star,r}(w) \in a_\star{}^\perp.$$

Thus $\mathfrak{m}^o_{a_\star}(E)$ is the transverse barycenter of the teacher mass whose two-sided latitude lies in E. When $d_w = 3$, the sets $\{|a_\star{}^\top w| = r\}$ are pairs of latitude circles; in higher dimension they are two-sided latitude spheres. The measure $\mathfrak{m}^o_{a_\star}$ records whether, on these slices, the teacher has a coherent first-order bias in some direction orthogonal to the antipodal axis.

The final assumption is a nondegeneracy condition for the escape direction. It requires the sliced transverse teacher moment to have a coherent component along one visible direction $e_\star$: after projecting each two-sided latitude slice onto $a_\star^\perp$, the teacher signal does not cancel across slices and is quantitatively positive away from the equator. This positive transverse margin is what drives the same-density perturbation escaping the antipodal saddle.

Assumption B.3 (Two-sided transverse imbalance)**.** *Let $\nu^o \coloneqq \mu^o_1 = \mu^o_2$. There exist a unit vector $e_\star \in a_\star{}^\perp$, a radius $r_{\mathrm{vis}} \in (0,1)$, and a constant $\gamma_\perp > 0$, such that the scalar signed measure*

$$\mathfrak{m}^o_{a_\star,e_\star}(E) \coloneqq \big\langle e_\star, \mathfrak{m}^o_{a_\star}(E)\big\rangle$$

is a positive finite measure on $[0,1]$. Equivalently,

$$\left\langle e_\star, \int_{\{w:\, |a_\star^\top w|\in E\}} \Pi_{a_\star{}^\perp} w \,\mathrm{d}\nu^o(w)\right\rangle \ge 0$$

for every Borel set $E \subset [0,1]$. Moreover, the imbalance is quantitatively strict on the visible latitudes:

$$\mathfrak{m}^o_{a_\star,e_\star}([r_{\mathrm{vis}},1]) = \left\langle e_\star, \int_{\{w:\, |a_\star{}^\top w|\in[r_{\mathrm{vis}},1]\}} \Pi_{a_\star{}^\perp} w \,\mathrm{d}\nu^o(w)\right\rangle \ge \gamma_\perp.$$

Remark B.1 (Interpretation of the transverse imbalance condition)**.** *The condition does not require pointwise asymmetry of the teacher density, but it requires $d_w \ge 2$ and excludes global point-reflection symmetry. Indeed, if $\nu^o = (-\mathrm{Id})_\#\nu^o$, then the integration domain is invariant under $w \mapsto -w$, whereas $\Pi_{a_\star^\perp} w$ is odd, and hence $\mathfrak{m}^o_{a_\star}(E) = 0$ for every Borel set $E \subset [0,1]$, contradicting $\gamma_\perp > 0$. It only asks for a one-dimensional signed projection of the sliced transverse moment to be nonnegative across all two-sided latitudes and nonzero on a visible range. Equivalently, the transverse slice barycenters*

$$m^o_{a_\star}(r) = \int_{\{|a_\star{}^\top w|=r\}} \Pi_{a_\star{}^\perp} w \,\mathrm{d}\nu^o_{a_\star,r}(w)$$

have a coherent component along $e_\star$. This prevents cancellations of the transverse teacher signal across latitudes and supplies the direction used later to construct a same-density escape perturbation from the antipodal saddle.

B.4.4 FIXED-ASSIGNMENT METRIC GEOMETRY

We next record the local spherical geometry of the antipodal metric regularizer.

All fixed-assignment formulas below are used locally, along spherical pushforward perturbations whose supports remain in the same two caps.

The cap separation makes the optimal assignment to the two Dirac atoms fixed.

Lemma B.1 (Fixed-assignment formula for the antipodal Dirac target)**.** ***Assumptions used.*** *assumption B.1.*

Let $\mu \in \mathcal{P}_{\pm}(r_{\mathrm{cap}})$. Then the optimal transport from μ to

$$\nu_\star^{\pm} = \frac{1}{2}\delta_{a_\star} + \frac{1}{2}\delta_{-a_\star}$$

is the fixed map $\pi_\star$ defined as

$$\pi_\star(w) \coloneqq \begin{cases} a_\star, & w \in K_+(r_{\mathrm{cap}}), \\ -a_\star, & w \in K_-(r_{\mathrm{cap}}), \end{cases} \tag{250}$$

Consequently,

$$W_{2,\mathbb{S}}^2(\mu, \nu_\star^{\pm}) = \int_{\mathbb{S}} d_{\mathbb{S}}(w, \pi_\star(w))^2 \,\mathrm{d}\mu(w) = 2\int_{\mathbb{S}} c_\star(w)\,\mathrm{d}\mu(w), \tag{251}$$

where the fixed-assignment cost is

$$c_\star(w) \coloneqq \frac{1}{2} d_{\mathbb{S}}(w, \pi_\star(w))^2. \tag{252}$$

The same identity remains valid for sufficiently small smooth pushforward perturbations of μ whose support stays in the same two caps.

Proof. For $w \in K_+(r_{\mathrm{cap}})$, one has

$$d_{\mathbb{S}}(w, a_\star) < r_{\mathrm{cap}} < \frac{\pi}{2}, \qquad d_{\mathbb{S}}(w, -a_\star) = \pi - d_{\mathbb{S}}(w, a_\star) > \frac{\pi}{2}.$$

Thus every point in the positive cap is strictly closer to $a_\star$ than to $-a_\star$. The analogous statement holds on the negative cap. Since the mass of each cap is $1/2$, the fixed assignment sends exactly the correct mass to each atom. Any alternative assignment would move some positive-cap mass to $-a_\star$, or some negative-cap mass to $a_\star$, increasing the transport cost. This proves optimality and the displayed formula. □

The following notation describes geodesic contraction toward the assigned antipodal atom.

Definition B.2. *The tangent space at $y \in \mathbb{S}^{d_w-1}$ is $T_y\mathbb{S}^{d_w-1} = \{\xi \in \mathbb{R}^{d_w} : y^\top \xi = 0\}$. The logarithm encodes a shortest spherical arc as a tangent vector; the exponential follows the arc specified by a tangent vector. If $y, w \in \mathbb{S}^{d_w-1}$ and $d_{\mathbb{S}}(y, w) < \pi$, we write*

$$\theta \coloneqq d_{\mathbb{S}}(y, w) = \arccos\big(y^\top w\big),$$

and define

$$\mathrm{LOG}_y w \coloneqq \begin{cases} \theta\, \dfrac{w - \cos\theta\, y}{\sin\theta}, & w \neq y, \\ 0, & w = y. \end{cases} \tag{253}$$

Conversely, for $\xi \in T_y\mathbb{S}^{d_w-1}$,

$$\mathrm{EXP}_y(\xi) \coloneqq \begin{cases} \cos(\|\xi\|)y + \sin(\|\xi\|)\dfrac{\xi}{\|\xi\|}, & \xi \neq 0, \\ y, & \xi = 0. \end{cases} \tag{254}$$

The vector $\mathrm{LOG}_y w$ has length θ, and

$$\mathrm{EXP}_y(\mathrm{LOG}_y w) = w \quad (d_{\mathbb{S}}(y, w) < \pi), \qquad \mathrm{LOG}_y(\mathrm{EXP}_y \xi) = \xi \quad (\|\xi\| < \pi).$$

At the antipode $-y$, there is no unique shortest direction, so the logarithm is not defined there. The exponential is defined for all tangent vectors, but the inverse identity just stated requires $\|\xi\| < \pi$. Since $r_{\mathrm{cap}} < \pi/2$, for every $w \in K_+(r_{\mathrm{cap}}) \cup K_-(r_{\mathrm{cap}})$, the vector

$$\xi_\star(w) \coloneqq \mathrm{LOG}_{\pi_\star(w)} w$$

is uniquely defined and satisfies

$$\|\xi_\star(w)\| = d_{\mathbb{S}}(w, \pi_\star(w)) < r_{\mathrm{cap}}.$$

We define the fixed-assignment spherical contraction toward the antipodal Dirac target by

$$C_s(w) \coloneqq \mathrm{EXP}_{\pi_\star(w)}\big(s\xi_\star(w)\big), \qquad 0 \le s \le 1.$$

Thus

$$C_1(w) = w, \qquad C_0(w) = \pi_\star(w), \qquad d_{\mathbb{S}}(C_s(w), \pi_\star(w)) = s\, d_{\mathbb{S}}(w, \pi_\star(w)).$$

The corresponding tangent vector at the current point is

$$q_\star(w) \coloneqq \frac{\partial}{\partial s}\Big|_{s=1} C_s(w) \in T_w \mathbb{S}^{d_w - 1},$$

with

$$\|q_\star(w)\| = d_{\mathbb{S}}(w, \pi_\star(w)).$$

The direction $-q_\star$ is the infinitesimal contraction toward the antipodal Dirac target. Moreover, because $\pi_\star(-w) = -\pi_\star(w)$, one has

$$C_s(-w) = -C_s(w), \qquad q_\star(-w) = -q_\star(w).$$

Lemma B.2 (Fixed-assignment contraction geometry)**.** *Let $w \in K_+(r_{\mathrm{cap}}) \cup K_-(r_{\mathrm{cap}})$, set $y = \pi_\star(w)$, and write $r = d_{\mathbb{S}}(w, y)$. If $r > 0$, define*

$$u := \frac{w - \cos r\, y}{\sin r}.$$

Then $y^\top u = 0$, $\|u\|_{\mathbb{R}^{d_w}} = 1$, and $w = \cos r\, y + \sin r\, u$. The contraction in definition B.2 satisfies

$$\begin{aligned} C_s(w) &= \cos(sr) y + \sin(sr) u, \\ \partial_s C_s(w) &= r[-\sin(sr) y + \cos(sr) u], && \|\partial_s C_s(w)\|_{\mathbb{R}^{d_w}} = r, \\ \partial_s^2 C_s(w) &= -r^2 C_s(w), && (I - C_s(w) C_s(w)^\top)\partial_s^2 C_s(w) = 0. \end{aligned} \tag{255}$$

Moreover,

$$q_\star(w) = \nabla_{\mathbb{S}} c_\star(w) = -\mathrm{LOG}_w \pi_\star(w), \tag{256}$$

and $q_\star(-w) = -q_\star(w)$. At $r = 0$, both sides of eq. (256) *are zero, and the identity is understood by continuous extension.*

Proof. For $r > 0$, using $w^\top y = \cos r$ gives

$$y^\top u = \frac{\cos r - \cos r}{\sin r} = 0, \qquad \|u\|^2_{\mathbb{R}^{d_w}} = \frac{1 - 2\cos^2 r + \cos^2 r}{\sin^2 r} = 1.$$

Thus $w = \cos r\, y + \sin r\, u$. Substituting this decomposition into eqs. (253) and (254) yields the first line of eq. (255); differentiating it gives the remaining lines. In particular, the ambient acceleration is $-r^2 C_s(w)$, whereas its tangent projection, the covariant acceleration on the sphere, is zero.

On either cap, $y = \pi_\star(w)$ is fixed and $r = \arccos\big(w^\top y\big)$. Therefore

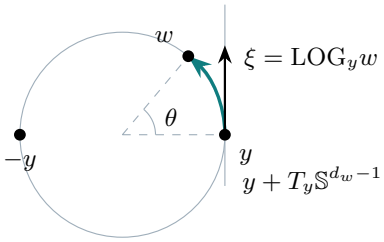
$$\nabla_{\mathbb{S}} r = -\frac{(I - ww^\top)y}{\sin r} = \frac{\cos r\, w - y}{\sin r}, \qquad \nabla_{\mathbb{S}} c_\star(w) = \frac{r}{\sin r}(\cos r\, w - y).$$

Substituting $w = \cos r\, y + \sin r\, u$ into the last expression gives

$$\frac{r}{\sin r}\big[\cos r(\cos r\, y + \sin r\, u) - y\big] = r[-\sin r\, y + \cos r\, u] = \partial_s C_s(w)|_{s=1}.$$

The formula for $\mathrm{LOG}_w y$ in eq. (253) gives the same vector with the opposite sign, proving eq. (256). Finally, $\pi_\star(-w) = -\pi_\star(w)$ and the explicit formulas give $C_s(-w) = -C_s(w)$ and $q_\star(-w) = -q_\star(w)$. If $r = 0$, the contraction is constant, while the smooth squared-distance cost has zero gradient; no differentiability of the unsquared distance at the pole is used. □

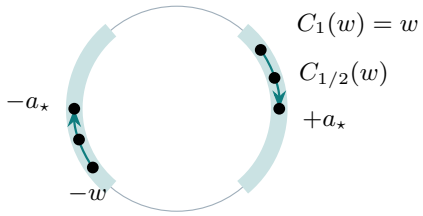
(a) Point ↔ tangent vector



$\|\xi\| = \theta < \pi, \quad \mathrm{EXP}_y \xi = w$
The antipode $-y$ is excluded from LOG.

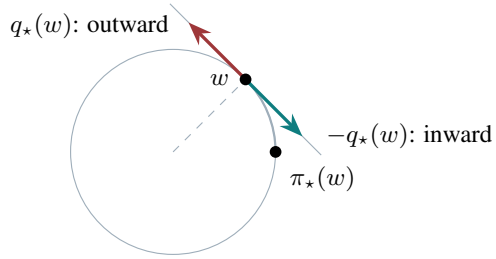
(b) Contraction in two fixed caps

Each cap has radius $r_{\mathrm{cap}} < \pi/2$.



$C_0(w) = \pi_\star(w), \quad C_1(w) = w$
Arrows: s decreases from 1 to 0.

(c) Opposite tangent directions at the endpoint



$q_\star(w) = \nabla_{\mathbb{S}} c_\star(w) = -\mathrm{LOG}_w \pi_\star(w)$
The inward vector is $\mathrm{LOG}_w \pi_\star(w)$.

Figure 3: Spherical logarithms, exponentials, and pole contraction (definition B.2 and lemma B.2). (a) A point w at distance $\theta < \pi$ from y corresponds to the tangent vector $\mathrm{LOG}_y w$ of length θ; EXP_y maps it back to w. Vectors are drawn with their tails at their base points. (b) Decreasing s from 1 to 0 in $C_s(w)$ moves each point to its assigned pole and preserves its cap. (c) At the endpoint w, $q_\star(w) = \nabla_{\mathbb{S}} c_\star(w) = -\mathrm{LOG}_w \pi_\star(w)$ points outward; its negative points toward the pole. The cost is $c_\star(w) = d_{\mathbb{S}}(w, \pi_\star(w))^2/2$. Each circle is a great-circle cross section of the unit sphere; these geometric paths do not represent a WGF trajectory.

B.4.5 TENSORIZED OBJECTIVE

We now rewrite the fitting loss as a sum of squared tensor residuals.

The sphere constraint gives a uniform bound on all moment tensors.

Remark B.2. *Under assumption B.1, the sphere constraint implies*

$$\|T_{i,n}\|_{\mathrm{F}} \le 1, \qquad \|T^o_{i,n}\|_{\mathrm{F}} \le 1, \qquad i = 1, 2,$$

for every $n \ge 0$. Under assumption B.2, the Hermite-moment expansions used below are finite sums.

The next lemma is the algebraic bridge from the neural-network fitting loss to finite tensor residuals.

Lemma B.3 (Hermite expansion of the metric-regularized objective)**.** ***Assumptions used.*** *assumptions B.1 and B.2.*

Define

$$\mathcal{M}_p \coloneqq \sum_{\substack{n,m\ge 0\\ 0\le r\le \min(n,m)\\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left(T_{1,n} \otimes_r T_{2,m}\right), \tag{257}$$

$$\mathcal{M}^o_p \coloneqq \sum_{\substack{n,m\ge 0\\ 0\le r\le \min(n,m)\\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left(T^o_{1,n} \otimes_r T^o_{2,m}\right), \tag{258}$$

$$c_{n,m,r} \coloneqq r!\binom{n}{r}\binom{m}{r}\sqrt{\frac{(n+m-2r)!}{n!\,m!}}, \qquad \Delta\mathcal{M}_p \coloneqq \mathcal{M}_p - \mathcal{M}^o_p. \tag{259}$$

These residuals have both output-matrix indices and spatial tensor indices: $\mathcal{M}_p, \mathcal{M}^o_p, \Delta\mathcal{M}_p \in \mathbb{R}^{d_h\times d_h} \otimes \mathrm{Sym}^p(\mathbb{R}^{d_w})$. For such a tensor B, write $B_{\alpha\beta}$ for its spatial tensor at output indices (α,β). For $U \in \mathrm{Sym}^p(\mathbb{R}^{d_w})$ and $A \in \mathbb{R}^{d_h\times d_h}$, the two partial pairings mean

$$\begin{aligned} \langle U, B\rangle_{\alpha\beta} &\coloneqq \langle U, B_{\alpha\beta}\rangle_{\mathrm{HS}}, && \langle U, B\rangle \in \mathbb{R}^{d_h\times d_h},\\ \langle B, A\rangle_{\mathrm{F}} &\coloneqq \sum_{\alpha,\beta=1}^{d_h} A_{\alpha\beta}B_{\alpha\beta}, && \langle B, A\rangle_{\mathrm{F}} \in \mathrm{Sym}^p(\mathbb{R}^{d_w}),\\ \|B\|^2_{\mathrm{HS}} &\coloneqq \sum_{\alpha,\beta=1}^{d_h} \|B_{\alpha\beta}\|^2_{\mathrm{HS}}. \end{aligned} \tag{260}$$

Thus pairing with $\mathcal{H}_p(z)$ contracts the spatial indices, whereas pairing with $a_n a_m^\top$ contracts the output indices. Then

$$F(\mu) = \frac{1}{2}\sum_{p\ge 0}\|\Delta\mathcal{M}_p\|^2_{\mathrm{HS}} + \frac{\lambda_{\mathrm{reg}}}{p_W} Z_{\pm}(\mu_1,\mu_2)^{p_W/2}. \tag{261}$$

The first term is the Hermite expansion of the fitting loss, and the second term is the antipodal Dirac metric regularizer from (244)*.*

Remark B.3. *Under assumption B.2, this tensor residual system becomes finite dimensional. The sums defining $\mathcal{M}_p$ and $\mathcal{M}^o_p$ reduce to indices $n, m \in \mathcal{O}$.*

Proof. **Step 1: multiply the finite Hermite expansions.** Throughout the proof, all Hermite tensors are understood with the normalization used in eqs. (227) and (228), so that different homogeneous chaoses are orthogonal in L^2 under the law of z.

For $i = 1, 2$, the Hermite expansion of the feature map gives

$$h_{\mu_i}(z) = \sum_{n\ge 0} a_n \left\langle \mathcal{H}_n(z), T_{i,n}\right\rangle, \qquad T_{i,n} = \int_{\mathbb{S}} w^{\otimes n}\,\mathrm{d}\mu_i(w). \tag{262}$$

Hence the matrix-valued product can be written as

$$h_{\mu_1}(z)h_{\mu_2}(z)^\top = \sum_{n,m\geq 0} a_n a_m^\top \langle \mathcal{H}_n(z), T_{1,n}\rangle \langle \mathcal{H}_m(z), T_{2,m}\rangle. \tag{263}$$

We now apply the Wiener-chaos product formula. For symmetric tensors $A \in (\mathbb{R}^{d_w})^{\otimes n}$ and $B \in (\mathbb{R}^{d_w})^{\otimes m}$, one has

$$\langle \mathcal{H}_n(z), A\rangle \langle \mathcal{H}_m(z), B\rangle = \sum_{r=0}^{\min(n,m)} c_{n,m,r} \Big\langle \mathcal{H}_{n+m-2r}(z), \operatorname{sym}(A \otimes_r B)\Big\rangle, \tag{264}$$

where

$$c_{n,m,r} = r!\binom{n}{r}\binom{m}{r}\sqrt{\frac{(n+m-2r)!}{n!m!}}.$$

Substituting $A = T_{1,n}$ and $B = T_{2,m}$, we obtain

$$h_{\mu_1}(z)h_{\mu_2}(z)^\top = \sum_{n,m\geq 0} \sum_{r=0}^{\min(n,m)} c_{n,m,r}(a_n a_m^\top) \Big\langle \mathcal{H}_{n+m-2r}(z), \operatorname{sym}(T_{1,n} \otimes_r T_{2,m})\Big\rangle. \tag{265}$$

Collecting all terms with the same Hermite degree $p = n + m - 2r$, we can rewrite this as

$$h_{\mu_1}(z)h_{\mu_2}(z)^\top = \sum_{p\geq 0} \langle \mathcal{H}_p(z), \mathcal{M}_p\rangle, \tag{266}$$

where the coefficient tensor is precisely

$$\mathcal{M}_p = \sum_{\substack{n,m\geq 0 \\ 0\leq r\leq \min(n,m) \\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \operatorname{sym}(T_{1,n} \otimes_r T_{2,m}). \tag{267}$$

The same calculation applied to the teacher measures gives

$$h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top = \sum_{p\geq 0} \langle \mathcal{H}_p(z), \mathcal{M}_p^o\rangle, \tag{268}$$

with

$$\mathcal{M}_p^o = \sum_{\substack{n,m\geq 0 \\ 0\leq r\leq \min(n,m) \\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \operatorname{sym}\left(T_{1,n}^o \otimes_r T_{2,m}^o\right). \tag{269}$$

Therefore

$$h_{\mu_1}(z)h_{\mu_2}(z)^\top - h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top = \sum_{p\geq 0} \langle \mathcal{H}_p(z), \Delta\mathcal{M}_p\rangle, \qquad \Delta\mathcal{M}_p = \mathcal{M}_p - \mathcal{M}_p^o. \tag{270}$$

Step 2: use orthogonality to sum the residual squares. We now take the L^2-norm in z, with Frobenius norm in the matrix output. Since distinct Hermite chaoses are orthogonal, all cross terms with $p \neq q$ vanish:

$$\begin{aligned}
&\mathbb{E}_z\Big[\big\|h_{\mu_1}(z)h_{\mu_2}(z)^\top - h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top\big\|_{\mathrm{F}}^2\Big] \\
&\quad= \sum_{p,q\geq 0} \mathbb{E}_z\Big[\Big\langle \langle \mathcal{H}_p(z), \Delta\mathcal{M}_p\rangle, \langle \mathcal{H}_q(z), \Delta\mathcal{M}_q\rangle\Big\rangle_{\mathrm{F}}\Big] \\
&\quad= \sum_{p\geq 0} \|\Delta\mathcal{M}_p\|_{\mathrm{HS}}^2.
\end{aligned} \tag{271}$$

Consequently,

$$\frac{1}{2}\mathbb{E}_z\Big[\big\|h_{\mu_1^o}(z)h_{\mu_2^o}(z)^\top - h_{\mu_1}(z)h_{\mu_2}(z)^\top\big\|_{\mathrm{F}}^2\Big] = \frac{1}{2}\sum_{p\geq 0} \|\Delta\mathcal{M}_p\|_{\mathrm{HS}}^2. \tag{272}$$

It remains only to identify the regularization term. By definition of $Z_\pm$, the metric regularization term in (244) is exactly

$$\frac{\lambda_{\mathrm{reg}}}{p_W} Z_\pm(\mu_1, \mu_2)^{p_W/2}.$$

This proves (261). □

Bridge to the geometry overview. This tensorized form is the coordinate system used in the curvature arguments below. Thus the model is reduced to a tensor residual system: the fitting loss is the squared Hilbert-Schmidt norm of the residuals $\Delta\mathcal{M}_p$, while the metric term controls distance to the antipodal pair. In the next subsection, the active odd moments are recombined into symmetric and anti-symmetric coordinates to describe the curvature transition.

B.5 Ambient first-order formulas

This subsection records the first-order ambient formulas used to analyze branch invariance, descent directions, and spherical Wasserstein gradients.

We first differentiate the tensorized fitting loss with respect to the active moment tensors.

Lemma B.4 (Moment derivatives of the fitting part and first variation of F)**.** ***Assumptions used.*** *assumptions B.1 and B.2.*

Let $\Delta\mathcal{M}_p$ be as in lemma B.3. The output contraction in (260) *is the spatial tensor*

$$\langle \Delta\mathcal{M}_p, a_n a_m^\top\rangle_{\mathrm{F}} = \sum_{\alpha,\beta=1}^{d_h} (a_n)_\alpha (a_m)_\beta (\Delta\mathcal{M}_p)_{\alpha\beta}.$$

Define the active fitting coefficients

$$G_{1,n}^{\mathrm{fit}} \coloneqq \frac{\partial F_{\mathrm{fit}}}{\partial T_{1,n}}, \qquad G_{2,m}^{\mathrm{fit}} \coloneqq \frac{\partial F_{\mathrm{fit}}}{\partial T_{2,m}}, \tag{273}$$

where

$$F_{\mathrm{fit}} \coloneqq \frac{1}{2}\sum_{p\geq 0} \|\Delta\mathcal{M}_p\|_{\mathrm{HS}}^2. \tag{274}$$

Then

$$G_{1,n}^{\mathrm{fit}} = \sum_{m\geq 0}\sum_{r\leq \min(n,m)} c_{n,m,r}\mathrm{sym}\left(\langle \Delta\mathcal{M}_{n+m-2r}, a_n a_m^\top\rangle_{\mathrm{F}} \otimes_{m-r} T_{2,m}\right), \tag{275}$$

$$G_{2,m}^{\mathrm{fit}} = \sum_{n\geq 0}\sum_{r\leq \min(n,m)} c_{n,m,r}\mathrm{sym}\left(\langle \Delta\mathcal{M}_{n+m-2r}, a_n a_m^\top\rangle_{\mathrm{F}} \otimes_{n-r} T_{1,n}\right). \tag{276}$$

The first variations of the fitting terms are

$$\frac{\delta F_{\mathrm{fit}}}{\delta\mu_1}(w) = \sum_{n\geq 0}\langle G_{1,n}^{\mathrm{fit}}, w^{\otimes n}\rangle, \tag{277}$$

$$\frac{\delta F_{\mathrm{fit}}}{\delta\mu_2}(w) = \sum_{m\geq 0}\langle G_{2,m}^{\mathrm{fit}}, w^{\otimes m}\rangle. \tag{278}$$

Proof. **Step 1: differentiate the tensor residuals.** For the data-fit term,

$$F_{\mathrm{fit}} \coloneqq \frac{1}{2}\sum_{p\geq 0} \|\Delta\mathcal{M}_p\|_{\mathrm{HS}}^2, \tag{279}$$

we have

$$\delta F_{\mathrm{fit}} = \sum_{p\geq 0}\langle \Delta\mathcal{M}_p, \delta\mathcal{M}_p\rangle_{\mathrm{HS}}. \tag{280}$$

If only $T_{1,n}$ is varied, then

$$\delta\mathcal{M}_p = \sum_{\substack{m\geq 0\\ r\leq\min(n,m)\\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top)\otimes \mathrm{sym}\left(\delta T_{1,n}\otimes_r T_{2,m}\right). \tag{281}$$

Therefore

$$\delta F_{\mathrm{fit}} = \sum_{\substack{p,m,r \\ p=n+m-2r}} c_{n,m,r} \Big\langle \Delta\mathcal{M}_p, (a_n a_m^\top) \otimes \mathrm{sym}\left(\delta T_{1,n} \otimes_r T_{2,m}\right) \Big\rangle_{\mathrm{HS}} \tag{282}$$

$$= \sum_{\substack{p,m,r \\ p=n+m-2r}} c_{n,m,r} \Big\langle \langle \Delta\mathcal{M}_p, a_n a_m^\top \rangle_{\mathrm{F}}, \mathrm{sym}\left(\delta T_{1,n} \otimes_r T_{2,m}\right) \Big\rangle_{\mathrm{HS}}. \tag{283}$$

Step 2: contract the output and moment indices. Set

$$B_{p,n,m,r} \coloneqq \langle \Delta\mathcal{M}_p, a_n a_m^\top \rangle_{\mathrm{F}} \in \mathrm{Sym}^{n+m-2r}(\mathbb{R}^{d_w}). \tag{284}$$

Then, componentwise,

$$\Big\langle B_{p,n,m,r}, \delta T_{1,n} \otimes_r T_{2,m} \Big\rangle_{\mathrm{HS}} = \sum_{I,J,K} (B_{p,n,m,r})_{I,J} (\delta T_{1,n})_{I,K} (T_{2,m})_{K,J} \tag{285}$$

$$= \Big\langle B_{p,n,m,r} \otimes_{m-r} T_{2,m}, \delta T_{1,n} \Big\rangle_{\mathrm{HS}}, \tag{286}$$

where I has length $n-r$, J has length $m-r$, and K has length r. Since $\delta T_{1,n}$ is symmetric, pairing against it only depends on the symmetric part of the tensor in the first slot, so

$$\Big\langle B_{p,n,m,r}, \mathrm{sym}\left(\delta T_{1,n} \otimes_r T_{2,m}\right) \Big\rangle_{\mathrm{HS}} = \Big\langle \mathrm{sym}\left(B_{p,n,m,r} \otimes_{m-r} T_{2,m}\right), \delta T_{1,n} \Big\rangle_{\mathrm{HS}}. \tag{287}$$

Substituting this into the previous display yields

$$\delta F_{\mathrm{fit}} = \left\langle \sum_{m\geq 0} \sum_{r\leq \min(n,m)} c_{n,m,r} \mathrm{sym}\left(\langle \Delta\mathcal{M}_{n+m-2r}, a_n a_m^\top \rangle_{\mathrm{F}} \otimes_{m-r} T_{2,m}\right), \delta T_{1,n} \right\rangle_{\mathrm{HS}}, \tag{288}$$

which is exactly the stated formula for the fit part of $G^{\mathrm{fit}}_{1,n}$. The formula for $G^{\mathrm{fit}}_{2,m}$ follows by the same argument with the two factors exchanged.

Step 3: recover the first measure variations. For a signed marginal perturbation $\delta\mu_i$, $\delta T_{i,n} = \int w^{\otimes n}\, \delta\mu_i(dw)$, so

$$\delta F_{\mathrm{fit}} = \sum_n \langle G^{\mathrm{fit}}_{i,n}, \delta T_{i,n} \rangle = \int \sum_n \langle G^{\mathrm{fit}}_{i,n}, w^{\otimes n} \rangle\, \delta\mu_i(dw).$$

This proves the stated first variations, up to constants that do not change the spherical gradients. □

The next corollary converts the moment derivatives into ambient and spherical particle-wise gradients.

Corollary B.1 (Ambient and spherical gradients)**.** ***Assumptions used.*** *assumptions B.1 and B.2.*

Under the notation of lemma B.4, the ambient gradients on the sphere are

$$\nabla_w \frac{\delta F_{\mathrm{fit}}}{\delta\mu_1}(w) = \sum_{n\geq 1} n\, G^{\mathrm{fit}}_{1,n} \otimes_{n-1} w^{\otimes(n-1)}, \tag{289}$$

$$\nabla_w \frac{\delta F_{\mathrm{fit}}}{\delta\mu_2}(w) = \sum_{m\geq 1} m\, G^{\mathrm{fit}}_{2,m} \otimes_{m-1} w^{\otimes(m-1)}. \tag{290}$$

Their spherical projections are

$$\nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{fit}}}{\delta\mu_1}(w) = \sum_{n\geq 1} n\Big(G^{\mathrm{fit}}_{1,n} \otimes_{n-1} w^{\otimes(n-1)} - \big(G^{\mathrm{fit}}_{1,n} \otimes_n w^{\otimes n}\big) w\Big) \tag{291}$$

$$\nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{fit}}}{\delta\mu_2}(w) = \sum_{m\geq 1} m\Big(G^{\mathrm{fit}}_{2,m} \otimes_{m-1} w^{\otimes(m-1)} - \big(G^{\mathrm{fit}}_{2,m} \otimes_m w^{\otimes m}\big) w\Big). \tag{292}$$

Proof. For a symmetric tensor $G \in \mathrm{Sym}^n(\mathbb{R}^{d_w})$ and direction u, the product rule and symmetry give

$$D_w\langle G, w^{\otimes n}\rangle[u] = \sum_{j=1}^{n}\langle G, w^{\otimes(j-1)} \otimes u \otimes w^{\otimes(n-j)}\rangle$$
$$= n\langle G \otimes_{n-1} w^{\otimes(n-1)}, u\rangle.$$

Moreover, $w^\top(G \otimes_{n-1} w^{\otimes(n-1)}) = G \otimes_n w^{\otimes n}$. Apply these identities to each $G_{i,n}^{\mathrm{fit}}$ in lemma B.4, and project the ambient gradient with $I - ww^\top$. This gives both displayed spherical formulas. The active sums are finite by assumption B.2. □

We then compute the first and second variations of the metric regularizer on the fixed-assignment cap class.

Lemma B.5 (First and second variations of the fixed-assignment metric regularizer)**.** ***Assumptions used.*** *assumption B.1.*

Let

$$F_{\mathrm{reg}}(\mu_1, \mu_2) = \frac{\lambda_{\mathrm{reg}}}{p_W} Z_\pm(\mu_1, \mu_2)^{p_W/2}, \qquad p_W = 2 + \delta_W,$$

where

$$Z_\pm(\mu_1, \mu_2) = W_2^2(\mu_1, \nu_\star^\pm) + W_2^2(\mu_2, \nu_\star^\pm).$$

On the fixed-assignment class of lemma B.1, define

$$c_\star(w) = \frac{1}{2} d_{\mathbb{S}}(w, \pi_\star(w))^2, \qquad D_\pm(\mu_1, \mu_2) \coloneqq \sum_{i=1}^{2} \int c_\star(w)\, \mathrm{d}\mu_i(w).$$

Then

$$Z_\pm = 2D_\pm.$$

The first variation of F_{reg}, up to an irrelevant additive constant, is

$$\frac{\delta F_{\mathrm{reg}}}{\delta \mu_i}(w) = \lambda_{\mathrm{reg}} Z_\pm^{p_W/2-1} c_\star(w), \qquad i = 1, 2. \tag{293}$$

Consequently, the Wasserstein gradient on the sphere is

$$\nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{reg}}}{\delta \mu_i}(w) = \lambda_{\mathrm{reg}} Z_\pm^{p_W/2-1} \nabla_{\mathbb{S}} c_\star(w). \tag{294}$$

The second variation formula below is the curvature contribution used in the pre-saddle and post-escape convexity arguments. For $Z_\pm > 0$ and a smooth spherical pushforward path generated by tangent fields $v = (v_1, v_2)$, as long as the fixed assignment remains valid for small $|h|$,

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{reg}}(\mu_{1,h}, \mu_{2,h}) = \lambda_{\mathrm{reg}} Z_\pm^{p_W/2-1} \sum_{i=1}^{2} \int \nabla_{\mathbb{S}}^2 c_\star(w)[v_i(w), v_i(w)]\, \mathrm{d}\mu_i(w) \tag{295}$$

$$+ \lambda_{\mathrm{reg}}(p_W - 2) Z_\pm^{p_W/2-2} \left(\sum_{i=1}^{2} \int \langle \nabla_{\mathbb{S}} c_\star(w), v_i(w)\rangle\, \mathrm{d}\mu_i(w) \right)^2, \tag{296}$$

where $\mu_{i,h} = (\Phi_{i,h})_\# \mu_i$, $\Phi_{i,h}(w) = \mathrm{EXP}_w(h v_i(w))$. The second term is nonnegative. If $Z_\pm = 0$, i.e. at $\mu_1 = \mu_2 = \nu_\star^\pm$, then its EXP-pushforward second derivative exists and equals zero; the exact central calculation is (298)*.*

Proof. **Step 1: differentiate the fixed-assignment cost.** For a signed perturbation of μ_i, the variation of $Z_\pm$ is

$$\delta Z_\pm = 2 \int c_\star(w)\, \delta\mu_i(\mathrm{d}w).$$

Therefore

$$\frac{\delta F_{\mathrm{reg}}}{\delta \mu_i}(w) = \lambda_{\mathrm{reg}} Z_{\pm}^{p_W/2-1} c_\star(w).$$

Taking the spherical gradient in w gives

$$\nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{reg}}}{\delta \mu_i}(w) = \lambda_{\mathrm{reg}} Z_{\pm}^{p_W/2-1} \nabla_{\mathbb{S}} c_\star(w).$$

Step 2: compute the second derivative at positive distance. Let

$$\mu_{i,h} = (\Phi_{i,h})_{\#}\mu_i, \qquad \Phi_{i,h}(w) = \mathrm{EXP}_w(h v_i(w)),$$

where $v_i(w) \in T_w \mathbb{S}^{d_w - 1}$. By (254),

$$\Phi_{i,h}(w) = \cos(h\|v_i(w)\|)w + \frac{\sin(h\|v_i(w)\|)}{\|v_i(w)\|} v_i(w).$$

The quotient is interpreted by continuity at $v_i(w) = 0$. Differentiating the sine and cosine gives

$$\dot{\Phi}_{i,0}(w) = v_i(w), \qquad \ddot{\Phi}_{i,0}(w) = -\|v_i(w)\|^2 w, \qquad \nabla_h \dot{\Phi}_{i,h}(w)|_{h=0} = (I - ww^\top)\ddot{\Phi}_{i,0}(w) = 0. \tag{297}$$

Thus the covariant acceleration, namely the tangential part of the ambient acceleration, vanishes.

Define

$$D_{\pm}(h) \coloneqq \sum_{i=1}^{2} \int c_\star(\Phi_{i,h}(w)) \, \mathrm{d}\mu_i(w), \qquad Z_{\pm}(h) = 2D_{\pm}(h).$$

By differentiating under the integral,

$$D'_{\pm}(0) = \sum_{i=1}^{2} \int \langle \nabla_{\mathbb{S}} c_\star(w), v_i(w) \rangle \, \mathrm{d}\mu_i(w).$$

Moreover, since the paths $h \mapsto \Phi_{i,h}(w)$ are geodesics,

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} c_\star(\Phi_{i,h}(w)) = \nabla^2_{\mathbb{S}} c_\star(w)[v_i(w), v_i(w)].$$

Thus

$$D''_{\pm}(0) = \sum_{i=1}^{2} \int \nabla^2_{\mathbb{S}} c_\star(w)[v_i(w), v_i(w)] \, \mathrm{d}\mu_i(w).$$

Consequently, $Z'_{\pm}(0) = 2D'_{\pm}(0),\ Z''_{\pm}(0) = 2D''_{\pm}(0)$. For $Z_{\pm} > 0$, differentiate

$$h \longmapsto F_{\mathrm{reg}}(\mu_{1,h}, \mu_{2,h}) = \frac{\lambda_{\mathrm{reg}}}{p_W} Z_{\pm}(h)^{p_W/2}.$$

The second derivative is

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{reg}}(\mu_{1,h}, \mu_{2,h}) = \frac{\lambda_{\mathrm{reg}}}{2} Z_{\pm}^{p_W/2-1} Z''_{\pm}(0) \\ + \frac{\lambda_{\mathrm{reg}}}{2} \left(\frac{p_W}{2} - 1\right) Z_{\pm}^{p_W/2-2} \big(Z'_{\pm}(0)\big)^2.\end{aligned}$$

Substituting $Z'_{\pm}(0) = 2D'_{\pm}(0)$ and $Z''_{\pm}(0) = 2D''_{\pm}(0)$, we obtain

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{reg}}(\mu_{1,h}, \mu_{2,h}) = \lambda_{\mathrm{reg}} Z_{\pm}^{p_W/2-1} D''_{\pm}(0) \\ + \lambda_{\mathrm{reg}}(p_W - 2) Z_{\pm}^{p_W/2-2} \big(D'_{\pm}(0)\big)^2.\end{aligned}$$

Using the formulas for $D'_{\pm}(0)$ and $D''_{\pm}(0)$, this is exactly

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{reg}}(\mu_{1,h}, \mu_{2,h}) =& \lambda_{\mathrm{reg}} Z_{\pm}^{p_W/2-1} \sum_{i=1}^{2} \int \nabla^2_{\mathbb{S}} c_\star(w)[v_i(w), v_i(w)] \, \mathrm{d}\mu_i(w) \\ &+ \lambda_{\mathrm{reg}}(p_W - 2) Z_{\pm}^{p_W/2-2} \left(\sum_{i=1}^{2} \int \langle \nabla_{\mathbb{S}} c_\star(w), v_i(w) \rangle \, \mathrm{d}\mu_i(w) \right)^2.\end{aligned}$$

The second term is nonnegative because $p_W = 2 + \delta_W > 2$.

Step 3: differentiate directly at the antipodal target. If $Z_\pm = 0$, then $\mu_0 = \nu_\star^\pm \otimes \nu_\star^\pm$. Each starting atom y is its own assigned pole. For sufficiently small $|h|$, eqs. (251) and (254) give, including zero velocities,

$$d_{\mathbb{S}}(\mathrm{EXP}_y(hv_i(y)), y)^2 = h^2\|v_i(y)\|^2, \qquad Z_\pm(\mu_h) = h^2 \sum_i \int \|v_i(y)\|^2 \, d\mu_{0,i}(y).$$

Thus, with the product norm $\|v\|^2_{L^2(\mu_0)} = \sum_i \int \|v_i\|^2 \, d\mu_{0,i}$,

$$F_{\mathrm{reg}}(\mu_h) = \frac{\lambda_{\mathrm{reg}}}{p_W}|h|^{p_W}\|v\|^{p_W}_{L^2(\mu_0)}. \tag{298}$$

For $h \neq 0$, direct differentiation yields

$$\frac{d}{dh}F_{\mathrm{reg}}(\mu_h) = \lambda_{\mathrm{reg}}\,\mathrm{sgn}(h)|h|^{p_W-1}\|v\|^{p_W}_{L^2(\mu_0)},$$
$$\frac{d^2}{dh^2}F_{\mathrm{reg}}(\mu_h) = \lambda_{\mathrm{reg}}(p_W-1)|h|^{p_W-2}\|v\|^{p_W}_{L^2(\mu_0)}.$$

At $h = 0$, the value difference quotient tends to zero. Dividing the first derivative at $h \neq 0$ by h gives $\lambda_{\mathrm{reg}}|h|^{p_W-2}\|v\|^{p_W}_{L^2(\mu_0)} \to 0$. Therefore the second derivative exists at zero and equals zero, since $p_W > 2$; the second derivative displayed above is also continuous there. □

Lemma B.6 (Full spherical product Hessian and spherical product pushforward second variation)**.** *Assume assumptions B.1 and B.2. Let $\mu = \mu_1 \otimes \mu_2 \in \mathcal{X}_\pm(r_{\mathrm{cap}})$, and let $v = (v_1, v_2)$ be a smooth spherical product field. Use the product norm $\|v\|^2_{L^2(\mu)} = \sum_i \int \|v_i(w)\|^2 \, d\mu_i(w)$. For sufficiently small $|h|$, require the paths*

$$\Phi_{i,h}(w) = \mathrm{EXP}_w(hv_i(w)), \qquad \mu_h = \bigotimes_{i=1}^{2}(\Phi_{i,h})_\# \mu_i$$

to remain in the fixed-assignment caps of (242)*. The partial measure derivatives of F_{fit} are those of its finite polynomial in the moments. In terms of $G^{\mathrm{fit}}_{i,n} = \partial F_{\mathrm{fit}}/\partial T_{i,n}$ from lemma B.4, this means*

$$\frac{\delta^2 F_{\mathrm{fit}}}{\delta\mu_i\delta\mu_j}(\mu; w, u) := \sum_{n,m\in\mathcal{O}} \left\langle D_{T_{j,m}} G^{\mathrm{fit}}_{i,n}[u^{\otimes m}], w^{\otimes n}\right\rangle_{\mathrm{HS}}. \tag{299}$$

For $Z_\pm(\mu) > 0$, define the full spherical product Hessian by

$$\begin{aligned}(\mathcal{H}_{\mathbb{S},\mu}v)_i(w) &= \nabla^2_{\mathbb{S}}\frac{\delta F}{\delta\mu_i}(\mu; w)v_i(w) \\ &\quad + \sum_{j=1}^{2}\int \nabla_{\mathbb{S},w}\nabla_{\mathbb{S},u}\frac{\delta^2 F}{\delta\mu_i\delta\mu_j}(\mu; w, u)v_j(u)\, d\mu_j(u).\end{aligned} \tag{300}$$

The first term differentiates the location with the measures fixed. The integral term differentiates the measures; its mixed derivative maps $T_u\mathbb{S}$ to $T_w\mathbb{S}$. At $Z_\pm(\mu) = 0$, the direct calculation eq. (298) *shows that the regularizer has zero pushforward second variation. Hence the full operator there is $\mathcal{H}_{\mathbb{S},\mu} = \mathcal{H}^{\mathrm{fit}}_{\mathbb{S},\mu}$, obtained from eq.* (300) *by using F_{fit} in both terms. Then the spherical product pushforward second variation exists and satisfies*

$$\begin{aligned}D^2_{W_2}F(\mu)[v, v] &:= \left.\frac{d^2}{dh^2}\right|_0 F(\mu_h) \\ &= \sum_{i=1}^{2}\int \langle v_i(w), (\mathcal{H}_{\mathbb{S},\mu}v)_i(w)\rangle \, d\mu_i(w) = \langle v, \mathcal{H}_{\mathbb{S},\mu}v\rangle_{L^2(\mu)}.\end{aligned} \tag{301}$$

This notation refers to the stated product pushforward; it does not require every such path to be an optimal Wasserstein geodesic. When the displayed pushforward path is a product Wasserstein geodesic, the same quantity is its usual geodesic second derivative. Thus the estimates below apply, in particular, along every admissible path of that form.

Proof. **Step 1: differentiate the fitting term in both its arguments.** The explicit EXP calculation (297) gives $\dot{\Phi}_{i,0} = v_i$ and zero covariant acceleration. The product first-variation formula in section B.2 gives

$$\frac{d}{dh}F_{\mathrm{fit}}(\mu_h) = \sum_i \int \left\langle \nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{fit}}}{\delta \mu_i}(\mu_h; \Phi_{i,h}(w)), \dot{\Phi}_{i,h}(w) \right\rangle d\mu_i(w).$$

The moment derivative along the path is the finite product rule

$$\left.\frac{d}{dh}\right|_0 T_{j,m}(\mu_h) = \int \sum_{\ell=1}^{m} u^{\otimes(\ell-1)} \otimes v_j(u) \otimes u^{\otimes(m-\ell)}\, d\mu_j(u).$$

Equivalently, the integrand is $D_u(u^{\otimes m})[v_j(u)]$. Hence the coefficient derivative is the finite-dimensional chain rule

$$\begin{aligned}\left.\frac{d}{dh}\right|_0 G^{\mathrm{fit}}_{i,n}(\mu_h) &= \sum_{j=1}^{2} \sum_{m\in\mathcal{O}} D_{T_{j,m}} G^{\mathrm{fit}}_{i,n} \left[\left.\frac{d}{dh}\right|_0 T_{j,m}(\mu_h) \right] \\ &= \sum_{j=1}^{2} \sum_{m\in\mathcal{O}} \int D_{T_{j,m}} G^{\mathrm{fit}}_{i,n} \left[D_u(u^{\otimes m})[v_j(u)] \right], d\mu_j(u).\end{aligned} \tag{302}$$

Insert this into the derivative of $\sum_n \langle G^{\mathrm{fit}}_{i,n}, w^{\otimes n}\rangle$ from eq. (277). Pairing eq. (302) with $w^{\otimes n}$, summing over n, and differentiating $u^{\otimes m}$ in the direction $v_j(u)$ is exactly the u-gradient of the kernel in eq. (299). Thus, for fixed w,

$$\left.\frac{d}{dh}\right|_0 \frac{\delta F_{\mathrm{fit}}}{\delta \mu_i}(\mu_h; w) = \sum_j \int \left\langle \nabla_{\mathbb{S},u} \frac{\delta^2 F_{\mathrm{fit}}}{\delta \mu_i \delta \mu_j}(\mu; w, u), v_j(u) \right\rangle d\mu_j(u).$$

All sums are finite by assumption B.2, so the sum–integral interchange in eq. (302) requires no limiting argument. This is the finite polynomial chain rule for the moments in lemmas B.3 and B.4. Taking its spherical gradient in w gives the integral term of (300) for F_{fit}. Differentiating the evaluation point gives its spatial-Hessian term. The remaining derivative of $\dot{\Phi}_{i,h}$ contributes $\langle \nabla_{\mathbb{S}}(\delta F_{\mathrm{fit}}/\delta\mu_i), \nabla_h \dot{\Phi}_{i,h}\rangle = 0$ at $h = 0$. Thus differentiating the first display proves (301) for F_{fit}. All differentiations under the integrals are valid for atomic as well as diffuse measures: the Hermite sums are finite and the supports are compact subsets of the smooth cap region.

Step 2: identify the regularizer contribution. Recall $c_\star(w) = d_{\mathbb{S}}(w, \pi_\star(w))^2/2$ from (252), and $Z_\pm = 2\sum_i \int c_\star\, d\mu_i$ from (251). For $Z_\pm > 0$, differentiating (293) gives

$$\begin{aligned}\frac{\delta F_{\mathrm{reg}}}{\delta \mu_i}(\mu; w) &= \lambda_{\mathrm{reg}} Z_\pm^{(p_W-2)/2} c_\star(w), \\ \frac{\delta^2 F_{\mathrm{reg}}}{\delta \mu_i \delta \mu_j}(\mu; w, u) &= \lambda_{\mathrm{reg}}(p_W - 2) Z_\pm^{(p_W-4)/2} c_\star(w) c_\star(u).\end{aligned}$$

The factor $p_W - 2$ is $2 \cdot (p_W - 2)/2$, since a variation of μ_j changes $Z_\pm$ by $2\int c_\star\, d(\delta\mu_j)$. Their spatial derivatives in (300), paired with v, give

$$\begin{aligned}&\lambda_{\mathrm{reg}} Z_\pm^{(p_W-2)/2} \sum_i \int \nabla^2_{\mathbb{S}} c_\star(w)[v_i(w), v_i(w)]\, d\mu_i(w) \\ &\quad + \lambda_{\mathrm{reg}}(p_W - 2) Z_\pm^{(p_W-4)/2} \left(\sum_i \int \langle \nabla_{\mathbb{S}} c_\star(w), v_i(w)\rangle\, d\mu_i(w) \right)^2.\end{aligned}$$

This equals the directly computed second derivative (295). For $Z_\pm = 0$, use (298) and its difference-quotient calculation instead: the second derivative is zero. No second measure variation of $Z_\pm^{p_W/2}$ at zero is used. Adding this contribution to Step 1 proves the identity. □

Remark B.4 (Relation to the full Hessian and the auxiliary comparison)**.** *The identity* (301) *is the spherical product counterpart of the full Hessian in section 2.1. Stationarity alone does not remove*

its spatial term. The normalized retractions used below have the same first and second derivatives as EXP at zero by (355)*; the finite-moment fitting chain rule therefore gives the same second variation. For the regularizer at the center, the corresponding direct calculation is* (369)*. This suggests a possible finite-sample curvature comparison of the kind studied in section A.10. Such a transfer theorem for this spherical product model remains open in the present work: the hypotheses of that auxiliary result are not verified here. The central calculation is a second-variation identity, not an application of the general C^3 theorem at the MF target.*

B.6 Symmetric and Antisymmetric Sub-Manifolds

This subsection introduces the symmetric and antisymmetric coordinates that separate the teacher branch from the antipodal saddle branch.

We first collect the active odd moments into a single finite-dimensional tensor space.

Define the active odd raw-moment space

$$\mathcal{H}_{\mathcal{O}} \coloneqq \mathcal{H}_{\mathcal{O}}^{\mathrm{raw}} = \bigoplus_{n\in\mathcal{O}} \mathrm{Sym}^n(\mathbb{R}^{d_w}), \qquad \|\mathbf{X}\|_{\mathcal{H}_{\mathcal{O}}}^2 \coloneqq \|\mathbf{X}\|_{\mathrm{raw}}^2 = \sum_{n\in\mathcal{O}} \|X_n\|_{\mathrm{HS}}^2. \tag{303}$$

The weighted norm records the size of active moments as seen through the activation coefficients:

$$\|\mathbf{X}\|_a^2 \coloneqq \sum_{n\in\mathcal{O}} \|a_n\|_2^2 \|X_n\|_{\mathrm{HS}}^2. \tag{304}$$

We now pass from the two raw moment sequences to symmetric and antisymmetric coordinates.

Definition B.3. *Introduce the two active student blocks and the teacher block*

$$\mathbf{T}_i \coloneqq (T_{i,n})_{n\in\mathcal{O}} \in \mathcal{H}_{\mathcal{O}}, \quad i=1,2, \quad \mathbf{T}^o \coloneqq (T_n^o)_{n\in\mathcal{O}} \in \mathcal{H}_{\mathcal{O}}, \tag{305}$$

where $T_{i,n} = \int w^{\otimes n}\, d\mu_i(w)$ and $T_n^o = \int w^{\otimes n}\, d\nu^o(w)$ as in assumption B.2. Recall the symmetric and antisymmetric blocks

$$\mathbf{S} \coloneqq (S_n)_{n\in\mathcal{O}}, \qquad \mathbf{A} \coloneqq (A_n)_{n\in\mathcal{O}}. \tag{306}$$

with $S_n = (T_{1,n} + T_{2,n})/2$ and $A_n = (T_{1,n} - T_{2,n})/2$. Thus

$$T_{1,n} = S_n + A_n, \qquad T_{2,n} = S_n - A_n, \qquad n \in \mathcal{O}. \tag{307}$$

The finite active moment map used in the branch geometry is

$$\Theta(\mu_1,\mu_2) \coloneqq \big(\mathbf{S}(\mu_1,\mu_2), \mathbf{A}(\mu_1,\mu_2)\big). \tag{308}$$

Only the moments with $n \in \mathcal{O}$ enter Θ; it is not asserted to be an injective coordinate chart of probability measures.

For each $p \geq 0$, we define the active ambient moment polynomial. In these coordinates, the active fitting residual becomes an ambient polynomial in $(\mathbf{S}, \mathbf{A})$.

$$\widetilde{\mathcal{M}}_p(\mathbf{S},\mathbf{A}) \coloneqq \sum_{\substack{n,m\in\mathcal{O} \\ 0\leq r\leq \min(n,m) \\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left((S_n + A_n) \otimes_r (S_m - A_m)\right). \tag{309}$$

The corresponding ambient active fitting functional is

$$\widetilde{\mathcal{J}}_{\mathrm{fit}}(\mathbf{S},\mathbf{A}) \coloneqq \frac{1}{2} \sum_{p\geq 0} \big\|\widetilde{\mathcal{M}}_p(\mathbf{S},\mathbf{A}) - \mathcal{M}_p^o\big\|_{\mathrm{HS}}^2. \tag{310}$$

Membership in either branch implies that its finite active moments lie on the corresponding coordinate axis.

Definition B.4. *For any measurable set A, let $-A \coloneqq \{-w \mid w \in A\}$. Consider the following two branches:*

1. ***Antisymmetric branch**: The antisymmetric branch is the candidate path toward the antipodal saddle.*

$$\mathcal{M}_- \coloneqq \{\mu_1 \otimes \mu_2 \mid \mu_1(-A) = \mu_2(A) \text{ for every measurable } A\}, \tag{311}$$

which is equivalent to

$$T_{1,n} = (-1)^n T_{2,n} \qquad (n \geq 0) \tag{312}$$

and this implies $\mathbf{S} = 0$.

2. ***Symmetric branch**: The symmetric branch contains the same-density teacher configuration.*

$$\mathcal{M}_+ \coloneqq \{\mu_1 \otimes \mu_2 \mid \mu_1 = \mu_2\}, \tag{313}$$

which is equivalent to

$$T_{1,n} = T_{2,n} \qquad (n \geq 0) \tag{314}$$

and this implies $\mathbf{A} = 0$.

We write the fitting loss restricted to each branch as a one-coordinate functional.

$$\mathcal{J}_+(\mathbf{S}) \coloneqq \widetilde{\mathcal{J}}_{\text{fit}}(\mathbf{S}, 0), \quad \mathcal{J}_-(\mathbf{A}) \coloneqq \widetilde{\mathcal{J}}_{\text{fit}}(0, \mathbf{A}). \tag{315}$$

The equivalences above use all raw orders $n \geq 0$, not only $n \in \mathcal{O}$. Here is the justification on the compact sphere. Equality of all tensor moments gives equality of every monomial integral and hence every polynomial integral. Restricted polynomials form an algebra containing constants and separating points of the sphere. By Stone–Weierstrass, for each continuous f and $\epsilon > 0$ there is a polynomial P with $\sup_{\mathbb{S}} |f - P| < \epsilon$. Thus, for two measures with the same raw moments,

$$\left| \int f \, d\mu_1 - \int f \, d\mu_2 \right| \leq \int |f - P| \, d\mu_1 + \int |f - P| \, d\mu_2 < 2\epsilon.$$

Letting $\epsilon \downarrow 0$ gives equality of the measures. For reflection $\mathsf{R}(w) = -w$, the identity

$$T_n(\mathsf{R}_{\#}\mu_1) = \int (-w)^{\otimes n} \, d\mu_1(w) = (-1)^n T_n(\mu_1)$$

gives the antisymmetric equivalence by the same argument. Consequently $\mu \in \mathcal{M}_-$ implies $\mathbf{S} = 0$ and $\mu \in \mathcal{M}_+$ implies $\mathbf{A} = 0$, but neither converse is inferred from the finite active blocks.

The first proposition identifies the distinguished active-origin point where the two branches meet.

Proposition B.1 (Active origin fiber and the distinguished Dirac saddle)**. *Assumptions used.** assumptions B.1 and B.2.*

Define the active-coordinate zero fiber

$$\mathcal{C}_0 \coloneqq \{\mu_1 \otimes \mu_2 \mid \Theta(\mu_1, \mu_2) = (0, 0)\}. \tag{316}$$

Then

$$\mu_\dagger \coloneqq \nu_\star^\pm \otimes \nu_\star^\pm \in \mathcal{M}_+ \cap \mathcal{M}_- \cap \mathcal{C}_0. \tag{317}$$

In addition, if

$$\mu_1 = \mu_2 = \nu_\star^\pm, \tag{318}$$

then the Wasserstein gradients vanish. In other words, $\mu_\dagger = \nu_\star^\pm \otimes \nu_\star^\pm$ *is a first-order stationary point.*

Proof. The proof only uses antipodal symmetry, oddness of the active moments, and the flatness of the metric regularizer at the antipodal target.

$\mu_\dagger$ is antipodally symmetric, equal in both factors, and has zero odd moments; hence it belongs to the displayed intersection.

If $\mu \in \mathcal{C}_0$, then $T_{1,n} = T_{2,n} = 0$ for all $n \in \mathcal{O}$. Therefore, on $\mathcal{C}_0$, the fitting first variation vanishes because of the definition of $G_{i,n}^{\text{fit}}$.

The full spherical gradient is only the regularization contribution

$$\nabla_{\mathbb{S}} \frac{\delta F}{\delta \mu_i} = \lambda_{\text{reg}} Z_\pm^{p_W/2-1} \nabla_{\mathbb{S}} c_\star(w), \qquad i = 1, 2. \tag{319}$$

If μ_1 and μ_2 are $\nu_\star^\pm$, then $Z_\pm = 0$ and the spherical gradient vanishes. □

The next proposition fixes the opposite endpoint of the cartoon: the exact-fit teacher lies on the same-density branch.

Proposition B.2 (The reference exact-fit point lies on the same-density branch)**.** ***Assumptions used.*** *assumptions B.1 and B.2.*

The fitting part satisfies

$$\widetilde{\mathcal{J}}_{\mathrm{fit}}(\mathbf{T}^o, 0) = 0 = \inf_{(\mathbf{S},\mathbf{A})\in\mathcal{H}_{\mathcal{O}}\times\mathcal{H}_{\mathcal{O}}} \widetilde{\mathcal{J}}_{\mathrm{fit}}(\mathbf{S}, \mathbf{A}). \tag{320}$$

Moreover, by the symmetric-teacher assumption, the reference teacher product measure belongs to the same-density branch:

$$\mu^o = \mu_1^o \otimes \mu_2^o \in \mathcal{M}_+, \qquad \mathbf{S}(\mu^o) = \mathbf{T}^o, \qquad \mathbf{A}(\mu^o) = 0. \tag{321}$$

In particular, the distinguished zero-fit global minimum of the data term lies on $\mathcal{M}_+$.

Proof. By assumption B.2, the teacher is symmetric across the two factors, so $\mu_1^o = \mu_2^o$ and therefore $\mu^o \in \mathcal{M}_+$. In the active coordinates this is exactly the identity $\mathbf{S}(\mu^o) = \mathbf{T}^o$ and $\mathbf{A}(\mu^o) = 0$.

Next, by the definition of $\mathcal{M}_p^o$ and of the ambient moment map,

$$\widetilde{\mathcal{M}}_p(\mathbf{T}^o, 0) = \mathcal{M}_p^o \qquad \forall p \geq 0. \tag{322}$$

Hence $\widetilde{\mathcal{J}}_{\mathrm{fit}}(\mathbf{T}^o, 0) = 0$. Since $\widetilde{\mathcal{J}}_{\mathrm{fit}}$ is a sum of squared Hilbert norms, it is nonnegative everywhere. □

B.7 The anti-symmetric branch and descent towards the saddle

This subsection proves invariance of the anti-symmetric branch and, under a local initial condition, cap confinement and convergence to the antipodal Dirac target.

B.7.1 Invariance of the anti-symmetric branch

We first show that the first variation respects the antipodal reflection symmetry.

Lemma B.7 (Reflection–exchange covariance of the full first variation)**.** ***Assumptions used.*** *assumptions B.1 and B.2.*

Let $\mathsf{R}(w) = -w$ and define the reflection–exchange involution

$$\mathfrak{R}(\mu_1, \mu_2) := (\mathsf{R}_{\#}\mu_2, \mathsf{R}_{\#}\mu_1).$$

For every admissible product state $\mu = (\mu_1, \mu_2)$,

$$F(\mathfrak{R}\mu) = F(\mu), \tag{323}$$

and its spherical gradient fields satisfy

$$\begin{aligned} g_1(\mathfrak{R}\mu; -w) &= -g_2(\mu; w), \\ g_2(\mathfrak{R}\mu; -w) &= -g_1(\mu; w), \end{aligned} \tag{324}$$

where $g_i = \nabla_{\mathbb{S}}(\delta F/\delta\mu_i)$. If $\mu \in \mathcal{M}_-$, then $\mathfrak{R}\mu = \mu$, so

$$\frac{\delta F}{\delta\mu_2}(-w) = \frac{\delta F}{\delta\mu_1}(w) + \mathrm{const}, \qquad g_2(\mu; -w) = -g_1(\mu; w). \tag{325}$$

Proof. **Step 1: transform the objective at an arbitrary state.** Oddness of the activation gives, for every marginal ν,

$$h_{\mathsf{R}_{\#}\nu}(z) = \int \sigma((-w)^\top z)\, d\nu(w) = -h_\nu(z).$$

Consequently,

$$h_{\mathsf{R}_{\#}\mu_2} h_{\mathsf{R}_{\#}\mu_1}^\top = h_{\mu_2} h_{\mu_1}^\top = (h_{\mu_1} h_{\mu_2}^\top)^\top.$$

The teacher matrix $h_{\nu^o} h_{\nu^o}^\top$ is symmetric by assumption B.2, and the Frobenius norm is invariant under transposition. Thus the fitting loss is unchanged. Reflection is a spherical isometry and $\mathsf{R}_\# \nu_\star^\pm = \nu_\star^\pm$. Pushing any transport plan forward by (R, R), and then applying this involution once more, gives

$$W_2(\mathsf{R}_\# \nu, \nu_\star^\pm) = W_2(\nu, \nu_\star^\pm).$$

The reflection exchanges the two summands in $Z_\pm$, which proves eq. (323) for every admissible state.

Step 2: differentiate the symmetry and specialize to the branch. Vary the second marginal of μ in eq. (323). The corresponding variation of the first marginal of $\mathfrak{R}\mu$ is reflected, and the first-variation identity is

$$\frac{\delta F}{\delta \mu_1}(\mathfrak{R}\mu; -w) = \frac{\delta F}{\delta \mu_2}(\mu; w) + \text{const.}$$

Here the additive constant is immaterial because first-variation potentials are defined only up to constants. Since $D\mathsf{R} = -I$, differentiating this identity gives, for every $v \in T_w \mathbb{S}^{d_w - 1}$,

$$\langle g_1(\mathfrak{R}\mu; -w), -v \rangle_{\mathbb{R}^{d_w}} = \langle g_2(\mu; w), v \rangle_{\mathbb{R}^{d_w}}.$$

Since this holds for every v, it proves the first identity in eq. (324); exchanging the factors proves the second. A state lies in $\mathcal{M}_-$ precisely when $\mathfrak{R}\mu = \mu$, so the gradient and potential identities in eq. (325) follow. □

We prove branch invariance by applying reflection and factor exchange to an arbitrary admissible product state before using uniqueness.

Proposition B.3 (Invariance of $\mathcal{M}_-$)**.** *Assume assumptions B.1 and B.2. Let $\mu_0 \in \mathcal{M}_-$. On every interval on which the characteristic product WGF is uniquely defined and remains in the fixed-assignment cap class $\mu_{i,t} \in \mathcal{P}_\pm(r_{\text{cap}})$, $i = 1, 2$,*

$$\mu_{2,t} = (-\mathrm{Id})_\# \mu_{1,t}, \qquad \mu_t \in \mathcal{M}_-.$$

Proof. Let $\mathfrak{R}$ be the reflection–exchange involution from lemma B.7. If $\Phi_{i,t}$ are the original characteristics, define

$$\widetilde{\Phi}_{1,t}(w) := -\Phi_{2,t}(-w), \qquad \widetilde{\Phi}_{2,t}(w) := -\Phi_{1,t}(-w).$$

Using the characteristic equation $\dot{\Phi}_{i,t} = -g_i(\mu_t; \Phi_{i,t})$ and the first identity in eq. (324),

$$\begin{aligned} \frac{d}{dt} \widetilde{\Phi}_{1,t}(w) &= g_2(\mu_t; \Phi_{2,t}(-w)) \\ &= -g_1(\mathfrak{R}\mu_t; \widetilde{\Phi}_{1,t}(w)). \end{aligned}$$

The second factor satisfies the analogous equation. Their marginal laws are $\mathfrak{R}\mu_t = (\mathsf{R}_\# \mu_{2,t}, \mathsf{R}_\# \mu_{1,t})$, and reflection exchanges the two equal-mass caps, so the transformed pair remains in the same admissible class.

Since $\mu_0 \in \mathcal{M}_-$, one has $\mathfrak{R}\mu_0 = \mu_0$; hence the transformed and original characteristic systems have the same initial state. Uniqueness on the stated interval gives $\mathfrak{R}\mu_t = \mu_t$, or equivalently $\mu_{2,t} = \mathsf{R}_\# \mu_{1,t}$, throughout that interval. □

B.7.2 Dissipation towards the antipodal saddle point on the anti-symmetric branch

We establish quantitative transport estimates and use them to prove local attraction to the antipodal target.

Fixed-assignment geometry and first variations. Recall the cap class $\mathcal{P}_\pm(r_{\text{cap}})$ and product class $\mathcal{X}_\pm(r_{\text{cap}}) = \mathcal{P}_\pm(r_{\text{cap}})^2$ from eqs. (241) and (242): each marginal has mass $1/2$ in each of the two open caps. The assigned pole and its cost from eqs. (250) and (252) are

$$\pi_\star(w) = \begin{cases} a_\star, & w \in K_+(r_{\text{cap}}), \\ -a_\star, & w \in K_-(r_{\text{cap}}), \end{cases} \qquad c_\star(w) = \frac{1}{2} r(w)^2, \qquad r(w) := d_{\mathbb{S}}(w, \pi_\star(w)).$$

The endpoint vector $q_\star$ of definition B.2 is illustrated in fig. 3. The fixed-assignment calculation in lemma B.2 and eqs. (255) and (256) shows that C_s is a constant-speed spherical geodesic with zero covariant acceleration and that

$$q_\star(w) = \nabla_{\mathbb{S}} c_\star(w) = -\mathrm{LOG}_w \pi_\star(w).$$

For subsequent derivative formulas define the gradient fields

$$g_i(\mu, w) := \nabla_{\mathbb{S}} \frac{\delta F}{\delta \mu_i}(\mu, w), \qquad g_{i,\mathrm{fit}}(\mu, w) := \nabla_{\mathbb{S}} \frac{\delta F_{\mathrm{fit}}}{\delta \mu_i}(\mu, w). \tag{326}$$

By the fixed-assignment formula eq. (251),

$$\begin{aligned} Z_\pm(\mu) &= \sum_{i=1}^{2} \int r(w)^2 \, \mathrm{d}\mu_i(w), \\ \|(q_\star, q_\star)\|^2_{L^2(\mu)} &= Z_\pm(\mu), \qquad \nabla_{W_2} Z_\pm(\mu) = 2(q_\star, q_\star). \end{aligned} \tag{327}$$

Here the product norm is $\|V\|^2_{L^2(\mu)} = \sum_{i=1}^2 \int \|V_i\|^2 \, \mathrm{d}\mu_i$. For a smooth product field $V = (V_1, V_2)$, write $DF(\mu)[V] := \sum_{i=1}^2 \int \langle g_i, V_i \rangle \, \mathrm{d}\mu_i$; the same convention applies to F_{fit}.

Constants for the local estimates. Under assumption B.2, write $b(z) := h_{\nu^o}(z)$. Let G_1, G_2 be independent standard real Gaussians. For a vector-valued random variable X, use $\|X\|_{L^4} := (\mathbb{E}\|X\|_2^4)^{1/4}$, and define

$$\begin{aligned} M_4 &:= \|\sigma(G_1)\|_{L^4}, \qquad B_4 := \|b(z)\|_{L^4_z}, \\ K_1 &:= \|\sigma'(G_1) G_2\|_{L^4}, \quad K_2 := \|\sigma''(G_1) G_2^2 - \sigma'(G_1) G_1\|_{L^4}. \end{aligned} \tag{328}$$

The constants in the branch estimates are

$$C_\star := \frac{K_2^2 (M_4^2 + B_4^2)}{32}, \qquad L_{\mathrm{fit}} := \frac{K_1^2 (M_4^2 + B_4^2)}{\sqrt{2}}. \tag{329}$$

All four Gaussian moments are finite because σ is a finite polynomial. By Gaussian rotational invariance and Minkowski's inequality,

$$\|h_\nu\|_{L^4_z} \le M_4 \quad \text{for every probability measure } \nu \text{ on } \mathbb{S}^{d_w - 1}, \qquad B_4 \le M_4. \tag{330}$$

These constants depend only on the fixed activation and teacher, and not on time, the student initialization, or particle number. The parameters $p_W, \lambda_{\mathrm{reg}}, r_{\mathrm{cap}}$ are fixed separately.

Lemma B.8 (Pole-contraction and fitting-field estimates)**.** *Assume assumptions B.1 and B.2. For every $\mu \in \mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$, the following two bounds hold:*

$$\begin{aligned} &\text{(i)} \quad DF_{\mathrm{fit}}(\mu)[(q_\star, q_\star)] \ge -C_\star Z_\pm(\mu)^2, \\ &\text{(ii)} \quad \sup_{w \in \mathbb{S}^{d_w - 1}} \|g_{i,\mathrm{fit}}(\mu, w)\|_{\mathbb{R}^{d_w}} \le L_{\mathrm{fit}} \sqrt{Z_\pm(\mu)}, \qquad i = 1, 2. \end{aligned} \tag{331}$$

The constants are those in eqs. (328) *and* (329)*. The simultaneous pole contraction is an admissible smooth transport perturbation at μ.*

Proof. Within this proof only, write $Z = Z_\pm(\mu)$ and $h = h_{\mu_1} = -h_{\mu_2}$. The teacher output is $h_{\nu^o}(z) = \int \sigma(w^\top z) \, \mathrm{d}\nu^o(w)$, where $\nu^o = \mu_1^o = \mu_2^o$ by assumption B.2.

Step 1: contraction, Gaussian derivatives, and the remainder. Recall the contraction from definition B.2 and define its output by

$$\begin{aligned} C_s(w) &= \mathrm{EXP}_{\pi_\star(w)}\big(s \mathrm{LOG}_{\pi_\star(w)} w\big), \quad \mu_{i,s} = (C_s)_\# \mu_i, \\ h_s(z) &= \int \sigma(C_s(w)^\top z) \, \mathrm{d}\mu_1(w), \qquad 0 \le s \le 1. \end{aligned} \tag{332}$$

Since $C_s(-w) = -C_s(w)$, reflection of the first pushed-forward marginal equals the second. Each point stays in its assigned cap, with $r(C_s(w)) = s r(w)$. The mass $1/2$ in each cap and oddness give

$h_0 = \frac{1}{2}\sigma(a_\star^\top z) + \frac{1}{2}\sigma(-a_\star^\top z) = 0$. Reflection also gives $\int r^2 \,\mathrm{d}\mu_2 = \int r(-w)^2 \,\mathrm{d}\mu_1 = \int r^2 \,\mathrm{d}\mu_1$. Consequently,

$$h_0 = 0, \qquad h_1 = h, \qquad \int r^2 \,\mathrm{d}\mu_1 = \frac{Z}{2}, \qquad Z_\pm(\mu_{1,s}, \mu_{2,s}) = s^2 Z. \tag{333}$$

The closed supports are compact subsets of the open caps; hence $sr(w) < r_{\mathrm{cap}}$ uniformly for s in a small two-sided neighborhood of 1. This proves admissibility, with product velocity $(q_\star, q_\star)$ at $s = 1$.

For $r(w) > 0$, eq. (255) gives

$$\|C_s(w)\|_{\mathbb{R}^{d_w}} = 1, \qquad \left\| \frac{\partial_s C_s(w)}{r(w)} \right\|_{\mathbb{R}^{d_w}} = 1, \qquad \left\langle C_s(w), \frac{\partial_s C_s(w)}{r(w)} \right\rangle_{\mathbb{R}^{d_w}} = 0. \tag{334}$$

Thus the geometry separates the first derivative's factor $r(w)$ from second-order terms carrying $r(w)^2$. Only for the next calculation, set

$$X_s := C_s(w)^\top z, \qquad Y_s := \frac{\partial_s C_s(w)^\top z}{r(w)}.$$

For $z \sim \mathcal{N}(0, I_{d_w})$, eq. (334) gives

$$\mathbb{E}_z X_s^2 = 1, \qquad \mathbb{E}_z Y_s^2 = 1, \qquad \mathbb{E}_z[X_s Y_s] = 0.$$

The pair (X_s, Y_s) is jointly Gaussian, so it has the law of independent standard normals (G_1, G_2). Recall from eq. (328) that

$$K_1 = \|\sigma'(G_1)G_2\|_{L^4}, \qquad K_2 = \|\sigma''(G_1)G_2^2 - \sigma'(G_1)G_1\|_{L^4}.$$

For $f_w(s, z) := \sigma(C_s(w)^\top z)$, the chain rule and $\partial_s^2 C_s = -r(w)^2 C_s$ from eq. (255) give

$$\begin{aligned} \partial_s f_w &= r(w)\sigma'(X_s)Y_s, \\ \partial_s^2 f_w &= r(w)^2\big[\sigma''(X_s)Y_s^2 - \sigma'(X_s)X_s\big], \\ \|\partial_s f_w\|_{L_z^4} &= r(w)\|\sigma'(G_1)G_2\|_{L^4} = K_1 r(w), \\ \|\partial_s^2 f_w\|_{L_z^4} &= r(w)^2\|\sigma''(G_1)G_2^2 - \sigma'(G_1)G_1\|_{L^4} = K_2 r(w)^2. \end{aligned} \tag{335}$$

When $r(w) = 0$, $C_s(w) = w$ is constant, so both derivatives vanish and the same norm identities hold without defining Y_s.

The endpoint identities $h_0 = 0$ and $h_1 = h$ in eq. (333) show that a linear output path would satisfy $h_1' = h_1 - h_0 = h$. We therefore measure the failure of this linearity by $e := h_1' - h$. Step 2 substitutes $h_1' = h + e$ into the fitting derivative, so the next calculation bounds $\|e\|_{L_z^4}$. For each fixed w, z, integration by parts gives

$$\int_0^1 s f_w''(s) \,\mathrm{d}s = [s f_w'(s)]_0^1 - \int_0^1 f_w'(s) \,\mathrm{d}s = f_w'(1) - f_w(1) + f_w(0).$$

Integrating this identity in w gives $\int f_w'(1) \,d\mu_1 = h_1'$, $\int f_w(1) \,d\mu_1 = h_1 = h$, and $\int f_w(0) \,d\mu_1 = h_0 = 0$. Hence

$$\begin{aligned} e := h_1' - h &= \int\int_0^1 s f_w''(s, \cdot) \,\mathrm{d}s \,\mathrm{d}\mu_1(w), \\ \|e\|_{L_z^4} &\le \int\int_0^1 s\|f_w''(s, \cdot)\|_{L_z^4} \,\mathrm{d}s \,\mathrm{d}\mu_1(w) \\ &\le K_2 \left(\int_0^1 s \,\mathrm{d}s\right) \int r(w)^2 \,\mathrm{d}\mu_1(w) \\ &= \frac{K_2}{2}\frac{Z}{2} = \frac{K_2 Z}{4}, \qquad h_1' = h + e, \end{aligned} \tag{336}$$

where the last equality uses eq. (333). The first inequality is Minkowski for the Bochner integral over s and w. The activation is a finite polynomial and the sphere is compact, so its Gaussian polynomial bounds justify these differentiations and integrals.

Step 2: the fitting derivative. The product path in eq. (332) has endpoint velocity $(q_\star, q_\star)$, so by definition of the directional derivative,

$$DF_{\mathrm{fit}}(\mu)[(q_\star, q_\star)] = \frac{d}{ds}\Big|_{s=1} F_{\mathrm{fit}}(\mu_{1,s}, \mu_{2,s}). \tag{337}$$

Recall $M_4 = \|\sigma(G_1)\|_{L^4}$ and $B_4 = \|h_{\nu^o}\|_{L^4_z} \le M_4$ from eqs. (328) and (330). In particular, $\|h\|_{L^4_z} \le \int \|\sigma(w^\top z)\|_{L^4_z} \mathrm{d}\mu_1(w) = M_4$. On the anti-symmetric branch, $h_{\mu_1} = h$ and $h_{\mu_2} = -h$, so the model output matrix is $-hh^\top$. Hence

$$F_{\mathrm{fit}}(\mu) = \frac{1}{2}\mathbb{E}_z\| - hh^\top - h_{\nu^o}h_{\nu^o}^\top\|_{\mathrm{F}}^2.$$

The Frobenius identities $\|hh^\top\|_{\mathrm{F}}^2 = \|h\|_2^4$ and $\langle hh^\top, h_{\nu^o}h_{\nu^o}^\top\rangle_{\mathrm{F}} = \langle h, h_{\nu^o}\rangle^2$ expand the branch objective as

$$F_{\mathrm{fit}}(\mu) = \frac{1}{2}\mathbb{E}_z\|h\|_2^4 + \mathbb{E}_z\langle h, h_{\nu^o}\rangle^2 + \frac{1}{2}\mathbb{E}_z\|h_{\nu^o}\|_2^4. \tag{338}$$

For the first two summands, before evaluating at $s = 1$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{2}\|h_s\|_2^4 = 2\|h_s\|_2^2\langle h_s, h_s'\rangle, \qquad \frac{\mathrm{d}}{\mathrm{d}s}\langle h_s, h_{\nu^o}\rangle^2 = 2\langle h_s, h_{\nu^o}\rangle\langle h_s', h_{\nu^o}\rangle.$$

The teacher term is constant. Evaluating at $s = 1$, using $h_1 = h$, and substituting $h_1' = h + e$ from eq. (336) into eq. (337) gives

$$\begin{aligned} DF_{\mathrm{fit}}(\mu)[(q_\star, q_\star)] = 2\mathbb{E}_z\big[&\|h\|_2^4 + \|h\|_2^2\langle h, e\rangle + \langle h, h_{\nu^o}\rangle^2 \\ &+ \langle h, h_{\nu^o}\rangle\langle e, h_{\nu^o}\rangle\big]. \end{aligned} \tag{339}$$

The two pointwise square completions are

$$2\|h\|_2^2(\|h\|_2^2 + \langle h, e\rangle) = 2\|h\|_2^2\|h + e/2\|_2^2 - \tfrac{1}{2}\|h\|_2^2\|e\|_2^2,$$

$$\begin{aligned} 2\langle h, h_{\nu^o}\rangle^2 + 2\langle h, h_{\nu^o}\rangle\langle e, h_{\nu^o}\rangle &= 2(\langle h, h_{\nu^o}\rangle + \tfrac{1}{2}\langle e, h_{\nu^o}\rangle)^2 - \tfrac{1}{2}\langle e, h_{\nu^o}\rangle^2 \\ &\ge -\tfrac{1}{2}\|e\|_2^2\|h_{\nu^o}\|_2^2. \end{aligned}$$

Discarding the nonnegative squares and applying Hölder with exponents $2, 2$ to each product of squared norms yields

$$\begin{aligned} DF_{\mathrm{fit}}(\mu)[(q_\star, q_\star)] &\ge -\tfrac{1}{2}\mathbb{E}_z[(\|h\|_2^2 + \|h_{\nu^o}\|_2^2)\|e\|_2^2] \\ &\ge -\tfrac{1}{2}(\|h\|_{L^4_z}^2 + \|h_{\nu^o}\|_{L^4_z}^2)\|e\|_{L^4_z}^2 \\ &\ge -\tfrac{1}{2}(M_4^2 + B_4^2)(K_2 Z/4)^2 = -C_\star Z^2, \end{aligned} \tag{340}$$

where $C_\star = K_2^2(M_4^2 + B_4^2)/32$, exactly as in eq. (329).

Step 3: the pointwise fitting field and output norm. For vectors $x, y \in \mathbb{R}^{d_h}$ and an increment a,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|(x + ta)y^\top - h_{\nu^o}h_{\nu^o}^\top\|_{\mathrm{F}}^2\Big|_{t=0} = \langle (xy^\top - h_{\nu^o}h_{\nu^o}^\top)y, a\rangle.$$

Thus, at $(x, y) = (h, -h)$, the first output gradient is $\|h\|_2^2 h + \langle h, h_{\nu^o}\rangle h_{\nu^o}$; the second is its negative. This differentiates one output at a time, unlike Step 2, which moves both marginals. For a unit tangent $u \in T_w\mathbb{S}^{d_w - 1}$, the vectors w, u are orthogonal unit vectors. Hence $(w^\top z, u^\top z)$ has the law of the independent standard Gaussian pair (G_1, G_2), and $\|\sigma'(w^\top z)(u^\top z)\|_{L^4_z} = K_1$ by eq. (328). Euclidean Cauchy–Schwarz, followed by Hölder with $1/(4/3) + 1/4 = 1$, gives

$$\begin{aligned} |\langle g_{i,\mathrm{fit}}(\mu, w), u\rangle| &\le \mathbb{E}_z\big[\big\|\|h\|_2^2 h + \langle h, h_{\nu^o}\rangle h_{\nu^o}\big\|_2\big\|\sigma'(w^\top z)(u^\top z)\big\|_2\big] \\ &\le K_1\big\|\|h\|_2^2 h + \langle h, h_{\nu^o}\rangle h_{\nu^o}\big\|_{L^{4/3}_z} \\ &\le K_1\big(\|h\|_{L^4_z}^3 + \|h\|_{L^4_z}\|h_{\nu^o}\|_{L^4_z}^2\big). \end{aligned}$$

For the last line, the triangle inequality and

$$\big\|\|h\|_2^3\big\|_{L^{4/3}_z} = (\mathbb{E}_z\|h\|_2^4)^{3/4} = \|h\|_{L^4_z}^3$$

handle the first term. For the second, pointwise Cauchy–Schwarz and Hölder with $1/(4/3) = 1/4 + 1/2$ give

$$\big\| |\langle h, h_{\nu^o}\rangle| \, \|h_{\nu^o}\|_2 \big\|_{L_z^{4/3}} \le \| \|h\|_2 \|h_{\nu^o}\|_2^2 \|_{L_z^{4/3}} \le \|h\|_{L_z^4} \|h_{\nu^o}\|_{L_z^4}^2.$$

Taking the supremum over unit tangents gives the same bound for $\|g_{i,\mathrm{fit}}(\mu, w)\|_{\mathbb{R}^{d_w}}$, because the fitting field belongs to $T_w \mathbb{S}^{d_w - 1}$. Finally, $h_0 = 0$ and the fundamental theorem of calculus give

$$h(z) = \int \int_0^1 \sigma'(C_s(w)^\top z)(\partial_s C_s(w)^\top z) \, \mathrm{d}s \, \mathrm{d}\mu_1(w).$$

Minkowski, eq. (335), and Cauchy–Schwarz against the probability measure μ_1 yield

$$\begin{aligned} \|h\|_{L_z^4} &\le \int \int_0^1 K_1 r(w) \, \mathrm{d}s \, \mathrm{d}\mu_1(w) \\ &\le K_1 \left(\int r(w)^2 \, \mathrm{d}\mu_1(w) \right)^{1/2} \left(\int 1 \, \mathrm{d}\mu_1 \right)^{1/2} = K_1 \sqrt{Z/2}. \end{aligned}$$

Combining this estimate with $\|h\|_{L_z^4} \le M_4$ proves

$$\|g_{i,\mathrm{fit}}(\mu, w)\|_{\mathbb{R}^{d_w}} \le K_1 (M_4^2 + B_4^2) K_1 \sqrt{Z/2} = L_{\mathrm{fit}} \sqrt{Z}, \qquad L_{\mathrm{fit}} = \frac{K_1^2 (M_4^2 + B_4^2)}{\sqrt{2}},$$

which is the constant in eq. (329). □

We now pass from the comparison transport to the gradient of the full objective and the actual product WGF.

Proposition B.4 (Quantitative radial dissipation near the target)**.** *Assume assumptions B.1 and B.2 and let $2 < p_W < 4$. Every state $\mu \in \mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$ satisfying*

$$0 < Z_\pm(\mu), \qquad C_\star Z_\pm(\mu)^{(4 - p_W)/2} \le \frac{\lambda_{\mathrm{reg}}}{2} \tag{341}$$

obeys

$$\begin{aligned} DF(\mu)[(q_\star, q_\star)] &\ge \frac{\lambda_{\mathrm{reg}}}{2} Z_\pm(\mu)^{p_W/2}, \\ \sum_{i=1}^2 \int \|g_i(\mu, w)\|_{\mathbb{R}^{d_w}}^2 \, \mathrm{d}\mu_i(w) &\ge \frac{\lambda_{\mathrm{reg}}^2}{4} Z_\pm(\mu)^{p_W - 1}. \end{aligned} \tag{342}$$

Consequently, on every time interval on which the product WGF remains in $\mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$, has positive distance, and satisfies the smallness condition in eq. (341)*,*

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_\pm(\mu_t) = -2 DF(\mu_t)[(q_\star, q_\star)] \le -\lambda_{\mathrm{reg}} Z_\pm(\mu_t)^{p_W/2}. \tag{343}$$

Among the states in $\mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$ satisfying the second inequality in eq. (341)*, the only stationary state is $\mu^\dagger := \nu_\star^\pm \otimes \nu_\star^\pm$.*

Proof. **Step 1: fix the radial margin and bound the squared gradient.** The contraction in eq. (332) has endpoint field $(q_\star, q_\star)$. By eq. (333), its distance is $s^2 Z_\pm(\mu)$, so

$$F_{\mathrm{reg}}(\mu_s) = \frac{\lambda_{\mathrm{reg}}}{p_W} s^{p_W} Z_\pm(\mu)^{p_W/2}, \qquad \frac{\mathrm{d}}{\mathrm{d}s} F_{\mathrm{reg}}(\mu_s) \Big|_{s=1} = \lambda_{\mathrm{reg}} Z_\pm(\mu)^{p_W/2}.$$

Adding the fitting bound in lemma B.8 and eq. (331) and using $Z_\pm(\mu)^2 = Z_\pm(\mu)^{(4-p_W)/2} Z_\pm(\mu)^{p_W/2}$ gives

$$\begin{aligned} DF(\mu)[(q_\star, q_\star)] &\ge \lambda_{\mathrm{reg}} Z_\pm(\mu)^{p_W/2} - C_\star Z_\pm(\mu)^2 \\ &= \left[\lambda_{\mathrm{reg}} - C_\star Z_\pm(\mu)^{(4-p_W)/2} \right] Z_\pm(\mu)^{p_W/2} \\ &\ge \frac{\lambda_{\mathrm{reg}}}{2} Z_\pm(\mu)^{p_W/2}. \end{aligned} \tag{344}$$

The exponent $(4 - p_W)/2$ is positive because $p_W < 4$; hence the $O(Z_\pm^{p_W/2})$ regularizer term dominates the $O(Z_\pm^2)$ fitting remainder sufficiently near the target.

Recall from eq. (327) that $\|(q_\star, q_\star)\|_{L^2(\mu;\mathbb{R}^{d_w})} = Z_\pm(\mu)^{1/2}$. Cauchy–Schwarz and eq. (344) therefore give

$$\frac{\lambda_{\mathrm{reg}}}{2} Z_\pm(\mu)^{p_W/2} \le DF(\mu)[(q_\star, q_\star)] \le \left(\sum_{i=1}^{2} \int \|g_i(\mu, w)\|_{\mathbb{R}^{d_w}}^2 \, \mathrm{d}\mu_i(w) \right)^{1/2} Z_\pm(\mu)^{1/2}.$$

Since $Z_\pm(\mu) > 0$, division by its square root and then squaring yield the second inequality in eq. (342).

Step 2: differentiate the distance along WGF. The WGF velocity is $-g_i$, whereas the outward cost gradient is $\nabla_{\mathbb{S}} c_\star = q_\star$ by eqs. (252) and (256). Thus the fixed-assignment formula gives

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_\pm(\mu_t) = 2 \sum_{i=1}^{2} \int \langle q_\star(w), -g_i(\mu_t, w) \rangle_{\mathbb{R}^{d_w}} \, \mathrm{d}\mu_{i,t}(w) = -2DF(\mu_t)[(q_\star, q_\star)].$$

Substitution of eq. (344) proves eq. (343). The minus sign explains why a positive outward directional derivative makes the negative gradient flow move toward the assigned poles.

Step 3: identify the stationary state. If $Z_\pm(\mu) > 0$, the squared-gradient lower bound in eq. (342) rules out stationarity. If $Z_\pm(\mu) = 0$, both marginals equal $\nu_\star^\pm$. Their output is

$$\tfrac{1}{2}\sigma(a_\star^\top z) + \tfrac{1}{2}\sigma((-a_\star)^\top z) = 0$$

by oddness of σ, so the fitting fields computed in Step 3 of lemma B.8 vanish. The regularization field $\lambda_{\mathrm{reg}} Z_\pm(\mu)^{(p_W-2)/2} q_\star$ also vanishes by eq. (294). Hence $\mu^\dagger$ is stationary. □

Second-order behavior at the target. For every $p_W > 2$, a smooth atom transport from $\mu^\dagger$ satisfies $Z_\pm(\mu_h) = O(h^2)$. Consequently,

$$F_{\mathrm{reg}}(\mu_h) = \frac{\lambda_{\mathrm{reg}}}{p_W} Z_\pm(\mu_h)^{p_W/2} = O(|h|^{p_W}) = o(h^2).$$

The regularizer has zero second variation at the target; the existence of this derivative for the normalized retractions is verified explicitly in eq. (369). The exponent condition here is $p_W > 2$; the next local-attraction result additionally requires $p_W < 3$.

Proposition B.5 (Cap confinement and convergence for a global characteristic flow)**.** *Assume assumptions B.1 and B.2. Fix $2 < p_W < 3$ and $\lambda_{\mathrm{reg}} > 0$. Choose $z_\star > 0$ such that*

$$C_\star z_\star^{(4-p_W)/2} \le \frac{\lambda_{\mathrm{reg}}}{2}, \qquad \frac{L_{\mathrm{fit}}}{\lambda_{\mathrm{reg}}} z_\star^{(3-p_W)/2} \le \frac{r_{\mathrm{cap}}}{2}. \tag{345}$$

Let $\mu_0 \in \mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$, and define

$$Z_0 := Z_\pm(\mu_0), \qquad R_0 := \max_{i=1,2} \sup_{w \in \operatorname{supp} \mu_{i,0}} r(w) < r_{\mathrm{cap}}. \tag{346}$$

Let $(\mu_t)_{t \ge 0}$ be the globally well-posed characteristic product WGF starting from μ_0: it exists for every $t \ge 0$ and is unique in the class of characteristic solutions. If $0 \le Z_0 \le z_\star$, then:

(i) The flow satisfies $\mu_t \in \mathcal{M}_- \cap \mathcal{X}_\pm(r_{\mathrm{cap}})$ for all $t \ge 0$, and

$$0 \le Z_\pm(\mu_t) \le \frac{Z_0}{\left[1 + \frac{\lambda_{\mathrm{reg}}(p_W-2)}{2} Z_0^{(p_W-2)/2} t\right]^{2/(p_W-2)}} \le Z_0. \tag{347}$$

(ii) Its support satisfies

$$\max_{i=1,2} \sup_{w \in \operatorname{supp} \mu_{i,t}} r(w) \le \overline{R} := \max\left\{ R_0, \frac{L_{\mathrm{fit}}}{\lambda_{\mathrm{reg}}} Z_0^{(3-p_W)/2} \right\} < r_{\mathrm{cap}}. \tag{348}$$

(iii) For almost every $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\sum_{i=1}^{2} \int \|g_i(\mu_t, w)\|^2_{\mathbb{R}^{d_w}} \, \mathrm{d}\mu_{i,t}(w) \leq -\frac{\lambda_{\mathrm{reg}}^2}{4} Z_{\pm}(\mu_t)^{p_W - 1} \leq 0. \tag{349}$$

The inequality is strict at every differentiability time for which $Z_{\pm}(\mu_t) > 0$. Moreover, $\mu_t \to \mu^\dagger$ in product Wasserstein distance, $\mathbf{S}(\mu_t) = 0$, and $\mathbf{A}(\mu_t) \to 0$.

If $Z_0 = 0$, the unique flow is the stationary target $\mu^\dagger$. No density or lower bound on atom masses is required.

The strict initial support bound in eq. (346) follows from eq. (241): the closed support is compact and contained in the two open caps. Both exponents in eq. (345) are positive, so such a $z_\star > 0$ exists for every fixed $\lambda_{\mathrm{reg}} > 0$. Global well-posedness specifies the time domain and uniqueness of the trajectory; cap confinement in eq. (348) is a conclusion proved below.

Proof. **Step 1: branch invariance and the mean-distance barrier.** If $Z_0 = 0$, then $\mu_0 = \mu^\dagger$. By proposition B.4, the full field vanishes there, so the constant curve is a characteristic product WGF. The assumed uniqueness makes it the trajectory in the proposition, and all claims follow. Hence assume $0 < Z_0 \leq z_\star$.

Let $\Phi_{i,t}$ be the characteristic maps of the assumed global flow and define the first cap-exit time

$$\tau_{\mathrm{cap}} := \inf \left\{ t \geq 0 : \max_{i=1,2} \sup_{w \in \operatorname{supp} \mu_{i,0}} r(\Phi_{i,t}(w)) \geq r_{\mathrm{cap}} \right\}, \qquad \inf \varnothing := \infty.$$

The strict initial support bound in eq. (346) and continuity of the characteristics give $\tau_{\mathrm{cap}} > 0$. Write $Z(t) = Z_{\pm}(\mu_t)$. On $[0, \tau_{\mathrm{cap}})$, proposition B.3 gives $\mu_t \in \mathcal{M}_-$. The fixed-assignment distance identity in eq. (343) shows that Z is continuously differentiable there: the characteristic maps, the finitely many active moments, and both gradient fields are continuous in time. Whenever $0 < Z(t) \leq Z_0$, the basin condition eq. (345) and eq. (343) give

$$\dot{Z}(t) \leq -\lambda_{\mathrm{reg}} Z(t)^{p_W/2} \leq 0. \tag{350}$$

In particular, $\dot{Z}(0) < 0$. If Z later exceeded Z_0, continuity would give a first upward return time $s > 0$ with $Z(s) = Z_0$ and $\dot{Z}(s) \geq 0$. This contradicts $\dot{Z}(s) \leq -\lambda_{\mathrm{reg}} Z_0^{p_W/2} < 0$. Thus, up to a possible cap exit,

$$Z(t) \leq Z_0, \qquad \dot{Z}(t) \leq -\lambda_{\mathrm{reg}} Z(t)^{p_W/2} \quad \text{whenever } Z(t) > 0.$$

On every interval on which $Z(t) > 0$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} Z(t)^{-(p_W-2)/2} = -\frac{p_W - 2}{2} Z(t)^{-p_W/2} \dot{Z}(t) \geq \frac{\lambda_{\mathrm{reg}}(p_W - 2)}{2}.$$

Integrating from 0 to t, and then taking the negative power $-2/(p_W - 2)$, gives the first upper bound in eq. (347). If the distance reaches zero at a finite time, the state is $\mu^\dagger$; the stationary solution and the assumed uniqueness keep it there thereafter. Thus the same bound remains valid without substituting $Z = 0$ into the preceding negative power.

Step 2: pointwise support confinement. For a characteristic $w_i(t)$, let $r_i(t) = d_{\mathbb{S}}(w_i(t), \pi_\star(w_i(t)))$. This is a pointwise pole distance, whereas $Z(t)$ is the sum of the two mean squared distances. Recall from definition B.2, lemma B.2, and eq. (256) that, with $r(w) = d_{\mathbb{S}}(w, \pi_\star(w))$,

$$q_\star(w) = \partial_s C_s(w)|_{s=1} = \nabla_{\mathbb{S}} c_\star(w) = -\mathrm{LOG}_w \pi_\star(w), \qquad \|q_\star(w)\|_{\mathbb{R}^{d_w}} = r(w).$$

Since $c_\star = r^2/2$, at every point with $r(w) > 0$,

$$q_\star(w) = r(w) \nabla_{\mathbb{S}} r(w), \qquad \nabla_{\mathbb{S}} r(w) = \frac{q_\star(w)}{r(w)}, \qquad \left\| \frac{q_\star(w)}{r(w)} \right\|_{\mathbb{R}^{d_w}} = 1.$$

Thus $q_\star/r$ is the outward unit radial tangent. By eqs. (256) and (294), the regularization field is

$$g_{i,\mathrm{reg}}(\mu, w) = \lambda_{\mathrm{reg}} Z_{\pm}(\mu)^{(p_W-2)/2} q_\star(w).$$

The fitting-field estimate in lemma B.8 and eq. (331) is uniform in the support point:

$$\|g_{i,\mathrm{fit}}(\mu_t, w_i(t))\|_{\mathbb{R}^{d_w}} \le L_{\mathrm{fit}} \sqrt{Z(t)}.$$

This is the aggregate output estimate proved in that lemma; it does not use the generally invalid pointwise comparison $r_i(t) \le \sqrt{Z(t)}$. For $r_i(t) > 0$, the characteristic equation $\dot{w}_i = -g_{i,\mathrm{fit}} - g_{i,\mathrm{reg}}$, the chain rule, and Cauchy–Schwarz therefore give

$$\begin{aligned}
\dot{r}_i(t) &= \left\langle \frac{q_\star(w_i(t))}{r_i(t)}, \dot{w}_i(t) \right\rangle_{\mathbb{R}^{d_w}} \\
&= -\left\langle \frac{q_\star(w_i(t))}{r_i(t)}, g_{i,\mathrm{fit}}(\mu_t, w_i(t)) \right\rangle_{\mathbb{R}^{d_w}} - \lambda_{\mathrm{reg}} Z(t)^{(p_W-2)/2} r_i(t) \\
&\le L_{\mathrm{fit}} \sqrt{Z(t)} - \lambda_{\mathrm{reg}} Z(t)^{(p_W-2)/2} r_i(t) \\
&= \lambda_{\mathrm{reg}} Z(t)^{(p_W-2)/2} \left[\frac{L_{\mathrm{fit}}}{\lambda_{\mathrm{reg}}} Z(t)^{(3-p_W)/2} - r_i(t) \right] \\
&\le \lambda_{\mathrm{reg}} Z(t)^{(p_W-2)/2} (\overline{R} - r_i(t)).
\end{aligned} \tag{351}$$

The last line uses $p_W < 3$, $Z(t) \le Z_0$, and the definition of $\overline{R}$ in eq. (348). At a pole, the upper right derivative satisfies

$$D^+ r_i(t) \le \|\dot{w}_i(t)\|_{\mathbb{R}^{d_w}} = \|g_{i,\mathrm{fit}}(\mu_t, w_i(t))\|_{\mathbb{R}^{d_w}} \le L_{\mathrm{fit}} \sqrt{Z(t)},$$

because the regularizer field vanishes there; hence the same differential bound holds. Multiplying by the integrating factor and using $r_i(0) \le R_0 \le \overline{R}$ yields

$$\begin{aligned}
r_i(t) &\le \overline{R} + (r_i(0) - \overline{R}) \exp\left[-\lambda_{\mathrm{reg}} \int_0^t Z(s)^{(p_W-2)/2} \, \mathrm{d}s \right] \\
&\le \overline{R} < r_{\mathrm{cap}}.
\end{aligned}$$

Here $\overline{R} < r_{\mathrm{cap}}$ follows from eqs. (345) and (346). The estimate is uniform over the compact initial supports for every $t < \tau_{\mathrm{cap}}$. If $\tau_{\mathrm{cap}} < \infty$, continuity of the characteristic maps and the preceding inequality, followed by $t \uparrow \tau_{\mathrm{cap}}$, would give

$$\max_{i=1,2} \sup_{w \in \operatorname{supp} \mu_{i,0}} r(\Phi_{i,\tau_{\mathrm{cap}}}(w)) \le \overline{R} < r_{\mathrm{cap}},$$

contradicting the definition of τ_{cap}. Hence $\tau_{\mathrm{cap}} = \infty$. Since the characteristic flow is globally defined by assumption, proposition B.3 now gives branch invariance for all $t \ge 0$, and the distance estimate from Step 1 also holds for all time.

Step 3: dissipation and convergence. For $Z(t) > 0$, the basin condition and $Z(t) \le Z_0 \le z_\star$ allow us to apply the squared-gradient estimate in eq. (342). The characteristic chain rule gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) &= \sum_{i=1}^{2} \int \langle g_i(\mu_t, w), -g_i(\mu_t, w) \rangle_{\mathbb{R}^{d_w}} \, \mathrm{d}\mu_{i,t}(w) \\
&= -\sum_{i=1}^{2} \int \|g_i(\mu_t, w)\|_{\mathbb{R}^{d_w}}^2 \, \mathrm{d}\mu_{i,t}(w) \\
&\le -\frac{\lambda_{\mathrm{reg}}^2}{4} Z(t)^{p_W - 1} < 0.
\end{aligned}$$

If $Z(t) = 0$, uniqueness keeps the flow at the stationary target, so the energy derivative is zero. This proves eq. (349) with the stated strictness.

The right-hand side of eq. (347) tends to zero, and the product Wasserstein distance from μ_t to $\mu^\dagger$ is $\sqrt{Z(t)}$. Hence $\mu_t \to \mu^\dagger$. Branch invariance gives $\mathbf{S}(\mu_t) = 0$. For each $n \in \mathcal{O}$, continuity of $w \mapsto w^{\otimes n}$ on the compact sphere gives

$$T_{i,n}(\mu_{i,t}) \longrightarrow \tfrac{1}{2} a_\star^{\otimes n} + \tfrac{1}{2} (-a_\star)^{\otimes n} = 0.$$

Since $\mathcal{O}$ is finite, $\mathbf{A}(\mu_t) \to 0$. □

B.8 Curvature away from the saddle on the antisymmetric branch

This subsection bounds the spherical product pushforward second variation at anti-symmetric branch points away from the antipodal target. On a fixed metric annulus, sufficiently strong regularization dominates the fitting curvature.

We first note that the antipodal Dirac target lies exactly at the active odd-moment origin by proposition B.1:

$$\mathbf{S}(\mu_\star^{\pm}) = 0, \qquad \mathbf{A}(\mu_\star^{\pm}) = 0.$$

The fitting term may have negative curvature, but on bounded anti-symmetric active sets this negative part is uniformly controlled; no metric annulus is needed for the following fitting estimate.

Lemma B.9 (Fitting-curvature lower bound on anti-symmetric active balls)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with the finite active odd-block representation of the fitting term.*

Fix $R_{\mathrm{act}} < \infty$. Consider anti-symmetric branch points

$$\mu = \mu_1 \otimes \mu_2 \in \mathcal{M}_-, \qquad \Theta(\mu) = (0, \mathbf{A}), \qquad \|\mathbf{A}\|_{\mathcal{H}_\mathcal{O}} \le R_{\mathrm{act}}.$$

This active-radius condition can always be met by taking $R_{\mathrm{act}} = \sqrt{|\mathcal{O}|}$. Indeed, $\|T_{i,n}\|_{\mathrm{HS}} \le \int \|w^{\otimes n}\|_{\mathrm{HS}} \,\mathrm{d}\mu_i(w) = 1$, so

$$\|A_n\|_{\mathrm{HS}} \le \frac{\|T_{1,n}\|_{\mathrm{HS}} + \|T_{2,n}\|_{\mathrm{HS}}}{2} \le 1, \qquad \|\mathbf{A}\|_{\mathcal{H}_\mathcal{O}}^2 = \sum_{n\in\mathcal{O}} \|A_n\|_{\mathrm{HS}}^2 \le |\mathcal{O}|.$$

Let $v = (v_1, v_2)$ be arbitrary smooth spherical tangent perturbations,

$$v_i(w) \in T_w \mathbb{S}^{d_w - 1}, \qquad w \cdot v_i(w) = 0.$$

The base point is on the anti-symmetric branch, but v is an arbitrary ambient product tangent field and need not be tangent to that branch. Let

$$R_{i,h}(w) \coloneqq \frac{w + h v_i(w)}{\|w + h v_i(w)\|}, \qquad \mu_{i,h} \coloneqq (R_{i,h})_{\#}\mu_i, \qquad \mu_h \coloneqq \mu_{1,h} \otimes \mu_{2,h}.$$

Then there exists a constant

$$C_{\mathrm{fit}}^{-}(R_{\mathrm{act}}) < \infty$$

depending only on R_{act}, the finite active block $\mathcal{O}$, the Hermite coefficients, and the teacher block, such that

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{fit}}(\mu_h) \ge -C_{\mathrm{fit}}^{-}(R_{\mathrm{act}}) \|v\|_{L^2(\mu)}^2, \tag{352}$$

where

$$\|v\|_{L^2(\mu)}^2 \coloneqq \|v_1\|_{L^2(\mu_1)}^2 + \|v_2\|_{L^2(\mu_2)}^2.$$

Consequently, the abstract lower bound

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{fit}}\big((R_{1,h})_{\#}\mu_1 \otimes (R_{2,h})_{\#}\mu_2\big) \ge -C_{\mathrm{fit}} \|v\|_{L^2(\mu;\mathbb{R}^{d_w})}^2$$

holds on the anti-symmetric active ball $\|\mathbf{A}\|_{\mathcal{H}_\mathcal{O}} \le R_{\mathrm{act}}$, with

$$C_{\mathrm{fit}} = C_{\mathrm{fit}}^{-}(R_{\mathrm{act}}).$$

Proof. **Step 1: write the finite-moment second-derivative formula.** The proof reduces the second variation to a finite-dimensional chain rule in $\Theta = (\mathbf{S}, \mathbf{A})$. Recall the finite active moment map from eq. (308) and the ambient fitting polynomial from eq. (310). Since the fitting term is represented by the finite active odd block, we have

$$F_{\mathrm{fit}}(\mu_h) = \widetilde{\mathcal{J}}_{\mathrm{fit}}\big(\Theta(\mu_h)\big),$$

where

$$\Theta(\mu_h) = \big(\mathbf{S}(h), \mathbf{A}(h)\big) \in \mathcal{H}_\mathcal{O} \times \mathcal{H}_\mathcal{O}.$$

At $h = 0$, since $\mu \in \mathcal{M}_-$,

$$\Theta(\mu) = (0, \mathbf{A}).$$

For $h \neq 0$, an ambient perturbation may leave the branch, so both components $\mathbf{S}(h)$ and $\mathbf{A}(h)$ are retained. Set

$$\dot{\Theta}[v] \coloneqq \frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} \Theta(\mu_h), \qquad \ddot{\Theta}[v] \coloneqq \frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} \Theta(\mu_h).$$

The ordinary chain rule gives

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{fit}}(\mu_h) = D^2 \widetilde{\mathcal{J}}_{\mathrm{fit}}(0, \mathbf{A})[\dot{\Theta}[v], \dot{\Theta}[v]] \tag{353}$$

$$+ \left\langle \nabla \widetilde{\mathcal{J}}_{\mathrm{fit}}(0, \mathbf{A}), \ddot{\Theta}[v] \right\rangle_{\mathcal{H}_{\mathcal{O}} \times \mathcal{H}_{\mathcal{O}}}. \tag{354}$$

The second term is required because the moment-coordinate path $h \mapsto \Theta(\mu_h)$ generally has the nonzero acceleration $\ddot{\Theta}[v]$.

Step 2: bound the first and second moment derivatives. The fields use the norms of $L^2(\mu_i; \mathbb{R}^{d_w})$; moment tensors use their Hilbert–Schmidt norms; and $\dot{\Theta}, \ddot{\Theta}$, and the fitting gradient use the product norm of $\mathcal{H}_{\mathcal{O}} \times \mathcal{H}_{\mathcal{O}}$. For each $n \in \mathcal{O}$,

$$T_{i,n}(h) = \int R_{i,h}(w)^{\otimes n} \, \mathrm{d}\mu_i(w).$$

Tangency $w^\top v_i(w) = 0$ gives $\|w + h v_i(w)\|^2 = 1 + h^2 \|v_i(w)\|^2$. Expanding its inverse square root in the normalized retraction yields

$$\begin{aligned} R_{i,h}(w) &= \frac{w + h v_i(w)}{\sqrt{1 + h^2 \|v_i(w)\|^2}} = w + h v_i(w) - \tfrac{1}{2} h^2 \|v_i(w)\|^2 w + O(h^3), \\ \dot{R}_{i,0}(w) &= v_i(w), \qquad \ddot{R}_{i,0}(w) = -\|v_i(w)\|^2 w, \qquad (I - w w^\top) \ddot{R}_{i,0}(w) = 0. \end{aligned} \tag{355}$$

These derivatives agree with those of $h \mapsto \mathrm{EXP}_w(h v_i(w))$ from eq. (254); the two curves need not agree beyond second order. Applying the product rule to $R_{i,h}(w)^{\otimes n}$, we obtain

$$\dot{T}_{i,n}[v_i] = n \int \mathrm{sym}\left(w^{\otimes (n-1)} \otimes v_i(w) \right) \mathrm{d}\mu_i(w), \tag{356}$$

$$\ddot{T}_{i,n}[v_i] = n(n-1) \int \mathrm{sym}\left(w^{\otimes (n-2)} \otimes v_i(w) \otimes v_i(w) \right) \mathrm{d}\mu_i(w) \tag{357}$$

$$- n \int \|v_i(w)\|^2 w^{\otimes n} \, \mathrm{d}\mu_i(w). \tag{358}$$

The term containing $n(n-1)$ is omitted when $n = 1$. Because $\|w\| = 1$, Cauchy–Schwarz yields

$$\|\dot{T}_{i,n}[v_i]\|_{\mathrm{HS}} \leq n \|v_i\|_{L^2(\mu_i)}, \tag{359}$$

and

$$\|\ddot{T}_{i,n}[v_i]\|_{\mathrm{HS}} \leq n^2 \|v_i\|^2_{L^2(\mu_i)}. \tag{360}$$

Recall the active coordinates $\Theta = (\mathbf{S}, \mathbf{A})$, where $S_n = (T_{1,n} + T_{2,n})/2$ and $A_n = (T_{1,n} - T_{2,n})/2$. The parallelogram identity gives

$$\|\dot{\Theta}\|^2 = \frac{1}{2} \sum_{n \in \mathcal{O}} (\|\dot{T}_{1,n}\|^2_{\mathrm{HS}} + \|\dot{T}_{2,n}\|^2_{\mathrm{HS}}), \qquad \|\ddot{\Theta}\|^2 = \frac{1}{2} \sum_{n \in \mathcal{O}} (\|\ddot{T}_{1,n}\|^2_{\mathrm{HS}} + \|\ddot{T}_{2,n}\|^2_{\mathrm{HS}}).$$

Here both norms on Θ are in $\mathcal{H}_{\mathcal{O}} \times \mathcal{H}_{\mathcal{O}}$. Using eqs. (359) and (360) and $a^2 + b^2 \leq (a+b)^2$ for nonnegative a, b, we obtain

$$\|\dot{\Theta}[v]\| \leq \left(\sum_{n \in \mathcal{O}} n^2 \right)^{1/2} \|v\|_{L^2(\mu)}, \tag{361}$$

$$\|\ddot{\Theta}[v]\| \leq \left(\sum_{n \in \mathcal{O}} n^4 \right)^{1/2} \|v\|^2_{L^2(\mu)}. \tag{362}$$

For example, the square of the second left-hand side is at most $\frac{1}{2}\sum_n n^4(\|v_1\|^4_{L^2(\mu_1)} + \|v_2\|^4_{L^2(\mu_2)})$, which is at most $\sum_n n^4\|v\|^4_{L^2(\mu)}$.

Step 3: choose the curvature bound and apply the chain rule. The set $\mathcal{O}$ is finite and each tensor block is finite-dimensional. Hence $\{(0, \mathbf{A}) : \|\mathbf{A}\|_{\mathcal{H}_\mathcal{O}} \le R_{\mathrm{act}}\}$ is compact. The fitting functional $\widetilde{\mathcal{J}}_{\mathrm{fit}}$ is a polynomial, so its first two derivatives are continuous and have finite suprema on this set. Define the fitting curvature constant directly by

$$\begin{aligned} C^-_{\mathrm{fit}}(R_{\mathrm{act}}) := \left(\sum_{n\in\mathcal{O}} n^2\right) \sup_{\|\mathbf{A}\|\le R_{\mathrm{act}}} \|D^2\widetilde{\mathcal{J}}_{\mathrm{fit}}(0,\mathbf{A})\|_{\mathrm{op}} \\ + \left(\sum_{n\in\mathcal{O}} n^4\right)^{1/2} \sup_{\|\mathbf{A}\|\le R_{\mathrm{act}}} \|\nabla\widetilde{\mathcal{J}}_{\mathrm{fit}}(0,\mathbf{A})\|. \end{aligned} \tag{363}$$

The gradient norm is the product $\mathcal{H}_\mathcal{O}$ norm. The Hessian norm in eq. (363) is the operator norm of

$$D^2\widetilde{\mathcal{J}}_{\mathrm{fit}}(0,\mathbf{A}) : \mathcal{H}_\mathcal{O} \times \mathcal{H}_\mathcal{O} \longrightarrow \mathcal{H}_\mathcal{O} \times \mathcal{H}_\mathcal{O}.$$

The chain rule eq. (353) and Cauchy–Schwarz now give

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}h^2} F_{\mathrm{fit}}(\mu_h)\Big|_0 &\ge -\|D^2\widetilde{\mathcal{J}}_{\mathrm{fit}}(0,\mathbf{A})\|_{\mathrm{op}}\|\dot\Theta\|^2 - \|\nabla\widetilde{\mathcal{J}}_{\mathrm{fit}}(0,\mathbf{A})\|\|\ddot\Theta\| \\ &\ge -C^-_{\mathrm{fit}}(R_{\mathrm{act}})\|v\|^2_{L^2(\mu)}. \end{aligned}$$

This proves eq. (352) with the constant defined in eq. (363). □

The metric regularizer relies on the positive Hessian of squared spherical distance inside each fixed cap.

Lemma B.10 (Spherical Hessian of the fixed squared-distance cost)**.** ***Assumptions used.*** *assumption B.1.*

Let

$$c_y(w) = \frac{1}{2} d_{\mathbb{S}}(w,y)^2, \qquad y \in \mathbb{S}^{d_w-1}.$$

On the geodesic ball $d_{\mathbb{S}}(w,y) \le r_{\mathrm{cap}} < \pi/2$, the spherical Hessian satisfies

$$\nabla^2_{\mathbb{S}} c_y(w)[v,v] \ge \kappa_{\mathrm{cap}}\|v\|^2, \qquad \kappa_{\mathrm{cap}} \coloneqq r_{\mathrm{cap}}\cot r_{\mathrm{cap}} > 0. \tag{364}$$

Consequently, for $c_\star(w) = \frac{1}{2} d_{\mathbb{S}}(w, \pi_\star(w))^2$,

$$\nabla^2_{\mathbb{S}} c_\star(w)[v,v] \ge \kappa_{\mathrm{cap}}\|v\|^2 \tag{365}$$

for every $w \in K_+(r_{\mathrm{cap}}) \cup K_-(r_{\mathrm{cap}})$ and every $v \in T_w\mathbb{S}^{d_w-1}$.

Proof. For $0 < r = d_{\mathbb{S}}(w,y) < \pi/2$, define $e = (y - \cos r\, w)/\sin r$. Since $w^\top y = \cos r$,

$$w^\top e = \frac{w^\top y - \cos r}{\sin r} = 0, \qquad \|e\|^2_{\mathbb{R}^{d_w}} = \frac{1 - 2\cos r\, y^\top w + \cos^2 r}{\sin^2 r} = 1.$$

Thus $e \in T_w\mathbb{S}^{d_w-1}$ is the unit tangent pointing from w toward y. Put $x(t) = y^\top \mathrm{EXP}_w(tv)$ and $f(x) = \frac{1}{2}(\arccos x)^2$. By eq. (254),

$$x(0) = \cos r, \quad x'(0) = y^\top v = \sin r\,\langle e, v\rangle, \quad x''(0) = -\|v\|^2\cos r,$$

and direct differentiation gives

$$f'(x) = -\frac{\arccos x}{\sqrt{1-x^2}}, \qquad f''(x) = \frac{1}{1-x^2} - \frac{x\arccos x}{(1-x^2)^{3/2}}.$$

Substituting $x = \cos r$ yields

$$f'(\cos r) = -\frac{r}{\sin r}, \qquad f''(\cos r) = \frac{\sin r - r\cos r}{\sin^3 r}.$$

For sufficiently small $|t|$, the EXP path does not reach the antipode of y, so the principal $\arccos$ gives the spherical distance. Hence

$$c_y(\mathrm{EXP}_w(tv)) = \frac{1}{2} d_{\mathbb{S}}(\mathrm{EXP}_w(tv), y)^2 = \frac{1}{2}[\arccos\big(y^\top \mathrm{EXP}_w(tv)\big)]^2 = f(x(t)).$$

The one-variable chain rule now gives

$$\begin{aligned} \nabla^2_{\mathbb{S}} c_y(w)[v,v] &= f''(\cos r)(x'(0))^2 + f'(\cos r) x''(0) \\ &= r\cot r\, \|v\|^2 + (1 - r\cot r)\langle v, e\rangle^2. \end{aligned} \tag{366}$$

To display the eigendirections, set $v_\parallel = \langle v, e\rangle e$ and $v_\perp = v - v_\parallel$. Since $\|v\|^2_{\mathbb{R}^{d_w}} = \|v_\parallel\|^2_{\mathbb{R}^{d_w}} + \|v_\perp\|^2_{\mathbb{R}^{d_w}}$,

$$\nabla^2_{\mathbb{S}} c_y(w)[v,v] = \|v_\parallel\|^2_{\mathbb{R}^{d_w}} + r\cot r\, \|v_\perp\|^2_{\mathbb{R}^{d_w}}.$$

The radial eigenvalue is therefore 1, and the orthogonal eigenvalue is $r\cot r$. Moreover,

$$(r\cot r)' = \frac{\sin r\cos r - r}{\sin^2 r} \le 0, \qquad r - \sin r\cos r = \int_0^r 2\sin^2 s\,\mathrm{d}s \ge 0, \qquad \lim_{r\to 0} r\cot r = 1.$$

The integral identity shows that $(r\cot r)' \le 0$. Therefore $1 \ge r\cot r \ge r_{\mathrm{cap}}\cot r_{\mathrm{cap}} > 0$ for $0 < r \le r_{\mathrm{cap}}$, so both eigenvalues are at least κ_{cap}. At $w = y$, the identity $c_y(\mathrm{EXP}_y(tv)) = t^2\|v\|^2_{\mathbb{R}^{d_w}}/2$ for small t gives Hessian $\|v\|^2_{\mathbb{R}^{d_w}}$ without introducing e. On each cap, $\pi_\star(w)$ is the fixed point $a_\star$ or $-a_\star$, so the calculation does not differentiate across a cap boundary. Substituting $y = a_\star$ or $y = -a_\star$ gives the claim for the fixed-assignment cost $c_\star(w) = d_{\mathbb{S}}(w, \pi_\star(w))^2/2$ from eqs. (241) and (252). □

Combining fixed assignment with the spherical distance Hessian gives a positive spherical product pushforward second variation for the metric regularizer.

Lemma B.11 (Metric-power regularizer Hessian on fixed-assignment caps)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with lemmas B.1 and B.10.*

Let $\mu_i \in \mathcal{P}_\pm(r_{\mathrm{cap}})$, and let $v = (v_1, v_2)$ be smooth tangent perturbations whose spherical retractions preserve the fixed-assignment caps for sufficiently small $|h|$. Then

$$D^2_{W_2} F_{\mathrm{reg}}(\mu_1,\mu_2)[v,v] \ge \lambda_{\mathrm{reg}}\,\kappa_{\mathrm{cap}} Z_\pm(\mu_1,\mu_2)^{\delta_W/2}\|v\|^2_{L^2(\mu)}. \tag{367}$$

In particular, if $Z_\pm(\mu_1,\mu_2) \ge \zeta^2_{\min}$, then

$$D^2_{W_2} F_{\mathrm{reg}}(\mu_1,\mu_2)[v,v] \ge \lambda_{\mathrm{reg}}\,\kappa_{\mathrm{cap}} \zeta^{\delta_W}_{\min}\|v\|^2_{L^2(\mu)}. \tag{368}$$

At $Z_\pm = 0$, the regularizer has zero quadratic variation because $p_W = 2 + \delta_W > 2$.

Proof. **Step 1: bound the positive-distance second derivative.** By eq. (355), the normalized retraction used here and the spherical EXP path have the same first and second derivatives at $h = 0$. Since the regularizer is smooth when $Z_\pm > 0$, its second derivative along the retraction is therefore the one computed for the EXP path in eq. (295). In that formula, the second term is nonnegative because $p_W - 2 = \delta_W > 0$. Substituting the pointwise Hessian bound eq. (365) into the first term gives

$$\begin{aligned} D^2_{W_2} F_{\mathrm{reg}}(\mu_1,\mu_2)[v,v] &\ge \lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}} Z_\pm^{p_W/2-1} \sum_{i=1}^2 \int \|v_i(w)\|^2_{\mathbb{R}^{d_w}}\,\mathrm{d}\mu_i(w) \\ &= \lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}} Z_\pm^{\delta_W/2}\|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}, \end{aligned}$$

which is eq. (367) for $Z_\pm > 0$. If $Z_\pm \ge \zeta^2_{\min}$, then $\delta_W > 0$ implies

$$Z_\pm^{\delta_W/2} \ge (\zeta^2_{\min})^{\delta_W/2} = \zeta^{\delta_W}_{\min},$$

and eq. (368) follows.

Step 2: compute the ordinary second derivative at the target. For $Z_\pm = 0$, both marginals equal $\nu^\pm_\star = (\delta_{a_\star} + \delta_{-a_\star})/2$. At a pole $y = \varepsilon a_\star$, tangency gives $y^\top R_{i,h}(y) = (1 + h^2\|v_i(y)\|^2)^{-1/2}$, so its

distance from the pole is $\arctan(|h|\|v_i(y)\|)$. The fixed-assignment formula eq. (251) consequently gives

$$Z_{\pm}(\mu_h) = \frac{1}{2}\sum_{i=1}^{2}\sum_{\varepsilon\in\{-1,1\}} \arctan^2(h\|v_i(\varepsilon a_\star)\|) = h^2\|v\|^2_{L^2(\mu)} + O(h^4). \tag{369}$$

If $\|v\|_{L^2(\mu)} = 0$, all four atoms stay fixed. Otherwise $a(h) := Z_{\pm}(\mu_h)/h^2$ extends to a smooth positive function at zero with $a(0) = \|v\|^2_{L^2(\mu)}$. Set $b(h) := (\lambda_{\mathrm{reg}}/p_W)a(h)^{p_W/2}$, so the regularizer is $|h|^{p_W}b(h)$. For $h \neq 0$, differentiating twice gives

$$F'_{\mathrm{reg}}(h) = p_W \operatorname{sgn}(h)|h|^{p_W-1}b(h) + |h|^{p_W}b'(h),$$
$$F''_{\mathrm{reg}}(h) = p_W(p_W-1)|h|^{p_W-2}b(h) + 2p_W \operatorname{sgn}(h)|h|^{p_W-1}b'(h) + |h|^{p_W}b''(h).$$

The difference quotients at zero and these formulas show that the regularizer is C^2, with both derivatives zero there, because $p_W > 2$. This proves the bound at the target as well. □

The next proposition is an abstract combination statement: at a fixed state, it adds any available fitting lower bound to the metric lower bound. On an anti-symmetric active ball, the fitting input has already been verified in lemma B.9; the subsequent corollary makes that substitution.

Proposition B.6 (Ambient second-variation bound on fixed metric annuli)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with lemma B.11.*

Fix $0 < \zeta_{\min} < \infty$, and consider a state in a fixed-assignment cap region satisfying

$$\mu_i \in \mathcal{P}_{\pm}(r_{\mathrm{cap}}), \qquad Z_{\pm}(\mu_1,\mu_2) \geq \zeta^2_{\min}.$$

Suppose that, on the region under consideration, the fitting lower bound

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F_{\mathrm{fit}}\big((R_{1,h})_{\#}\mu_1 \otimes (R_{2,h})_{\#}\mu_2\big) \geq -C_{\mathrm{fit}}\|v\|^2_{L^2(\mu)} \tag{370}$$

is available for every smooth tangent perturbation $v = (v_1, v_2)$ whose retractions stay in the fixed-assignment caps for small $|h|$. Then the full objective satisfies the following pointwise bound at that state:

$$\langle v, \mathcal{H}_{\mathbb{S},\mu}v\rangle_{L^2(\mu;\mathbb{R}^{d_w})} \geq \Big(\lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}}\zeta^{\delta_W}_{\min} - C_{\mathrm{fit}}\Big)\|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}. \tag{371}$$

In particular, if

$$\lambda_{\mathrm{reg}} > \frac{C_{\mathrm{fit}}}{\kappa_{\mathrm{cap}}\zeta^{\delta_W}_{\min}},$$

then the spherical product pushforward second variation is strictly positive in every nonzero admissible tangent direction at each point of the region.

Proof. By lemma B.6, the spherical product pushforward second variation is

$$D^2_{W_2}F(\mu)[v,v] = \langle v, \mathcal{H}_{\mathbb{S},\mu}v\rangle_{L^2(\mu;\mathbb{R}^{d_w})}.$$

The two inputs give

$$D^2_{W_2}F_{\mathrm{fit}}(\mu)[v,v] \geq -C_{\mathrm{fit}}\|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}, \qquad \text{by eq. (370),}$$
$$D^2_{W_2}F_{\mathrm{reg}}(\mu)[v,v] \geq \lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}}\zeta^{\delta_W}_{\min}\|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}, \qquad \text{by eq. (368).}$$

Adding these inequalities proves eq. (371). □

Remark B.5 (Pointwise scope)**.** *The annulus need not be geodesically convex. A strong-convexity inequality between two distinct states follows only along a comparison geodesic that remains in a region where eq.* (371) *holds at every point.*

Specializing this bound to the anti-symmetric branch gives positive ambient curvature at the prescribed fixed-assignment annular states.

Corollary B.2 (Ambient curvature at anti-symmetric annular states)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with proposition B.6.*

Fix $R_{\mathrm{act}} < \infty$. Let $\mu_- \in \mathcal{M}_-$ satisfy $\|\mathbf{A}(\mu_-)\|_{\mathcal{H}_\mathcal{O}} \le R_{\mathrm{act}}$ and

$$\mu_{-,i} \in \mathcal{P}_\pm(r_{\mathrm{cap}}), \qquad Z_\pm(\mu_{-,1}, \mu_{-,2}) \ge \zeta_{\min}^2.$$

Then eq. (371) *holds with $C_{\mathrm{fit}} = C_{\mathrm{fit}}^-(R_{\mathrm{act}})$ from eqs.* (352) *and* (363)*. In particular, every nonzero admissible product tangent field has positive spherical product pushforward second variation when $\lambda_{\mathrm{reg}}\kappa_{\mathrm{cap}}\zeta_{\min}^{\delta_W} > C_{\mathrm{fit}}^-(R_{\mathrm{act}})$. The field need not be tangent to $\mathcal{M}_-$.*

B.9 Descent direction to escape the saddle point on the antisymmetric branch

This subsection uses the transverse teacher imbalance (assumption B.3) to construct a same-density perturbation that escapes from the antipodal saddle.

The imbalance assumption says that the teacher has a coherent first-order transverse bias across two-sided latitude slices around the antipodal axis. Here we show that this bias is visible to the active odd Hermite features, and therefore produces a same-density tangent direction that decreases the fitting loss to second order at the antipodal Dirac point.

Remark B.6 (Comparison with the zero-output saddle argument of Yamamoto et al. (2025))**.** *The algebraic mechanism is the same as the known zero-output saddle calculation in the matrix-decomposition example (Proposition G.1 in Yamamoto et al. (2025)): when the student output vanishes, a perturbation that creates a first-order feature g changes the product output only at order h^2, producing the quadratic cross term*

$$-\mathbb{E}_z[(h^o(z)^\top g(z))^2].$$

However, the affine mixture curve used there does not preserve our product spherical ansatz. We therefore rebuild the argument with a Dirac-supported spherical tangent perturbation at the antipodal pair. The transverse imbalance assumption is precisely the condition ensuring that one such tangent feature is visible to the teacher.

The first step turns the sliced transverse imbalance assumption into a nonzero active-weighted transverse teacher moment.

Lemma B.12 (Quantitative control of the active-weighted transverse moment)**.** ***Assumptions used.*** *definition B.1 and assumption B.3.*

Set $a = a_\star$, the anchor in that imbalance assumption. Define

$$\kappa_\mathcal{O}(r) \coloneqq \frac{1}{2}\sum_{n\in\mathcal{O}} n\|a_n\|_2^2 r^{n-1}, \qquad r \in [0,1].$$

Assume at least one active coefficient $\|a_n\|_2$ is nonzero. Since every $n \in \mathcal{O}$ is odd,

$$\kappa_\mathcal{O}(r) > 0 \qquad \forall r \in (0,1].$$

Let

$$b_a^o \coloneqq \int_{\mathbb{S}^{d_w-1}} \kappa_\mathcal{O}(|a^\top w|)\Pi_{a^\perp} w \,\mathrm{d}\nu^o(w) \in a^\perp.$$

Equivalently,

$$b_a^o = \int_{[0,1]} \kappa_\mathcal{O}(r)\,\mathrm{d}\mathfrak{m}_a^o(r).$$

Define

$$\kappa_{\min}(r_{\mathrm{vis}}) \coloneqq \inf_{r\in[r_{\mathrm{vis}},1]} \kappa_\mathcal{O}(r).$$

Then

$$\kappa_{\min}(r_{\mathrm{vis}}) > 0$$

and

$$\|b_a^o\| \ge \langle e_\star, b_a^o\rangle \ge \kappa_{\min}(r_{\mathrm{vis}})\gamma_\perp.$$

In particular,

$$b_a^o \ne 0.$$

Proof. We first rewrite the weighted transverse moment using the latitude disintegration. By definition B.1, for every bounded Borel function $\psi : [0,1] \to \mathbb{R}$,

$$\int_{\mathbb{S}^{d_w-1}} \psi(r_a(w))\Pi_{a^\perp} w \, \mathrm{d}\nu^o(w) = \int_{[0,1]} \psi(r) \, \mathrm{d}\mathfrak{m}_a^o(r).$$

Taking

$$\psi(r) = \kappa_{\mathcal{O}}(r)$$

gives

$$b_a^o = \int_{[0,1]} \kappa_{\mathcal{O}}(r) \, \mathrm{d}\mathfrak{m}_a^o(r).$$

Therefore,

$$\langle e_\star, b_a^o \rangle = \int_{[0,1]} \kappa_{\mathcal{O}}(r) \, \mathrm{d}\mathfrak{m}_{a,e_\star}^o(r).$$

Next, since every $n \in \mathcal{O}$ is odd, $n-1$ is even. Hence

$$r^{n-1} > 0 \qquad \forall r \in (0,1]$$

for every active degree n. Because at least one active coefficient is nonzero,

$$\kappa_{\mathcal{O}}(r) > 0 \qquad \forall r \in (0,1].$$

Since $r_{\mathrm{vis}} \in (0,1)$ and $\kappa_{\mathcal{O}}$ is continuous,

$$\kappa_{\min}(r_{\mathrm{vis}}) = \inf_{r \in [r_{\mathrm{vis}},1]} \kappa_{\mathcal{O}}(r) > 0.$$

By assumption B.3, $\mathfrak{m}_{a,e_\star}^o$ is a positive finite measure on $[0,1]$. Thus

$$\langle e_\star, b_a^o \rangle = \int_{[0,1]} \kappa_{\mathcal{O}}(r) \, \mathrm{d}\mathfrak{m}_{a,e_\star}^o(r) \geq \int_{[r_{\mathrm{vis}},1]} \kappa_{\mathcal{O}}(r) \, \mathrm{d}\mathfrak{m}_{a,e_\star}^o(r).$$

Using the definition of $\kappa_{\min}(r_{\mathrm{vis}})$, we obtain

$$\int_{[r_{\mathrm{vis}},1]} \kappa_{\mathcal{O}}(r) \, \mathrm{d}\mathfrak{m}_{a,e_\star}^o(r) \geq \kappa_{\min}(r_{\mathrm{vis}}) \mathfrak{m}_{a,e_\star}^o([r_{\mathrm{vis}},1]).$$

The quantitative strict imbalance gives

$$\mathfrak{m}_{a,e_\star}^o([r_{\mathrm{vis}},1]) \geq \gamma_\perp.$$

Consequently,

$$\langle e_\star, b_a^o \rangle \geq \kappa_{\min}(r_{\mathrm{vis}}) \gamma_\perp.$$

Finally, since $\|e_\star\| = 1$,

$$\|b_a^o\| \geq \langle e_\star, b_a^o \rangle.$$

This proves the claim. □

We next encode how a tangent displacement of the antipodal atoms is seen by the teacher feature.

Definition B.5 (Tangential visibility matrix at an antipodal Dirac pair)**.** *Let*

$$\nu_a^\pm := \frac{1}{2}\delta_a + \frac{1}{2}\delta_{-a}.$$

For $s \in a^\perp$, define the Dirac-supported active differential

$$\mathcal{U}_a(s)_n := \frac{n}{2} \mathrm{sym}\left(a^{\otimes(n-1)} \otimes s\right), \qquad n \in \mathcal{O}.$$

Let $g_{a,s}$ be the corresponding first-order student feature,

$$g_{a,s}(z) := \sum_{n \in \mathcal{O}} a_n \langle \mathcal{H}_n(z), \mathcal{U}_a(s)_n \rangle_{\mathrm{HS}}. \tag{372}$$

This uses the finite expansion in (226)*. Recall the teacher from assumption B.2:*

$$h^o(z) = h_{\nu^o}(z) = \sum_{n\in\mathcal{O}} a_n \langle \mathcal{H}_n(z), T_n^o\rangle_{\mathrm{HS}}, \qquad T_n^o = \int w^{\otimes n}\, d\nu^o(w).$$

Equivalently, define the linear map

$$J_a(z) : a^\perp \to \mathbb{R}^{d_h}, \qquad J_a(z)s = g_{a,s}(z).$$

For the teacher feature h^o, define the positive semidefinite operator on $a^\perp$

$$\mathsf{B}_a^o \coloneqq \mathbb{E}_z\left[J_a(z)^\top h^o(z) h^o(z)^\top J_a(z)\right].$$

We say that the teacher is tangentially visible at the antipodal pair $\{\pm a\}$ if

$$\lambda_{\max}(\mathsf{B}_a^o) > 0.$$

The nonzero transverse teacher moment implies that some tangent direction has nonzero correlation with the teacher output.

Lemma B.13 (Active-weighted transverse moment implies tangential visibility)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with definition B.5.*

Let

$$b_a^o \coloneqq \int_{\mathbb{S}^{d_w-1}} \kappa_{\mathcal{O}}(|a^\top w|)\Pi_{a^\perp} w\, \mathrm{d}\nu^o(w).$$

Then

$$\mathsf{B}_a^o \succeq b_a^o (b_a^o)^\top.$$

Consequently,

$$\lambda_{\max}(\mathsf{B}_a^o) \geq \|b_a^o\|^2.$$

In particular, if $b_a^o \neq 0$, then the teacher is tangentially visible at $\{\pm a\}$.

Proof. For $s \in a^\perp$, Hermite orthogonality gives

$$\mathbb{E}_z\left[g_{a,s}(z)^\top h^o(z)\right] = \sum_{n\in\mathcal{O}} \|a_n\|_2^2 \left\langle \mathcal{U}_a(s)_n, T_n^o\right\rangle_{\mathrm{HS}}.$$

By the definition of $\mathcal{U}_a(s)$,

$$\langle \mathcal{U}_a(s)_n, T_n^o\rangle_{\mathrm{HS}} = \frac{n}{2}\left\langle \mathrm{sym}\left(a^{\otimes(n-1)}\otimes s\right), T_n^o\right\rangle_{\mathrm{HS}}.$$

Since T_n^o is symmetric,

$$\left\langle \mathrm{sym}\left(a^{\otimes(n-1)}\otimes s\right), T_n^o\right\rangle_{\mathrm{HS}} = \left\langle s, T_n^o \otimes_{n-1} a^{\otimes(n-1)}\right\rangle.$$

Using

$$T_n^o = \int_{\mathbb{S}^{d_w-1}} w^{\otimes n}\, \mathrm{d}\nu^o(w),$$

and $s \in a^\perp$, we obtain

$$\left\langle s, T_n^o \otimes_{n-1} a^{\otimes(n-1)}\right\rangle = \int_{\mathbb{S}^{d_w-1}} (a^\top w)^{n-1}\langle s, \Pi_{a^\perp} w\rangle\, \mathrm{d}\nu^o(w).$$

Since n is odd, $n-1$ is even, hence

$$(a^\top w)^{n-1} = |a^\top w|^{n-1}.$$

Therefore

$$\mathbb{E}_z\left[g_{a,s}(z)^\top h^o(z)\right] = \langle s, b_a^o\rangle.$$

Set

$$X(z) \coloneqq J_a(z)^\top h^o(z) \in a^\perp.$$

Then

$$\mathsf{B}_a^o = \mathbb{E}_z[X(z)X(z)^\top], \qquad \mathbb{E}_z[X(z)] = b_a^o.$$

By Jensen's inequality in the Loewner order,

$$\mathsf{B}_a^o = \mathbb{E}_z[X(z)X(z)^\top] \succeq \mathbb{E}_z[X(z)]\mathbb{E}_z[X(z)]^\top = b_a^o(b_a^o)^\top.$$

Thus

$$\lambda_{\max}(\mathsf{B}_a^o) \geq \|b_a^o\|^2.$$

□

Finally, the visible tangent direction yields a same-density perturbation with negative second-order fitting curvature at the antipodal saddle. The proposition is an existence statement for a negative-curvature direction. It does not by itself prove that the Hessian-guided perturbation selects this particular same-density direction; that selection issue belongs to the perturbation analysis rather than to the geometric verification.

Proposition B.7 (Dirac-supported transverse descent from unweighted transverse imbalance)**.** ***Assumptions used.*** *assumptions B.1 to B.3.*

Assume $\mathcal{O} \neq \varnothing$. Set $a = a_\star$, so that the perturbed Dirac pair is the fixed regularizer target. Let

$$\nu_a^\pm := \frac{1}{2}\delta_a + \frac{1}{2}\delta_{-a}, \qquad \mu^\dagger := \nu_a^\pm \otimes \nu_a^\pm.$$

By lemma B.12, lemma B.13,

$$\gamma_{\mathrm{esc}} := \lambda_{\max}(\mathsf{B}_a^o) \geq \kappa_{\min}(r_{\mathrm{vis}})^2\gamma_\perp^2 > 0.$$

Let $s_\star \in a^\perp$ be a unit eigenvector associated with γ_{esc}. Then there exists a same-density spherical pushforward perturbation

$$\mu_h = \mu_{1,h} \otimes \mu_{2,h}, \qquad \mu_{1,h} = \mu_{2,h}, \qquad \mu_{i,0} = \nu_a^\pm,$$

such that

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} F(\mu_h) = 0$$

and

$$F(\mu_h) = F(\mu^\dagger) - h^2\gamma_{\mathrm{esc}} + o(h^2).$$

In particular,

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} F(\mu_h) = -2\gamma_{\mathrm{esc}} < 0.$$

Moreover, for the metric regularizer

$$F_{\mathrm{reg}}(\mu_1, \mu_2) = \frac{\lambda_{\mathrm{reg}}}{p_W} Z_\pm(\mu_1, \mu_2)^{p_W/2}, \qquad p_W = 2 + \delta_W,$$

one has

$$F_{\mathrm{reg}}(\mu_h) = O(|h|^{2+\delta_W}) = o(h^2).$$

Thus the metric regularizer does not affect the quadratic descent term at the antipodal Dirac saddle.

Proof. **Step 1: compute the moments at the antipodal pair.** We first compute the active odd moments at the antipodal Dirac pair. For every odd $n \in \mathcal{O}$,

$$T_n(\nu_a^\pm) = \frac{1}{2}a^{\otimes n} + \frac{1}{2}(-a)^{\otimes n} = 0.$$

Therefore

$$\Theta(\mu^\dagger) = (0, 0).$$

Step 2: construct the visible spherical perturbation.

Recall that

$$a^\perp = \{s \in \mathbb{R}^{d_w} : \langle s, a\rangle_{\mathbb{R}^{d_w}} = 0\} = T_a\mathbb{S}^{d_w-1} = T_{-a}\mathbb{S}^{d_w-1}, \qquad \Pi_{a^\perp} = I - aa^\top,$$

where the projection is the one used in definition B.1. Thus the visible unit vector $s_\star \in a^\perp$ from the statement is a common tangent direction at both poles. Set

$$v(w) := \frac{1}{2}(I - ww^\top)s_\star, \qquad u_+ := v(a) = \frac{s_\star}{2}, \qquad u_- := v(-a) = \frac{s_\star}{2}. \tag{373}$$

Indeed, $w^\top v(w) = 0$, so $v(w) \in T_w\mathbb{S}^{d_w-1}$. Moreover, substituting $w = a$ and $w = -a$ and using $a^\top s_\star = 0$ gives

$$v(a) = \tfrac{1}{2}(I - aa^\top)s_\star = \tfrac{1}{2}s_\star, \qquad v(-a) = \tfrac{1}{2}(I - aa^\top)s_\star = \tfrac{1}{2}s_\star.$$

Hence the two atoms move in the same ambient transverse direction and $u_+ + u_- = s_\star$. The factor $1/2$ in v compensates for the two atom masses $1/2$, so that their summed first-order moment change is exactly the differential $\mathcal{U}_a(s_\star)$ defined in definition B.5. Use it in both marginals and define $\mu_{1,h} = \mu_{2,h} := (R_h)_\# \nu_a^\pm$, where, for $w \in \{\pm a\}$,

$$R_h(w) = \frac{w + hv(w)}{\|w + hv(w)\|}.$$

Since $v(w) \in T_w\mathbb{S}^{d_w-1}$, we have $w^\top v(w) = 0$, and hence

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} R_h(w) = v(w).$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} R_h(w)^{\otimes n} = n\,\mathrm{sym}\left(w^{\otimes(n-1)} \otimes v(w)\right).$$

Consequently,

$$\dot{T}_n[v] := \frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} T_n((R_h)_\# \nu_a^\pm) \tag{374}$$

$$= \frac{n}{2}\mathrm{sym}\left(a^{\otimes(n-1)} \otimes u_+\right) + \frac{n}{2}\mathrm{sym}\left((-a)^{\otimes(n-1)} \otimes u_-\right). \tag{375}$$

Because n is odd, $n-1$ is even, and therefore

$$(-a)^{\otimes(n-1)} = a^{\otimes(n-1)}.$$

Hence

$$\dot{T}_n[v] = \frac{n}{2}\mathrm{sym}\left(a^{\otimes(n-1)} \otimes (u_+ + u_-)\right) = \frac{n}{2}\mathrm{sym}\left(a^{\otimes(n-1)} \otimes s_\star\right) = \mathcal{U}_a(s_\star)_n.$$

Since $\mu_{1,h} = \mu_{2,h}$, we have

$$\dot{\mathbf{S}}[v] = \frac{\dot{\mathbf{T}}_1[v] + \dot{\mathbf{T}}_2[v]}{2} = \mathcal{U}_a(s_\star), \qquad \dot{\mathbf{A}}[v] = \frac{\dot{\mathbf{T}}_1[v] - \dot{\mathbf{T}}_2[v]}{2} = 0.$$

Thus this path leaves the active origin in the symmetric coordinate $\mathbf{S}$ with velocity $\mathcal{U}_a(s_\star)$, while its anti-symmetric coordinate $\mathbf{A}$ remains zero.

Step 3: expand the fitting term. Now we compute the model block. By the active Hermite product formula,

$$\widetilde{\mathcal{M}}_p(\mathbf{S}, \mathbf{A}) = \sum_{\substack{n,m\in\mathcal{O} \\ 0\le r\le \min(n,m) \\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left((S_n + A_n) \otimes_r (S_m - A_m)\right).$$

Along the same-density perturbation above,

$$\mathbf{A}(h) = 0$$

for every h. The normalized atom maps constructed from the field in eq. (373) are smooth near $h = 0$, so their finitely many tensor moments are C^2 in h. Step 1 gives $\mathbf{S}(0) = 0$, and the derivative calculation in Step 2 gives $\dot{\mathbf{S}}(0) = \mathcal{U}_a(s_\star)$. Taylor's formula therefore gives

$$\mathbf{S}(h) = h\,\mathcal{U}_a(s_\star) + O(h^2).$$

Therefore

$$\widetilde{\mathcal{M}}_p(\Theta(\mu_h)) = \widetilde{\mathcal{M}}_p(\mathbf{S}(h), 0) \tag{376}$$

$$= \sum_{\substack{n,m\in\mathcal{O}\\ 0\le r\le \min(n,m)\\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left(S_n(h) \otimes_r S_m(h)\right) \tag{377}$$

$$= h^2 \underbrace{\sum_{\substack{n,m\in\mathcal{O}\\ 0\le r\le \min(n,m)\\ p=n+m-2r}} c_{n,m,r}(a_n a_m^\top) \otimes \mathrm{sym}\left(\mathcal{U}_a(s_\star)_n \otimes_r \mathcal{U}_a(s_\star)_m\right)}_{=:\mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star))} + o(h^2). \tag{378}$$

We now expand the fitting term. Recall that

$$F_{\mathrm{fit}}(\mu_h) = \frac{1}{2}\sum_{p\ge 0}\left\|\widetilde{\mathcal{M}}_p(\Theta(\mu_h)) - \mathcal{M}_p^o\right\|_{\mathrm{HS}}^2.$$

Using

$$\widetilde{\mathcal{M}}_p(\Theta(\mu_h)) = h^2\mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star)) + o(h^2),$$

we get

$$\frac{1}{2}\left\|\widetilde{\mathcal{M}}_p(\Theta(\mu_h)) - \mathcal{M}_p^o\right\|_{\mathrm{HS}}^2 = \frac{1}{2}\left\|-\mathcal{M}_p^o + h^2\mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star)) + o(h^2)\right\|_{\mathrm{HS}}^2 \tag{379}$$

$$= \frac{1}{2}\|\mathcal{M}_p^o\|_{\mathrm{HS}}^2 - h^2\left\langle \mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star)), \mathcal{M}_p^o\right\rangle_{\mathrm{HS}} + o(h^2). \tag{380}$$

Summing over $p \ge 0$ proves

$$F_{\mathrm{fit}}(\mu_h) = F_{\mathrm{fit}}(\mu^\dagger) - h^2\sum_{p\ge 0}\left\langle \mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star)), \mathcal{M}_p^o\right\rangle_{\mathrm{HS}} + o(h^2).$$

Step 4: identify the visibility coefficient. It remains to identify the escape coefficient with the visibility matrix. Use the first-order student feature $g_{a,s}(z)$ defined by the finite Hermite sum in (372). By the definition of the Hermite active blocks,

$$g_{a,s}(z)g_{a,s}(z)^\top$$

has p-th active Hermite block

$$\mathcal{Q}_{p,+}(\mathcal{U}_a(s)).$$

Similarly,

$$h^o(z)h^o(z)^\top$$

has p-th active Hermite block $\mathcal{M}_p^o$. Therefore, by orthogonality of the Hermite blocks,

$$\sum_{p\ge 0}\left\langle \mathcal{Q}_{p,+}(\mathcal{U}_a(s)), \mathcal{M}_p^o\right\rangle_{\mathrm{HS}} = \mathbb{E}_z\left[\left\langle g_{a,s}(z)g_{a,s}(z)^\top, h^o(z)h^o(z)^\top\right\rangle_{\mathrm{F}}\right] \tag{381}$$

$$= \mathbb{E}_z\left[\left(h^o(z)^\top g_{a,s}(z)\right)^2\right]. \tag{382}$$

Now write $g_{a,s}(z) = J_a(z)s$. Then

$$\mathbb{E}_z\left[\left(h^o(z)^\top g_{a,s}(z)\right)^2\right] = \mathbb{E}_z\left[s^\top J_a(z)^\top h^o(z)h^o(z)^\top J_a(z)s\right] \tag{383}$$

$$= s^\top\mathbb{E}_z\left[J_a(z)^\top h^o(z)h^o(z)^\top J_a(z)\right]s \tag{384}$$

$$= s^\top \mathsf{B}_a^o s. \tag{385}$$

Taking $s = s_\star$ gives

$$\sum_{p\ge 0}\left\langle \mathcal{Q}_{p,+}(\mathcal{U}_a(s_\star)), \mathcal{M}_p^o\right\rangle_{\mathrm{HS}} = s_\star^\top \mathsf{B}_a^o s_\star = \gamma_{\mathrm{esc}}.$$

Hence

$$F_{\mathrm{fit}}(\mu_h) = F_{\mathrm{fit}}(\mu^\dagger) - h^2\gamma_{\mathrm{esc}} + o(h^2).$$

Step 5: compute the regularizer contribution.

Recall that $s_\star$ is the chosen unit transverse direction, $s_\star^\top a_\star = 0$, and $u_\pm = s_\star/2$. At the assigned pole $y = \pm a_\star$, the normalized perturbation satisfies

$$R_h(y) = \frac{y + (h/2)s_\star}{\sqrt{1 + h^2/4}}, \qquad y^\top R_h(y) = \frac{1}{\sqrt{1 + h^2/4}}.$$

The angle lies in $[0, \pi/2)$. Because $\cos(\arctan x) = 1/\sqrt{1+x^2}$ for $x \geq 0$,

$$d_{\mathbb{S}}(R_h(y), y) = \arccos \frac{1}{\sqrt{1 + h^2/4}} = \arctan(|h|/2).$$

For sufficiently small $|h|$ both atoms retain their cap assignments. Each marginal has two atoms of mass $1/2$, and there are two marginals; therefore eq. (251) gives

$$Z_\pm(\mu_h) = \sum_{i=1}^{2} \sum_{\varepsilon \in \{-1,1\}} \frac{1}{2} \arctan^2(|h|/2) = 2\arctan^2(h/2) = O(h^2). \tag{386}$$

For $h \neq 0$, the regularizer has the explicit factorization

$$F_{\mathrm{reg}}(\mu_h) = \frac{\lambda_{\mathrm{reg}}}{p_W} |h|^{p_W} \left(\frac{2\arctan^2(h/2)}{h^2} \right)^{p_W/2}.$$

The factor in parentheses extends to a smooth positive function at zero, with value $1/2$, because $\arctan(h/2) = h/2 + O(h^3)$. The derivative calculation following eq. (369) then gives $F'_{\mathrm{reg}}(h) = O(|h|^{p_W - 1})$ and $F''_{\mathrm{reg}}(h) = O(|h|^{p_W - 2})$. Since $p_W = 2 + \delta_W > 2$, the regularizer is C^2 at zero, both derivatives there are zero, and $F_{\mathrm{reg}}(\mu_h) = o(h^2)$. Thus the ordinary second derivative of the full objective exists and equals $-2\gamma_{\mathrm{esc}}$, as well as satisfying the displayed expansion. Combining the fitting expansion and the regularizer estimate proves the claim. □

B.10 Ambient positive curvature near the same-density teacher

This subsection proves positive spherical product pushforward second variation at, and near, the exact-fit same-density teacher. It does not assert that an escaping gradient-flow trajectory reaches this neighborhood, or that the teacher is stationary for the regularized objective.

The descent direction in proposition B.7 creates a same-sign odd mode. This motivates the same-density branch

$$\mathcal{M}_+ \coloneqq \{\mu_1 \otimes \mu_2 \mid \mu_1 = \mu_2\}. \tag{387}$$

On $\mathcal{M}_+$ we have $A_n = 0$ for every $n \in \mathcal{O}$ and $S_n = T_{1,n} = T_{2,n}$. By proposition B.2, the reference zero-fit point of the data term lies on this branch.

B.10.1 Invariance of the same-density branch

We first check that the same-density branch is dynamically consistent with the product WGF.

Proposition B.8 (Invariance of $\mathcal{M}_+$ under the antipodal metric objective)**.** ***Assumptions used.*** *assumptions B.1 and B.2, together with the exchange symmetry of F in (μ_1, μ_2).*

If $\mu_0 = \mu_{1,0} \otimes \mu_{2,0} \in \mathcal{M}_+$ and μ_t is the unique corresponding product WGF solution on an interval $[0, T)$ in the fixed-assignment cap class $\mu_{i,t} \in \mathcal{P}_\pm(r_{\mathrm{cap}})$, $i = 1, 2$, then

$$\mu_t \in \mathcal{M}_+ \qquad \forall t \in [0, T). \tag{388}$$

Proof. **Step 1: exchange the two factors in the objective.** If $\mu_1 = \mu_2$, then $T_{1,n} = T_{2,n}$ for every $n \in \mathcal{O}$.

The fitting part is invariant under interchanging the two factors because $\mu_1^o = \mu_2^o$.

The metric regularizer is exchange-invariant because it depends on the two factors only through

$$Z_\pm(\mu_1, \mu_2) = W_2^2(\mu_1, \nu_\star^\pm) + W_2^2(\mu_2, \nu_\star^\pm).$$

Step 2: transform the characteristics and apply uniqueness. In particular, exchange symmetry gives

$$F(\mu_1, \mu_2) = F(\mu_2, \mu_1), \qquad g_1(\mu_1, \mu_2; w) = g_2(\mu_2, \mu_1; w),$$

where $g_i = \nabla_{\mathbb{S}}(\delta F/\delta \mu_i)$, as in eq. (326). For the swapped pair $\widetilde{\mu}_{1,t} = \mu_{2,t}$, $\widetilde{\mu}_{2,t} = \mu_{1,t}$, its first continuity equation is

$$\partial_t \widetilde{\mu}_{1,t} - \nabla_{\mathbb{S}} \cdot (g_1(\widetilde{\mu}_{1,t}, \widetilde{\mu}_{2,t}; \cdot)\widetilde{\mu}_{1,t}) = \partial_t \mu_{2,t} - \nabla_{\mathbb{S}} \cdot (g_2(\mu_{1,t}, \mu_{2,t}; \cdot)\mu_{2,t}) = 0.$$

The other equation follows by exchanging indices. Since $\mu_{1,0} = \mu_{2,0}$, the swapped pair has the same initial state. Uniqueness on the stated interval gives $\widetilde{\mu}_{i,t} = \mu_{i,t}$, hence $\mu_{1,t} = \mu_{2,t}$. □

B.10.2 Positive curvature at and near the same-density teacher

Let

$$\nu^o := \mu_1^o = \mu_2^o, \qquad \mu_+^o := \nu^o \otimes \nu^o \in \mathcal{M}_+. \tag{389}$$

At this state the student output equals the teacher output, so the fitting residual vanishes. The next proposition combines the resulting nonnegative fitting curvature with the strict metric curvature and shows that the margin persists locally.

Proposition B.9 (Ambient positive curvature near a same-density teacher)**.** ***Assumptions used.*** *assumptions B.1 and B.2. Suppose*

$$\nu^o \in \mathcal{P}_\pm(r_{\mathrm{cap}}), \qquad Z_o := Z_\pm(\mu_+^o) = 2W_{2,\mathbb{S}}^2(\nu^o, \nu_\star^\pm) > 0.$$

Then there exist $\rho_o > 0$ and $c_o > 0$, depending only on the fixed model parameters and the teacher, such that the following holds. For every $\mu = \mu_1 \otimes \mu_2 \in \mathcal{X}_\pm(r_{\mathrm{cap}})$ satisfying

$$W_{2,\mathbb{S}}^2(\mu_1, \nu^o) + W_{2,\mathbb{S}}^2(\mu_2, \nu^o) < \rho_o^2,$$

and every smooth product tangent field $v = (v_1, v_2)$ whose perturbation paths preserve the fixed-assignment caps for small parameter values,

$$D_{W_2}^2 F(\mu)[v, v] \geq c_o \|v\|_{L^2(\mu; \mathbb{R}^{d_w})}^2. \tag{390}$$

No density assumption on ν^o is required.

Proof. **Step 1: positive curvature at the exact-fit teacher.** For smooth tangent fields v_i on the two factors, let

$$R_{i,h}(w) := \frac{w + h v_i(w)}{\|w + h v_i(w)\|_{\mathbb{R}^{d_w}}}, \qquad \nu_{i,h} := (R_{i,h})_\# \nu^o, \qquad \mu_h := \nu_{1,h} \otimes \nu_{2,h}.$$

The normalized retraction has the same second jet as the spherical EXP path by eq. (355), so its second derivative computes $D_{W_2}^2 F(\mu_+^o)[v, v]$. Set

$$Q(h, z) := h_{\nu_{1,h}}(z) h_{\nu_{2,h}}(z)^\top, \qquad B(z) := h_{\nu^o}(z) h_{\nu^o}(z)^\top.$$

Exact fitting means $Q(0, z) = B(z)$. Therefore, differentiating the squared residual twice gives

$$\begin{aligned} D_{W_2}^2 F_{\mathrm{fit}}(\mu_+^o)[v, v] &= \frac{\mathrm{d}^2}{\mathrm{d}h^2}\Big|_{h=0} \frac{1}{2} \|Q(h) - B\|_{L_z^2(\mathrm{F})}^2 \\ &= \|\dot{Q}(0)\|_{L_z^2(\mathrm{F})}^2 \\ &= \sum_{p \geq 0} \left\| \frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} \widetilde{\mathcal{M}}_p(\Theta(\mu_h)) \right\|_{\mathrm{HS}}^2 \geq 0. \end{aligned} \tag{391}$$

The last equality is the finite active Hermite expansion from lemma B.3 and eq. (310).

By lemma B.11, the metric term at the same state satisfies

$$D^2_{W_2} F_{\mathrm{reg}}(\mu^o_+)[v,v] \geq \lambda_{\mathrm{reg}} \kappa_{\mathrm{cap}} Z_o^{(p_W-2)/2} \|v\|^2_{L^2(\mu^o_+;\mathbb{R}^{d_w})}.$$

The coefficient is positive because $Z_o > 0$. Adding this inequality to eq. (391) proves strict ambient curvature at the teacher itself.

Step 2: a lower bound at a nearby cap-supported state. Fix $\mu = \mu_1 \otimes \mu_2 \in \mathcal{X}_{\pm}(r_{\mathrm{cap}})$, and use the same normalized retractions based at each μ_i. Write

$$h_i(h,z) := \int \sigma(R_{i,h}(w)^\top z)\, \mathrm{d}\mu_i(w), \qquad V_i := \|v_i\|_{L^2(\mu_i;\mathbb{R}^{d_w})}.$$

Dots below denote derivatives at $h = 0$. From eq. (355), $\dot{R}_{i,0} = v_i$ and $\ddot{R}_{i,0} = -\|v_i\|^2_{\mathbb{R}^{d_w}} w$, so differentiation under the integral gives

$$\begin{aligned} \dot{h}_i(z) &= \int \sigma'(w^\top z)(v_i(w)^\top z)\, \mathrm{d}\mu_i(w), \\ \ddot{h}_i(z) &= \int \left[\sigma''(w^\top z)(v_i(w)^\top z)^2 - \|v_i(w)\|^2_{\mathbb{R}^{d_w}} \sigma'(w^\top z)(w^\top z)\right] \mathrm{d}\mu_i(w). \end{aligned} \tag{392}$$

When $v_i(w) \neq 0$, the vectors w and $v_i(w)/\|v_i(w)\|_{\mathbb{R}^{d_w}}$ are orthogonal unit vectors because $v_i(w) \in T_w \mathbb{S}^{d_w-1}$. Hence their projections of $z \sim N(0, I_{d_w})$ are independent standard Gaussians. Substitution into the definitions of M_4, K_1, K_2 in eq. (328), followed by Minkowski and Cauchy–Schwarz, yields

$$\|h_i(0)\|_{L^4_z} \leq M_4, \qquad \|\dot{h}_i\|_{L^4_z} \leq K_1 V_i, \qquad \|\ddot{h}_i\|_{L^4_z} \leq K_2 V_i^2. \tag{393}$$

At $v_i(w) = 0$, the corresponding integrands in eq. (392) vanish, so the same bounds hold without division by zero. The finite polynomial activation and compact sphere justify the differentiations and Gaussian integration.

For $Q(h,z) := h_1(h,z) h_2(h,z)^\top$, the product rule gives

$$\dot{Q} = \dot{h}_1 h_2^\top + h_1 \dot{h}_2^\top, \qquad \ddot{Q} = \ddot{h}_1 h_2^\top + 2\dot{h}_1 \dot{h}_2^\top + h_1 \ddot{h}_2^\top.$$

Since $\|xy^\top\|_{\mathrm{F}} = \|x\|_2 \|y\|_2$, Hölder and eq. (393) give, term by term,

$$\begin{aligned} \|\ddot{h}_1 h_2^\top\|_{L^2_z(\mathrm{F})} &\leq M_4 K_2 V_1^2, \\ \|2\dot{h}_1 \dot{h}_2^\top\|_{L^2_z(\mathrm{F})} &\leq 2K_1^2 V_1 V_2, \\ \|h_1 \ddot{h}_2^\top\|_{L^2_z(\mathrm{F})} &\leq M_4 K_2 V_2^2. \end{aligned}$$

Using $2V_1V_2 \leq V_1^2 + V_2^2$, we obtain

$$\|\ddot{Q}\|_{L^2_z(\mathrm{F})} \leq (M_4 K_2 + K_1^2) \|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}. \tag{394}$$

Recall $B(z) = h_{\nu^o}(z) h_{\nu^o}(z)^\top$, and set

$$E(\mu) := \|h_{\mu_1} h_{\mu_2}^\top - B\|_{L^2_z(\mathrm{F})}, \qquad C := M_4 K_2 + K_1^2.$$

The constant C is proof-local and depends only on the fixed activation. Differentiating $F_{\mathrm{fit}}(\mu_h) = \frac{1}{2}\|Q(h) - B\|^2_{L^2_z(\mathrm{F})}$, dropping the nonnegative $\|\dot{Q}\|^2_{L^2_z(\mathrm{F})}$ term, and using eq. (394) gives

$$\begin{aligned} D^2_{W_2} F_{\mathrm{fit}}(\mu)[v,v] &= \|\dot{Q}\|^2_{L^2_z(\mathrm{F})} + \langle Q(0) - B, \ddot{Q} \rangle_{L^2_z(\mathrm{F})} \\ &\geq -C E(\mu) \|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}. \end{aligned}$$

Adding the regularizer estimate in eq. (367) yields the key nearby bound

$$D^2_{W_2} F(\mu)[v,v] \geq \left[\lambda_{\mathrm{reg}} \kappa_{\mathrm{cap}} Z_{\pm}(\mu)^{(p_W-2)/2} - C E(\mu)\right] \|v\|^2_{L^2(\mu;\mathbb{R}^{d_w})}. \tag{395}$$

Step 3: keep the coefficient positive in a teacher neighborhood. Let

$$d := \left[W_{2,\mathbb{S}}^2(\mu_1, \nu^o) + W_{2,\mathbb{S}}^2(\mu_2, \nu^o)\right]^{1/2}.$$

For sphere points w, y, let $c : [0,1] \to \mathbb{S}^{d_w-1}$ be a shortest constant-speed arc from y to w. Then $\|c'(s)\|_{\mathbb{R}^{d_w}} = d_{\mathbb{S}}(w,y)$ and $c(s)^\top c'(s) = 0$. Thus, when the speed is nonzero, $c(s)$ and $c'(s)/\|c'(s)\|_{\mathbb{R}^{d_w}}$ are orthogonal unit vectors. The same Gaussian calculation used in eq. (393) gives

$$\|\sigma'(c(s)^\top z)(c'(s)^\top z)\|_{L_z^4} = K_1 \|c'(s)\|_{\mathbb{R}^{d_w}}.$$

The identity is immediate when the speed is zero. Integrating along the arc and applying Minkowski gives

$$\|\sigma(w^\top z) - \sigma(y^\top z)\|_{L_z^4} \le K_1 d_{\mathbb{S}}(w,y).$$

For an optimal coupling π_i of μ_i and ν^o, another use of Minkowski and Cauchy–Schwarz therefore yields

$$\begin{aligned}\|h_{\mu_i} - h_{\nu^o}\|_{L_z^4} &\le K_1 \int d_{\mathbb{S}}(w,y)\, \mathrm{d}\pi_i(w,y) \\ &\le K_1 W_{2,\mathbb{S}}(\mu_i, \nu^o).\end{aligned}$$

Subtracting the two product outputs,

$$h_{\mu_1} h_{\mu_2}^\top - B = (h_{\mu_1} - h_{\nu^o}) h_{\mu_2}^\top + h_{\nu^o}(h_{\mu_2} - h_{\nu^o})^\top.$$

Since $\|h_{\mu_i}\|_{L_z^4}, \|h_{\nu^o}\|_{L_z^4} \le M_4$ by eq. (330), Hölder and the preceding feature bound give $E(\mu) \le \sqrt{2} M_4 K_1 d$. Moreover,

$$d = W_{2,\mathcal{M}}(\mu, \mu_+^o), \qquad \sqrt{Z_\pm(\mu)} = W_{2,\mathcal{M}}(\mu, \mu^\dagger), \qquad \sqrt{Z_o} = W_{2,\mathcal{M}}(\mu_+^o, \mu^\dagger),$$

by the product metric in eq. (231) and the definition of $Z_\pm$ in eq. (251), together with the definition of Z_o in the proposition statement. Therefore the reverse triangle inequality gives $\left|\sqrt{Z_\pm(\mu)} - \sqrt{Z_o}\right| \le d$. Combining the two estimates yields

$$E(\mu) \le \sqrt{2} M_4 K_1 d, \qquad \left|\sqrt{Z_\pm(\mu)} - \sqrt{Z_o}\right| \le d. \tag{396}$$

As $d \to 0$, eq. (396) shows that $E(\mu) \to 0$ and $Z_\pm(\mu) \to Z_o$. Hence the bracket in eq. (395) converges to the positive number $\lambda_{\mathrm{reg}} \kappa_{\mathrm{cap}} Z_o^{(p_W-2)/2}$. Choose $\rho_o > 0$ so that, whenever $d < \rho_o$, the regularizer coefficient is at least three quarters of this limit and $CE(\mu)$ is at most one quarter of it. Then eq. (395) gives eq. (390) with

$$c_o := \frac{1}{2} \lambda_{\mathrm{reg}} \kappa_{\mathrm{cap}} Z_o^{(p_W-2)/2} > 0.$$

The neighborhood is relative to $\mathcal{X}_\pm(r_{\mathrm{cap}})$: small Wasserstein distance alone does not force the entire support to remain in the fixed-assignment caps. None of the estimates uses a teacher density. □

If $Z_o = 0$, the proposition does not apply. In that case the metric regularizer is Hessian-flat at the antipodal Dirac target by eq. (298).

C NUMERICAL SANITY CHECKS

We examine the non-spikiness condition in Assumption 5.2(iii) and the behavior of adaptive PWGF with each system making its own perturbation decisions. Following Yamamoto et al. (2025, Algorithm 1), the trigger uses the L^2 norm of the gradient under that system's current distribution. We use a smooth matrix-factorization model, separate from the spherical model of Section 4.

C.1 MODEL AND POPULATION–PARTICLE COMPARISON

Objective and initial law. A particle is a matrix $W \in \mathbb{R}^{32\times 2}$, identified with $\mathbb{R}^{64}$ by row-major vectorization; matrix norms below are Frobenius norms. With entrywise $\tanh$, set

$$F(\nu) = f(A_\nu), \qquad f(A) = \frac{1}{4}\|AA^\top - M_*\|_F^2, \qquad A_\nu = \int \tanh(W)\,\nu(dW). \tag{397}$$

Here $M_* = u_1u_1^\top + \frac{1}{2}u_2u_2^\top$, where $(u_1)_a = 1/\sqrt{32}$ and $(u_2)_a = (-1)^{a-1}/\sqrt{32}$ for $1 \le a \le 32$. Thus the objective factorizes the mean transformed particle. Generate 2^{14} matrices Z_a with independent standard normal entries once (seed 20260925), and fix the realized population law

$$P_a(0-) = Z_a, \quad P_{a+2^{14}}(0-) = -Z_a, \qquad \mu_{0-} = 2^{-15}\sum_{a=1}^{2^{14}}(\delta_{Z_a} + \delta_{-Z_a}).$$

The notation $0-$ denotes initialization before a possible jump at $t = 0$. For each of three trials, sample labels I_i independently and uniformly from $\{1, \dots, 2^{15}\}$, independently of the population innovations, and set $X^i_{0-} = P_{I_i}(0-)$ before that jump. We use $N \in \{2048, 4096, 8192, 16384, 32768\}$, with nested label prefixes within each trial. Sampling is with replacement, including at $N = 32768$.

Continuous dynamics. The specified population has 32768 atoms, all of which are evolved; its integrals are full atom averages. The empirical system uses only its sampled particles:

$$\mu_t = 2^{-15}\sum_{a=1}^{2^{15}}\delta_{P_a(t)}, \qquad \widehat{\mu}^N_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^i_t}, \qquad Y^i_t = P_{I_i}(t).$$

The ODEs and their shared velocity formula are

$$\begin{aligned} \dot{P}_a(t) &= -\nabla_W F(\mu_t, P_a(t)), \qquad \dot{X}^i_t = -\nabla_W F(\widehat{\mu}^N_t, X^i_t), \\ -\nabla_W F(\nu, W) &= -(1 - \tanh^2 W) \odot [(A_\nu A_\nu^\top - M_*)A_\nu], \end{aligned} \tag{398}$$

where $\odot$ is entrywise multiplication. In particular, for $F_N(\mathbf{X}) = F(\widehat{\mu}^N)$, the particle velocity is $-N\nabla_{X^i}F_N$. We integrate both systems by RK4; the population is a fixed discrete measure, not a Monte Carlo approximation to a continuous law.

Gaussian jumps. For a matrix W, write $J_W = \mathrm{diag}(\mathrm{vec}(1 - \tanh^2 W))$ and $S_\nu = \int J_U^2\,\nu(dU)$. Since $\nabla f(A) = (AA^\top - M_*)A$, differentiation in a matrix direction $B \in \mathbb{R}^{32\times 2}$ gives $\mathrm{D}(\nabla f)(A)[B] = (BA^\top + AB^\top)A + (AA^\top - M_*)B$. Thus the symmetric matrix $L_A \in \mathbb{R}^{64\times 64}$ representing the Hessian of f satisfies

$$L_A\,\mathrm{vec}(B) = \mathrm{vec}\big((AA^\top - M_*)B + B(A^\top A) + AB^\top A\big).$$

Thus the integral Hessian and its Gaussian covariance from Section 2.1 are

$$\nabla^2_W F(\nu; W, V) = J_W L_{A_\nu} J_V, \qquad K_\nu(W, V) = J_W L_{A_\nu} S_\nu L_{A_\nu}^\top J_V. \tag{399}$$

Since $\mathrm{vec}(\mathrm{D}\tanh(V)[B]) = J_V\,\mathrm{vec}(B)$, the chain rule in eq. (398) yields the first identity. The kernel composition in Section 2.1 gives $K_\nu(W, V) = \int (J_W L_{A_\nu} J_U)(J_U L_{A_\nu} J_V)\,\nu(dU) = J_W L_{A_\nu} S_\nu L_{A_\nu} J_V$, where $S_\nu = \int J_U^2\,\nu(dU)$ and $L_{A_\nu} = L_{A_\nu}^\top$. A draw of the field is $\mathrm{vec}(\xi_\nu(W)) = J_W L_{A_\nu} S_\nu^{1/2} z$, using one $z \sim \mathcal{N}(0, I_{64})$ for all W; its covariance is $K_\nu(W, V)$ because $S_\nu^{1/2}\mathbb{E}[zz^\top]S_\nu^{1/2} = S_\nu$. The Gaussian rank is at most 64 even without finite support. At their respective trigger times, the population and particle jumps have marginal laws

$$\begin{aligned} P_a^+ &= P_a^- + \eta\xi_{\mu^-}(P_a^-), \\ \mathrm{vec}(X^{i,+} - X^{i,-}) &= \frac{\eta}{\sqrt{N}}\sum_{j=1}^{N}\nabla^2_W F(\widehat{\mu}^{N,-}; X^{i,-}, X^{j,-})g_j, \qquad g_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{64}). \end{aligned} \tag{400}$$

The implementation uses equivalent 64-dimensional factors. The population path is generated first. Its j-th innovation is coupled, by covariance-factor alignment with independent Gaussian completion, to the empirical system's j-th attempt. An unmatched empirical attempt uses a fresh independent field with its empirical covariance. This coupling uses attempt index, not common physical time; the reference particles do not enter the controller.

C.2 Adaptive controller and recorded diagnostics

Each system follows the gradient-trigger, wait, and decrease-test pattern of Yamamoto et al. (2025, Algorithm 1). The controller acts separately for the population and each particle system:

1. **Inspect and perturb.** At $t = 0, 0.1, \ldots, 99.9$, if the system has not stopped and no waiting window is active, test its pre-jump state using

$$\text{population:} \quad \left(2^{-15} \sum_{a=1}^{2^{15}} \|\nabla_W F(\mu_{t-}, P_a(t-))\|_F^2 \right)^{1/2} \leq 0.01,$$

$$\text{particles:} \quad \left(\frac{1}{N} \sum_{i=1}^{N} \|\nabla_W F(\widehat{\mu}_{t-}^N, X_{t-}^i)\|_F^2 \right)^{1/2} \leq 0.01.$$

   These are the respective L^2-gradient norms, evaluated from eq. (398); each system uses its own measure in both the gradient and the average. When its inequality holds, save the pre-jump measure and apply the corresponding jump in eq. (400) with $\eta = 0.2$. Otherwise, continue the flow to the next inspection time.

2. **Wait and test the decrease.** After a jump at s, evolve for 10 time units without another jump, then compare the objective with the saved pre-jump value:

$$\begin{cases} \text{resume inspection at the next grid tick,} & F(\nu_{s-}) - F(\nu_{s+10}) > 0.001, \\ \text{stop and return } \nu_{s-}, & F(\nu_{s-}) - F(\nu_{s+10}) \leq 0.001. \end{cases}$$

   Here ν is the measure of the system being tested. A passed decrease test does not trigger another jump at $s + 10$; inspection resumes at $s + 10.1$ if time remains before the cap of 100. The returned candidate ν_{s-} is distinct from the executed state ν_{s+10}.

Parameter	Value	Parameter	Value
Jump amplitude η	0.2	Gradient threshold	0.01
Waiting time	10	Decrease threshold	0.001
RK4 step	0.05	Inspection interval	0.1
Time cap	100	Trials	3

Non-spikiness. At common physical times while both systems remain active, we compute the paired error from Section 2.4 and its cubic ratio:

$$r_t = \left(\frac{1}{N} \sum_i \|X_t^i - Y_t^i\|_F^2 \right)^{1/2}, \qquad s_t = \frac{N^{-1} \sum_i \|X_t^i - Y_t^i\|_F^3}{r_t^3}. \tag{401}$$

We record inspection times and both sides of each jump, and omit s_t when $r_t \leq 10^{-12} \max\{1, (N^{-1} \sum_i \|Y_t^i\|_F^2)^{1/2}\}$. Across all 15 conditions, the largest recorded ratio is 1.017416. Figure 4 therefore supports non-spiky paired errors in these runs: the empirical analogue of Assumption 5.2(iii) is close to one at recorded times. This is not a bound over all continuous times or random paths.

Adaptive schedules. All 15 particle systems and three population paths stopped before the cap; the empirical returned objectives ranged from 1.12×10^{-4} to 1.41×10^{-4}. Figure 5 illustrates autonomous decisions based on each system's own L^2-gradient test: all $N = 2048, 4096$ systems skip the population's initial jump, and trial 2 at $N = 16384$ makes a third attempt without a population counterpart. The resulting differences in physical jump times can produce temporary same-time

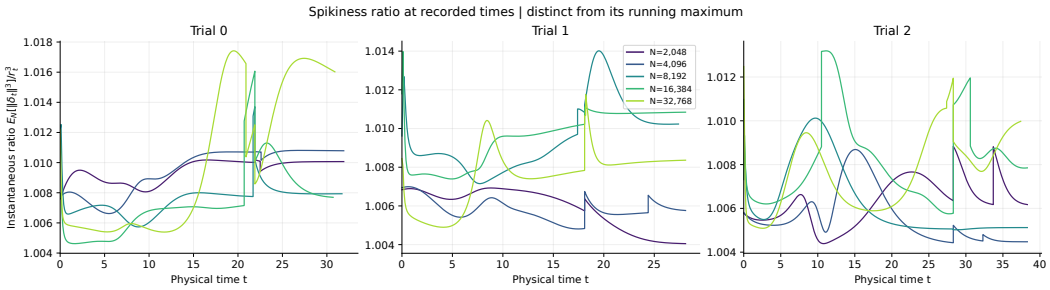


Figure 4: The instantaneous cubic ratio s_t in eq. (401), with all three trials and five particle counts. Curves end when either system stops; near-zero denominators are omitted. The ratio measures the distribution of paired errors, not particle positions.

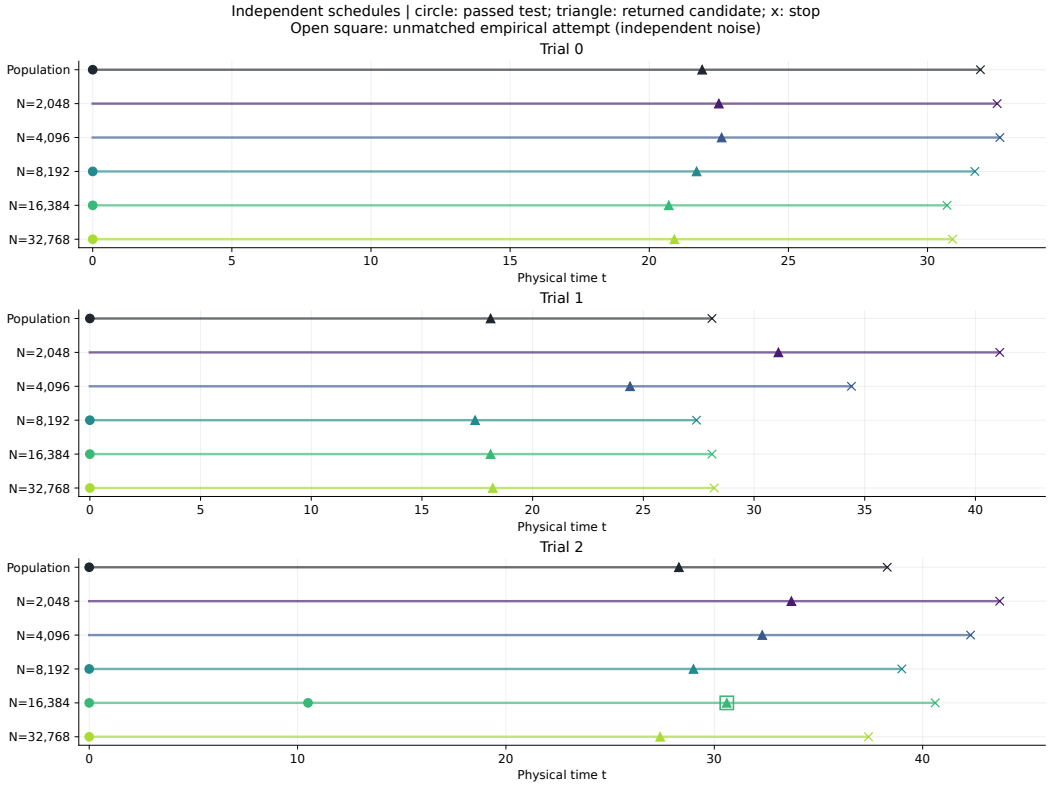


Figure 5: Independently chosen adaptive schedules, showing every system. Circles mark jumps whose decrease test passes; triangles mark the saved pre-jump candidates returned after a failed test; crosses mark stopping times. The open square marks an unmatched empirical attempt. Coupling by attempt index does not force jump times to coincide.

discrepancies: at $N = 32768$, recorded peaks of r_t range from 0.584 to 0.601. Discrepancies can be much larger at smaller N and are not monotone in N, despite the near-unit spikiness ratios. These diagnostics complement Theorem 3.1; its guarantee remains for a prescribed common schedule.

Numerical checks and reproduction. Halving the RK4 step preserved all event times and counts; halving the inspection interval preserved counts but shifted some event times by 0.05. The latter changed individual returned objectives by up to 4.89×10^{-6}, so small objective gaps are not used to infer a rate. Code, configurations, exact seeds, and archived scalar records are provided in the accompanying anonymous supplementary package.

C.3 PAIRED TRAJECTORIES AND OBJECTIVE VALUES

Figure 6 shows the paired error r_t from eq. (401) and the objective gap $|F(\widehat{\mu}_t^N) - F(\mu_t)|$ at the same physical time. Figure 7 shows the objective values themselves, using eq. (397), until each system stops. Together they show particle discrepancies and optimization progress under separately triggered jumps.

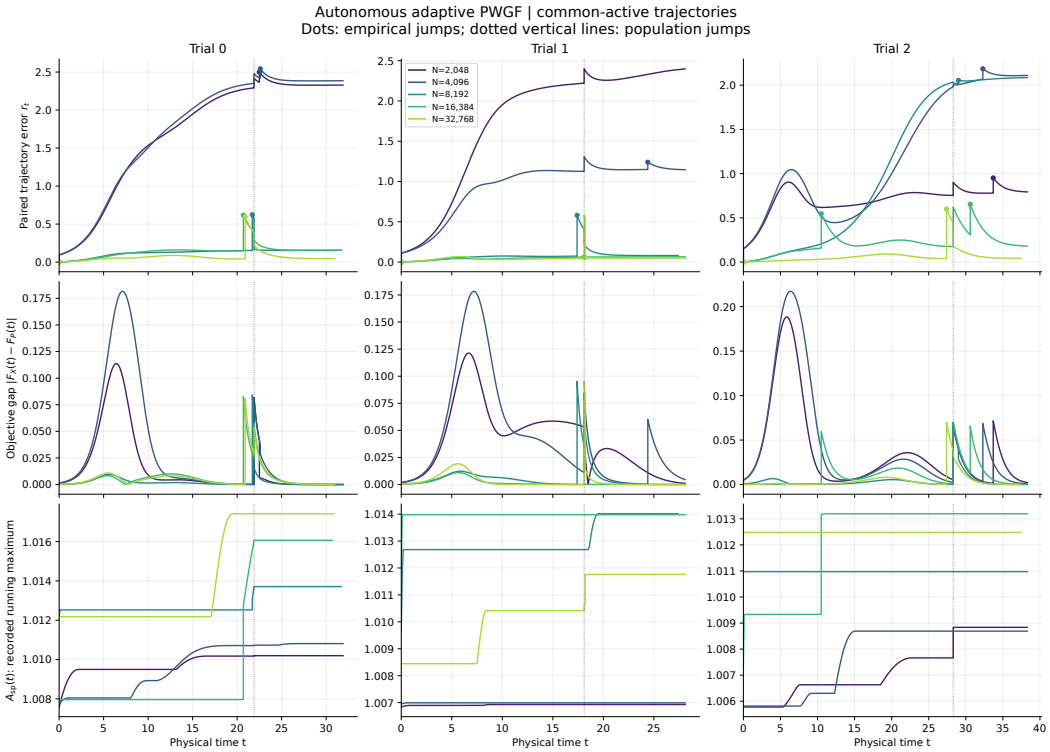


Figure 6: Common-time diagnostics for all 15 conditions, with one column per trial. Top: paired RMS error r_t. Middle: $|F_X(t) - F_P(t)|$, where $F_X(t) = F(\widehat{\mu}_t^N)$ and $F_P(t) = F(\mu_t)$. Bottom: the running maximum of the recorded ratio s_t, labeled $A_{\mathrm{sp}}(t)$, rather than its instantaneous value in fig. 4. Dots in the top row mark empirical jumps; dotted vertical lines mark population jumps. Curves end when either system stops; vertical segments connect recorded pre/post-jump values.

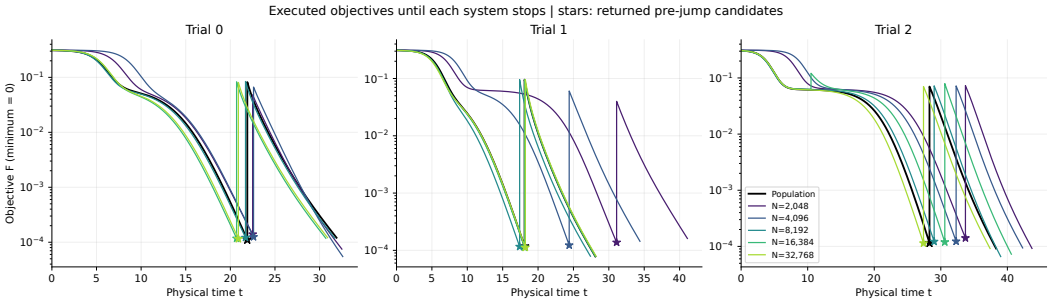


Figure 7: Objective values $F(\mu_t)$ (black) and $F(\widehat{\mu}_t^N)$ (colored) for every trial and particle count. Each curve continues to that system's own stopping time. Stars mark the saved pre-jump candidates at their candidate times, rather than the final executed states. The minimum of eq. (397) is zero.