STATISTICAL BIAS AND EFFECTIVE DYNAMICS OF JKO SCHEMES WITH ESTIMATED GROUND GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Can smaller time steps remove errors caused by estimating a transport geometry from finite data? In smooth finite-dimensional local models, optimal-plan response contributes to a bias that can survive exact solves and vanishing steps. A candidate-dependent frozen-plan comparison isolates its endpoint coefficient; the local transport tensor determines its velocity limit. A two-support example proves nonvanishing at fixed entropic regularization. Under independent refreshment, finite-horizon averaging separates drift, fluctuations, and numerical error. In a five-support test, the independently computed coefficient predicts endpoint contrasts within 1.12% and probability-mass contrasts within 1.97% at the smallest perturbation. Plan response partially cancels the frozen response. Split-sample correction reduces leading coordinate-mean bias; on two frozen-feature tasks it improves finite-bank reference tracking by 7.3–24.7% while worsening prediction. Flow selection, accurate solution, and predictive quality are distinct objectives.

1 INTRODUCTION

A variational update can be solved accurately and still approximate the wrong target flow. Consider estimating a Mahalanobis ground metric from a batch covariance before each Jordan–Kinderlehrer–Otto (JKO) step. This changes which movements of mass are inexpensive. As the step size decreases, fresh estimation noise can average out while a systematic drift remains. *How much of that drift comes from the optimal transport plan adapting to the estimated geometry?*

The question connects JKO-based distribution learning (Mokrov et al., 2021; Xu et al., 2023) with data-dependent transport costs (Jawanpuria et al., 2025). Random inverse preconditioning already causes bias. To identify the additional plan response, we compare the full update with a control that freezes each candidate's coupling only with respect to geometry. At the reference geometry, the objectives and their first geometry derivatives coincide. Their second-order difference isolates the contribution of plan adaptation.

We establish three results in smooth finite-dimensional local models:

- **Endpoint separation.** A full–frozen identity cancels shared estimator bias and first-order response, leaving an explicit plan-motion coefficient (Theorem 2).
- **Persistent velocity.** A uniform small-step expansion identifies the coefficient per unit time. An entropic two-support example proves it can be nonzero at fixed regularization (Theorem 3; Proposition 4).
- **Magnitude and dynamics.** Independent coefficient tests support both predictions in one five-support geometry and reveal substantial bias cancellation. Under independent refreshment, finite-horizon averaging separates drift, fluctuations, discretization, and solver error (Theorem 5).

Gaussian feature tasks address a separate consequence: correcting mean bias improves tracking of a finite-bank reference but worsens prediction. A target flow and a predictive objective must therefore be evaluated separately.

Relation to existing work. JKO specifies dynamics through an energy and transport geometry (Jordan et al., 1998; Ambrosio et al., 2008). Entropic OT and Sinkhorn divergences yield smooth

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.

penalties (Cuturi, 2013; Peyré & Cuturi, 2019; Feydy et al., 2019), with established software (Flamary et al., 2021). Statistical OT studies empirical marginals (Genevay et al., 2019; Fatras et al., 2020; Goldfeld et al., 2024a;b); statistical JKO studies estimated model parameters (Wu & Wang, 2025); inexact JKO studies optimization error (Di Marino et al., 2025). Here the estimated input is the ground geometry.

Our tools are classical: the delta method, implicit optimization differentiation, and Sinkhorn sensitivity (Luise et al., 2018; Eisenberger et al., 2022; Blondel et al., 2022). Cost-curvature analysis concerns the transport value and inverse-cost identification (Peyré et al., 2026); we propagate the full–frozen curvature difference through the outer minimizer and identify its velocity limit. Stochastic preconditioning already has inversion and gradient–preconditioner biases (Nayak et al., 2026). Our contribution is a transport-specific decomposition within that broader phenomenon, not a new generic bias-correction principle. Appendix B.2 makes the nested-optimization connection explicit. Unlike intrinsic stochastic Wasserstein flows (Ren et al., 2025), randomness here enters through estimated geometry.

2 The Estimand and the Variational Model

Let $\Theta \subset \mathbb{R}^p$ parameterize an injective smooth family $\theta \mapsto \mu_\theta$. Let $g \in \Gamma \subset \mathbb{R}^{d_g}$ specify the ground geometry. For a smooth energy F_t and a specified transport penalty D_g, write

$$\mathsf{R}_{\tau,t}(\theta, g) = \arg\min_{u \in \Theta_{\mathrm{loc}}} \Psi_{\tau,t}(u; \theta, g), \qquad \Psi_{\tau,t} = F_t(u) + \frac{D_g(u, \theta)}{2\tau}. \tag{1}$$

Here $\Theta_{\mathrm{loc}} \subset \Theta$ is a fixed neighborhood in which the selected interior minimizer is unique for the parameters considered. No global uniqueness is assumed. For $D_g = W_g^2(\mu_u, \mu_\theta)$, this is JKO restricted to the model. A divergence penalty defines a related proximal scheme with its own limiting local metric. The finite entropic results below concern that metric at fixed regularization.

Population and reference geometry. The population estimand is a functional $g_t^\star = \mathcal{G}(P_t)$ of a feature distribution P_t. It need not minimize prediction error. Writing $C(P) = \mathrm{Cov}_{X\sim P}(X)$ for features in $\mathbb{R}^d$, one example is

$$G(P) = \frac{d\,[C(P) + \rho I]^{-1}}{\mathrm{tr}([C(P) + \rho I]^{-1})}, \qquad \rho > 0, \tag{2}$$

with $\widehat{G}_B$ obtained by inserting the declared sample-covariance estimator. The divisor is $B - 1$ in the feature experiments; literal substitution of the empirical measure uses B. This changes the leading estimator bias (Appendix A.2). Trace normalization fixes scale, and the ridge makes inversion smooth. Neither rule generally gives an unbiased estimate of $G(P)$.

On finite datasets, a disjoint large reference bank supplies g_t^{ref}. Its trajectory is an empirical approximation to the target flow, conditional on that bank and the chosen model. We reserve $g_t^\star$ for the population functional. A constant Mahalanobis cost becomes Euclidean under a linear change of variables, so bias in a Gaussian location example can arise entirely from inverse preconditioning. The full–frozen comparison below distinguishes this source from coupling response.

Coordinates and observables. An endpoint mean depends on its coordinates. We therefore also study scalar observables φ, which describe a fixed measured quantity under changes of chart. Energy noise can be included as another estimated argument; independence removes mixed covariances but does not center a nonlinear update.

Entropic finite transport. For fixed u, θ, abbreviate the strictly positive probability vectors by $a = a(u)$ and $b = b(\theta)$. Define

$$T_{\varepsilon,g}(a, b) = \min_{\pi \in \Pi(a,b)} \langle C(g), \pi \rangle + \varepsilon\, \mathrm{KL}(\pi \| a \otimes b), \tag{3}$$

$$S_{\varepsilon,g}(a, b) = T_{\varepsilon,g}(a, b) - \tfrac{1}{2} T_{\varepsilon,g}(a, a) - \tfrac{1}{2} T_{\varepsilon,g}(b, b). \tag{4}$$

Finite entropic examples use $D_g = S_{\varepsilon,g}$; Gaussian examples use their Gaussian Wasserstein penalty. Raw $T_{\varepsilon,g}$ can move a zero-energy endpoint, violating the near-identity condition below. Self-subtraction removes this issue, but fixed-ε Sinkhorn geometry need not equal classical W_2 geometry. We assert no $\varepsilon \to 0$ limit (Feydy et al., 2019; Carlier et al., 2017).

3 Statistical Response of a JKO Endpoint

Fix (τ, t, θ) and set $u^\star = \mathsf{R}_\tau(\theta, g^\star)$. The following conditions separate differentiability of the variational problem from concentration of its statistical input.

Assumption 1 (Stable response branch and estimator moments)**.** The selected minimizer in (1) is a unique interior branch on an open convex neighborhood U of $g^\star$. Its stationarity map $q = \nabla_u \Psi_\tau$ is C^3, $q_u \succeq \kappa I$ along the branch for some $\kappa > 0$, and the needed derivatives are bounded. For $\Delta_B = \widehat{g}_B - g^\star$, with $\widehat{g}_B \in U$ almost surely, let β and Σ denote its leading bias and second-moment coefficients:

$$\mathbb{E}\Delta_B = B^{-1}\beta + o(B^{-1}), \qquad \mathbb{E}[\Delta_B \Delta_B^\top] = B^{-1}\Sigma + o(B^{-1}), \tag{5}$$
$$\mathbb{E}\|\Delta_B\|^3 = o(B^{-1}), \qquad \mathbb{E}\|\Delta_B\|^4 = O(B^{-2}).$$

Uniform statements require these conditions uniformly over the state set considered. A localized alternative with explicit tail control is given in Appendix A.

The energy may be nonconvex away from this branch. A ridge makes covariance inversion smooth but does not imply outer stability. Appendix A.2 verifies estimator moments for Gaussian sampling; Appendix F lists the remaining conditions. Throughout, H is the outer Hessian, A_g the local transport tensor defined in Section 4.1, and $\mathcal{V}_\tau$ the endpoint covariance coefficient.

Let $H = q_u(u^\star; \theta, g^\star)$ and let $U_h = D_g \mathsf{R}_\tau[h]$ be the endpoint change per unit perturbation in direction h. Implicit differentiation gives

$$U_h = -H^{-1} q_g[h], \tag{6}$$
$$D_g^2 \mathsf{R}_\tau[h,k] = -H^{-1}\{q_{gg}[h,k] + q_{ug}[U_h, k] + q_{ug}[U_k, h] + q_{uu}[U_h, U_k]\}. \tag{7}$$

All derivatives are evaluated at $(u^\star, \theta, g^\star)$; $h, k \in \mathbb{R}^{d_g}$ and $U_h, U_k \in \mathbb{R}^p$. The two mixed terms make (7) symmetric in h, k.

Theorem 1 (Mean, covariance, and observable response)**.** *Under Assumption 1, put $J = D_g \mathsf{R}_\tau$ and $\mathcal{V}_\tau = J\Sigma J^\top$. Then*

$$\mathbb{E}[\mathsf{R}_\tau(\theta, \widehat{g}_B) - u^\star] = B^{-1} b_\tau + o(B^{-1}), \qquad \mathrm{Cov}(\mathsf{R}_\tau(\theta, \widehat{g}_B)) = B^{-1}\mathcal{V}_\tau + o(B^{-1}), \tag{8}$$

where $b_\tau = J\beta + \frac{1}{2}\sum_{i,j} \Sigma_{ij} D_g^2 \mathsf{R}_\tau[e_i, e_j]$. For a C^3 scalar observable with bounded derivatives on the branch,

$$\mathbb{E}\varphi(\mathsf{R}_\tau(\theta, \widehat{g}_B)) - \varphi(u^\star) = B^{-1}\left[D\varphi(u^\star) b_\tau + \tfrac{1}{2}\operatorname{tr}(\mathcal{V}_\tau D^2 \varphi(u^\star))\right] + o(B^{-1}). \tag{9}$$

The estimator bias $J\beta$ and the pure response term must both be retained. For centered perturbations $g^\star + \sigma\xi$ of fixed covariance, the corresponding leading mean effect is quadratic in σ. This is the prediction tested by E1. The statistical argument is the delta method (van der Vaart, 1998); the variational content is the derivative of the outer minimizer in (7).

The coefficient of a fixed scalar observable is unchanged by smooth reparameterization (Proposition 7, Appendix A.3). Individual coordinate-mean biases need not be invariant. Joint energy and geometry estimation uses their joint parameter vector; independent cross-fitting removes mixed covariance terms but leaves pure geometry response.

4 An Exact Separation of Coupling Motion

The full-minus-frozen experiment needs an objective-level definition. Freezing a single plan while changing its marginals would make the comparison infeasible. Instead, for *each candidate pair* (a, b), solve (3) at $g^\star$, obtaining $\pi_\star(a, b)$, and define

$$T^{\mathrm{fr}}_{\varepsilon, g}(a, b) = \langle C(g), \pi_\star(a, b)\rangle + \varepsilon \mathrm{KL}(\pi_\star(a, b) \| a \otimes b). \tag{10}$$

Form S^{fr} with all three terms in (4) frozen this way. Plans may move with a, b but not with g. Let $\mathsf{R}^{\mathrm{fr}}_\tau$ minimize the resulting outer objective. Let b^{fr}_τ be its mean-response coefficient from Theorem 1. This is an analysis control anchored at $g^\star$, not an estimator of the unknown geometry. It need not be a globally nonnegative divergence; only the stated stable local branches are used.

Theorem 2 (Plan-motion response identity)**.** *For smooth $C(g)$, select full and frozen branches satisfying Assumption 1 through the same interior minimizer at $g^\star$. They have the same outer Hessian H and endpoint derivative $J = D_g \mathsf{R}_\tau$. Writing $\operatorname{tr}_\Sigma D_g^2 S = \sum_{i,j} \Sigma_{ij} \partial_{g_i g_j} S$, their leading mean-response coefficients satisfy*

$$b_\tau^{\mathrm{act}} := b_\tau - b_\tau^{\mathrm{fr}} = -\frac{1}{4\tau} H^{-1} \nabla_u \operatorname{tr}_\Sigma D_g^2 (S_{\varepsilon,g} - S_{\varepsilon,g}^{\mathrm{fr}})(a(u), b(\theta))\big|_{u=u^\star, g=g^\star}. \tag{11}$$

Consequently, $b_\tau^{\mathrm{act}} \neq 0$ exactly when the gradient on the right is nonzero. The first-order endpoint covariance is identical for the two schemes.

At $g^\star$, the objectives and their first ground derivatives agree as functions of u; their second ground derivatives generally differ. Consequently, all terms in (7) except q_{gg} cancel, as does $J\beta$. Equation (11) is an exact identity between leading response coefficients, not a finite-B equality of biases. Cross and self terms may still cancel, so plan motion alone need not give a nonzero endpoint contrast.

There is an explicit way to compute this response. Fix the marginals, vectorize the plan, and let N be a full-column-rank basis for zero-row-sum, zero-column-sum perturbations. For $c_h = \operatorname{vec}(D_g C[h])$ and $H_\pi = \varepsilon \operatorname{diag}(1/\pi_\star)$, differentiation of the constrained optimum gives

$$D_g \pi_\star[h] = -N(N^\top H_\pi N)^{-1} N^\top c_h, \qquad (D_g^2 T - D_g^2 T^{\mathrm{fr}})[h,k] = -c_h^\top N (N^\top H_\pi N)^{-1} N^\top c_k. \tag{12}$$

Apply this formula separately to all three terms of S, using linear solves rather than explicit inverses. For affine $C(g)$, the optimized raw transport value has a negative semidefinite cost Hessian; the self-subtracted combination need not share its sign. The response can vanish for geometric or symmetry reasons. For a Gaussian location family with fixed covariance, optimal transport is translation, giving a plan-motion negative control while leaving ordinary inverse-metric bias possible.

The shared first-order covariance also cancels the observable-Hessian terms: a scalar φ has leading contrast $D\varphi(u^\star) b_\tau^{\mathrm{act}}/B$ (Appendix B.5). For affine costs, $D_g^2 S^{\mathrm{fr}} = 0$. For nonlinear costs, explicit $D_g^2 C$ terms cancel between full and frozen, leaving (12). Our matrix-exponential, trace-normalized experiment uses this general difference formula.

4.1 When coupling response survives a vanishing step

An endpoint contrast accumulates over approximately $1/\tau$ steps per unit time. We therefore distinguish the endpoint coefficient b_τ^{act} from the velocity coefficient b_v^{act}. For either penalty, $A_g(x) = \frac{1}{2} \nabla_{uu}^2 D_g(u,x)|_{u=x}$, with the second endpoint held fixed; A_g^{fr} denotes its frozen counterpart.

Theorem 3 (Coupling contribution to the effective field)**.** *Use the full and frozen Sinkhorn penalties on a common marginal chart a, with smooth symmetric $C(g)$. On a compact interior state set and a neighborhood of $g^\star$, assume both penalties have bounded mixed derivatives through order six and uniformly positive diagonal tensors; let F be C^4 there. Select the local minimizing branches issuing from $u = x$ at $\tau = 0$. Set $A = A_{g^\star}(x)$ and define the symmetric $p \times p$ matrix*

$$Q_\Sigma(x) = \sum_{i,j=1}^{d_g} \Sigma_{ij}\, \partial_{g_i g_j} (A_g - A_g^{\mathrm{fr}})(x)\Big|_{g=g^\star}.$$

Then

$$A_{g^\star}^{\mathrm{fr}}(x) = A, \qquad D_g A_{g^\star}^{\mathrm{fr}}(x) = D_g A_{g^\star}(x), \tag{13}$$

$$b_\tau^{\mathrm{act}}(x) = \tau b_v^{\mathrm{act}}(x) + O(\tau^2), \qquad b_v^{\mathrm{act}}(x) = \tfrac{1}{2} A^{-1} Q_\Sigma A^{-1} \nabla F(x). \tag{14}$$

For a fixed noise coefficient Σ, the remainder is uniform on a smaller compact state set and ground neighborhood. In particular, $b_v^{\mathrm{act}} = 0$ exactly when $Q_\Sigma A^{-1} \nabla F = 0$. A quantitative nonzero lower bound is given in Appendix B.6.

To connect the two coefficients, define $\mathcal{L}_\Sigma(u,x) = \operatorname{tr}_\Sigma D_g^2 (D_g - D_g^{\mathrm{fr}})(u,x)|_{g^\star}$. Define $K = \nabla_{uu}^2 \mathcal{L}_\Sigma(u,x)|_{u=x}$. Commuting the stated mixed derivatives gives $Q_\Sigma(x) = K/2$, as used in S4. For affine costs the frozen second ground derivative vanishes. The mean contrast is $O(\tau/B)$ per step, hence can contribute $O(B^{-1})$ per unit time. The proof controls its remainder in $C^2(g)$ before differentiating.

Proposition 4 (A nonzero entropic example)**.** *Let $a(p) = (p, 1-p)$, $D_g(p,x) = S_{\varepsilon,g}(a(p), a(x))$, and let $C(g)$ have zero diagonal and off-diagonal entries $g > 0$. At $x = 1/2$, for scalar noise coefficient $\Sigma = s^2 > 0$,*

$$A_g(1/2) = \varepsilon \sinh(g/\varepsilon), \qquad b_v^{\mathrm{act}}(1/2) = \frac{s^2 F'(1/2)}{2\varepsilon^3 \sinh(g^\star/\varepsilon)}. \tag{15}$$

For $F'(1/2) \neq 0$ this is nonzero, retains its sign under sufficiently small perturbations within the smooth affine-cost class, and satisfies $|b_\tau^{\mathrm{act}}| \geq \tau |b_v^{\mathrm{act}}|/2$ for all sufficiently small τ.

The example proves that a coupling contribution can survive in the dynamics at fixed ε. A Gaussian location family gives the complementary zero control because its translation coupling is independent of the metric. Ordinary inverse-metric bias can occur in either model. Thus a nonzero total bias alone cannot distinguish the two mechanisms.

5 From Endpoint Response to Effective Dynamics

The one-step theorem fixes τ. A dynamical limit additionally needs uniform control and a specified sampling protocol. Here geometry is refreshed independently at each step. Reusing one estimate or an EMA gives different temporal dependence (Appendix C.4).

Assumption 2 (Near-identity updates and independent refresh)**.** On a fixed horizon $[0, T]$ and an invariant state region, the exact update satisfies

$$\mathsf{R}_{\tau,t}(x,g) = x + \tau v_t(x,g) + \tau^2 r_{\tau,t}(x,g), \qquad \sup_{t,x,g,\tau} \|r_{\tau,t}\| < \infty. \tag{16}$$

The fields are uniformly Lipschitz in x and regular in t, with uniform $(2+\delta)$ moments. Geometry draws are fresh at each step, independent of the past with a prescribed law $Q_{t,B}$. The averaged field is $\bar{v}_{t,B}(x) = \mathbb{E}_{Q_{t,B}} v_t(x,g)$ and is Lipschitz in time. For adapted numerical endpoint errors s_k, define

$$e_\tau = \left\| \sum_{k < T/\tau} \|s_k\| \right\|_{L^2}. \tag{17}$$

The initial state is deterministic; the exact and numerical paths stay in the stated region. Constants are uniform in B for any joint B, τ limit.

For a smooth penalty with $\nabla_u D_g(x,x) = 0$ and $A_g(x) \succ 0$, implicit expansion yields $v_t(x,g) = -A_g(x)^{-1} \nabla F_t(x)$. Stability and tensor positivity must be checked for that penalty; zero-energy identity alone is insufficient. At fixed ε, the Sinkhorn tensor need not equal the Wasserstein tensor.

Theorem 5 (Averaging on a fixed horizon)**.** *Under Assumption 2, set $t_k = k\tau$ and let X^τ linearly interpolate $X_{k+1} = \mathsf{R}_{\tau,t_k}(X_k, \widehat{g}_{k,B}) + s_k$, and let $\dot{\bar{x}}_B(t) = \bar{v}_{t,B}(\bar{x}_B(t))$, with the same initial state. If $V_B = \sup_{t,x} \mathbb{E}\|v_t(x,\widehat{g}_B) - \bar{v}_{t,B}(x)\|^2 < \infty$, then*

$$\mathbb{E} \sup_{t \leq T} \|X^\tau(t) - \bar{x}_B(t)\|^2 \leq C_T\{\tau V_B + \tau^2 + e_\tau^2\}. \tag{18}$$

If the moment and differentiability conditions of Theorem 1 apply uniformly to v_t, then

$$\bar{v}_{t,B} = v_t(\cdot, g_t^\star) + B^{-1} b_{v,t} + o(B^{-1}), \qquad b_{v,t} = D_g v_t[\beta_t] + \tfrac{1}{2} \operatorname{tr}_{\Sigma_t} D_g^2 v_t, \tag{19}$$

and $V_B = O(B^{-1})$. For equal initial states, with $x^\star$ the population flow,

$$\left(\mathbb{E} \sup_{t \leq T} \|X^\tau(t) - x^\star(t)\|^2 \right)^{1/2} \leq C_T\{B^{-1} + \sqrt{\tau/B} + \tau + e_\tau\} + o(B^{-1}). \tag{20}$$

The four terms in (20) give a finite-horizon error budget for batch estimation, fluctuations, time discretization, and numerical solution. At fixed B, smaller τ approaches $\bar{x}_B$, which can differ from the population path. Targeting a prescribed flow therefore requires control of B and solve accuracy as well as the time step; constants must be calibrated for the chosen model.

The coupling contribution has a corresponding trajectory interpretation. Under the uniform field expansions, $B(\bar{x}_B^{\mathrm{full}} - \bar{x}_B^{\mathrm{fr}}) \to z$, where

$$\dot{z} = D_x v_t^\star(x^\star(t))z + b_{v,t}^{\mathrm{act}}(x^\star(t)), \qquad z(0) = 0. \tag{21}$$

If $b_{v,0}^{\mathrm{act}}(x_0) \neq 0$, the two averaged trajectories separate at order B^{-1} for sufficiently short fixed times. Appendix B.8 states the conditions and proves a local lower bound. Other bias terms can still cancel in a comparison with the population flow.

Under the additional martingale limit conditions in Corollary 8, $\tau^{-1/2}(X^\tau - \bar{x}_B)$ converges to a linear diffusion. Resolving this fluctuation scale requires $e_\tau = o(\sqrt{\tau})$, a stronger requirement than consistency of the averaged path. The argument uses standard stochastic approximation (Kushner & Yin, 2003); the complete statement and proof are in Appendix C.

5.1 Response correction and its limits

Split an even-sized geometry batch into independent halves and use the same current state and fixed energy for three solves:

$$u_B = \mathsf{R}_\tau(\theta, \widehat{g}_B), \quad u_i = \mathsf{R}_\tau(\theta, \widehat{g}_{B/2}^{(i)}), \quad u_B^{\mathrm{RC}} = 2u_B - \tfrac{1}{2}(u_1 + u_2). \tag{22}$$

The full batch is the union of the two halves. Thus B distinct geometry observations and three proximal solves are required; the three endpoints are dependent. This is a split-sample jackknife applied to the entire variational map, using classical bias extrapolation (Quenouille, 1956; Efron, 1979). Gaussian states use mean and matrix-log covariance coordinates; finite simplex states use log-ratio coordinates. The final chart inverse preserves feasibility. In a general chart, extrapolation must remain in its domain.

Theorem 6 (What response correction cancels)**.** *Suppose the same estimator rule at sample sizes $n = B, B/2$ has $m_n := \mathbb{E}\mathsf{R}_\tau(\theta, \widehat{g}_n) = u^\star + b_\tau/n + o(n^{-1})$ and endpoint variance $O(n^{-1})$. Then*

$$\mathbb{E}u_B^{\mathrm{RC}} - u^\star = o(B^{-1}), \qquad \mathrm{Cov}(u_B^{\mathrm{RC}}) = O(B^{-1}). \tag{23}$$

If, in addition, $m_n = u^\star + b_\tau/n + c_\tau/n^2 + o(n^{-2})$, then $\mathbb{E}u_B^{\mathrm{RC}} - u^\star = -2c_\tau/B^2 + o(B^{-2})$. If the full and half samples share an asymptotically linear representation with influence ψ and an L^2 remainder $o(n^{-1/2})$, as specified in (34)*, then $B\,\mathrm{Cov}(u_B^{\mathrm{RC}})$ and $B\,\mathrm{Cov}(u_B)$ both converge to* $\mathrm{Cov}(\psi)$.

Mean cancellation follows from $\mathbb{E}u_B^{\mathrm{RC}} = 2m_B - m_{B/2}$. Equal leading covariance requires the shared influence representation. The B^{-2} rate needs the stronger mean expansion; E4 does not resolve that rate.

For a nonlinear observable, the curvature term $\frac{1}{2}\,\mathrm{tr}(\mathcal{V}_\tau D^2\varphi)$ in (9) generally remains at order B^{-1} after coordinate-mean correction. One can instead extrapolate the *observable values* to cancel their leading bias. That procedure estimates a scalar quantity and differs from applying the observable to a corrected endpoint.

Recursive correction. To infer an $O(B^{-2})$ drift error, require a uniform expansion directly for the small-step fields, or a uniform endpoint expansion separating a $O(\tau/B^2)$ statistical remainder from the $O(\tau^2)$ discretization remainder. Under these extra conditions the averaged corrected field is $2\bar{v}_B - \bar{v}_{B/2} = v^\star + O(B^{-2})$; the bound in (20) replaces B^{-1} by B^{-2} while retaining fluctuation, discretization, and solver terms. A fixed-τ endpoint expansion alone cannot justify division by τ or interchange of the two limits.

Fidelity and risk. For a Euclidean chart and fixed target, mean-squared error is exactly squared mean bias plus variance. Reducing one component need not improve the sum at finite B. For a smooth prediction loss ℓ,

$$\mathbb{E}\ell(u_B) - \ell(u^\star) = B^{-1}\{D\ell(u^\star)b_\tau + \tfrac{1}{2}\,\mathrm{tr}(\mathcal{V}_\tau D^2\ell(u^\star))\} + o(B^{-1}). \tag{24}$$

Removing the first term may raise the loss when that term was negative. A reference variational endpoint is not generally a minimizer of held-out risk. This is a logical explanation of why a fidelity

improvement need not transfer; it does not by itself identify the cause of the observed classification degradation. Accuracy is nonsmooth and is reported empirically without applying (24) to it.

Appendix D.4 gives an exact Gaussian location control with ordinary inverse-metric bias and zero plan-motion contrast.

6 Experiments

Entropic experiments test the endpoint and velocity coefficients in Theorems 2–3. Gaussian tasks instead test correction, reference fidelity, and prediction. Appendix E maps each claim to its evidence and reports all controls, failures, and baselines.

6.1 Predicting the coupling contribution

The five-support problem uses positive probability vectors, the energy $F(u) = \frac{1}{2}\|a(u) - a_{\text{target}}\|^2$, $\varepsilon = 0.1$, and $\tau = 0.05$. Endpoints are expressed in four log-ratio coordinates. Each of five seeded direction sets contains four symmetric metric perturbations and their negatives. The ground map applies a matrix exponential around the reference metric, trace normalization, and a fixed cost scale. We therefore evaluate the general full–frozen formula in Theorem 2. The empirical covariance Σ_m uses the actual perturbation directions.

For $L(g) = \mathsf{R}_\tau(\theta, g) - \mathsf{R}^{\text{fr}}_\tau(\theta, g)$, we compare

$$d_\sigma = \frac{1}{2m}\sum_{j=1}^{m}\{L(g^\star + \sigma h_j) + L(g^\star - \sigma h_j) - 2L(g^\star)\}, \qquad d_\sigma/\sigma^2 \longrightarrow b^{\text{act}}_\tau(\Sigma_m). \tag{25}$$

The notation g denotes the declared perturbation chart, including its map to costs. The prediction is computed from the constrained-plan derivative (12), its endpoint derivative, and the outer Hessian. It is not fitted to the perturbed endpoint contrasts. Recalculation from the saved plan and Hessian intermediates reproduces the stored coefficient vectors within $4 \cdot 10^{-13}$.

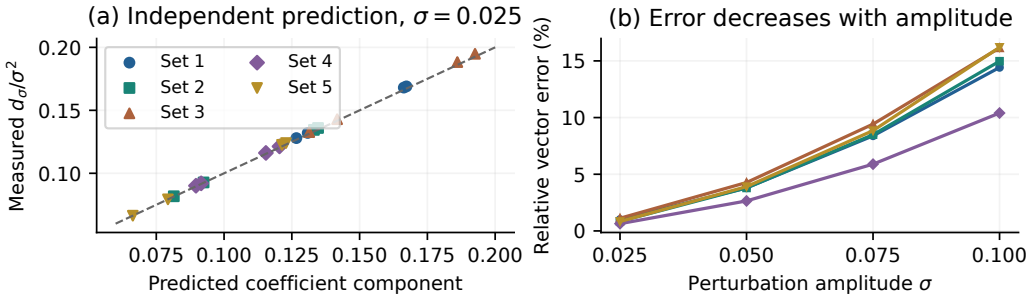

Figure 1: Direct test of the plan-motion coefficient (S3). Each color is one of five direction sets on the same five-support problem. Left: all four components of the independently predicted coefficient versus the measured normalized contrast at $\sigma = 0.025$; the dashed line is equality. Right: relative vector error decreases as the perturbation amplitude decreases. These direction sets vary the noise covariance, not the support geometry.

At $\sigma = 0.025$, relative vector errors range from 0.64% to 1.12%; at $\sigma = 0.05$, they range from 2.65% to 4.27% (Figure 1). Across both amplitudes, the direction cosine is at least 0.99987. Comparing the same predictions with all three archived solver profiles gives a worst error of 4.35%. The error grows to 10.4–16.2% at $\sigma = 0.1$, consistent with the finite-amplitude remainder becoming visible. This is quantitative evidence for the coefficient, beyond the earlier quadratic-scaling test.

Plan-response residuals are below $6.2 \cdot 10^{-12}$; an independent plan finite difference agrees within $1.9 \cdot 10^{-8}$. Hessian-step sensitivity reaches 6.02%. Appendix E.4 reports the precision checks and unresolved cases; these are not certified error bounds.

Does plan response matter quantitatively? Separating the same saved endpoints into full and frozen normalized mean responses shows substantial cancellation. At $\sigma = 0.025$, across all five direction sets and three solver profiles, the active/full norm ratio is 0.908–0.927, but the two vectors point almost oppositely (cosine below -0.9993). The full response has only 0.519–0.524 times the frozen-response norm. Thus plan adaptation can offset the remaining bias; it need not amplify it. These are finite-amplitude, chart-specific comparisons, not percentages of population bias (Appendix E.5).

Does the contrast change observable quantities? Reanalysis of the same endpoints tests all five probability masses, the energy, and Shannon entropy. At $\sigma = 0.025$, their worst coefficient errors across all direction sets and solver profiles are 1.96%, 1.50%, and 1.44%, respectively. Each prediction uses $D\varphi(u^\star)b_\tau^{\mathrm{act}}$; the shared covariance terms cancel. These checks connect the log-ratio response to fixed distributional quantities (Appendix E.6).

6.2 Testing the small-step coefficient

The next test computes b_v^{act} independently from the diagonal tensor and its ground response, then compares it with $b_\tau^{\mathrm{act}}/\tau$ for $\tau \in \{0.05, 0.025, 0.0125, 0.00625\}$. Neither coefficient is extrapolated from the other. All five direction sets use the same state and energy. As Figure 2 shows, the relative discrepancy decreases at every step reduction, reaching 0.66–0.71% at the smallest step. The direction cosine over all twenty comparisons is at least 0.99979. The two-support analytic check gives the same qualitative convergence and provides a separate control for the factors of two in the tensor and response definitions.

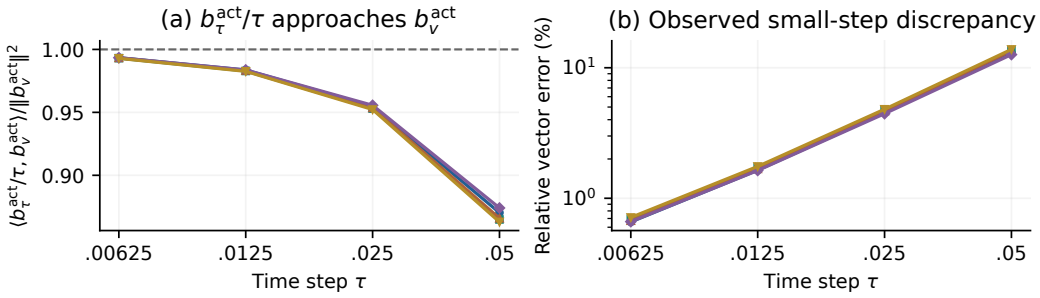

Figure 2: Testing persistence in the velocity (S4). Left: the projection of $b_\tau^{\mathrm{act}}/\tau$ onto b_v^{act}, divided by $\|b_v^{\mathrm{act}}\|^2$, approaches one. Right: the full relative vector discrepancy decreases across all five direction sets. The velocity prediction comes from diagonal derivatives. The smallest-step reference solve was refined; achieved residuals and derivative sensitivity are reported below.

The diagonal tensor has minimum eigenvalue 0.00784. Finer derivative steps change the velocity coefficient by at most 1.10%, close to the smallest observed discrepancy. At the smallest τ, a failed initial solve was refined and both records retained. Thus S4 supports the coefficient bridge without certifying sub-percent accuracy or convergence of entire trajectories.

Trajectory-level check. E3 provides a separate dynamical diagnostic from 200 independent trajectories per configuration. For the stationary model, fitting terminal variance over $\tau \le 0.1$ gives slope 1.017 with a 95% bootstrap interval $[0.905, 1.133]$. The full-range slope is 1.116; harder moving settings retain mean discrepancies. This supports the variance scaling in the tested small-step regime, without establishing the diffusion-process limit (Appendix E.8).

6.3 What bias correction improves

Binary and Gaussian controls separate mean correction from variance reduction. At $B = 64$, RC reduces binary signed mean bias from $2.94 \cdot 10^{-4}$ to $6.68 \cdot 10^{-6}$; the corrected interval includes

zero, but the B^{-2} rate remains unresolved. Archived covariance results show the same bias–MSE distinction (Appendix E.7).

On MNIST and Fashion-MNIST (LeCun et al., 1998; Xiao et al., 2017), ten Gaussian class states track 40 angles from 0° to 90°, using a frozen CNN, 16-dimensional PCA, and disjoint reference/metric banks. Geometry uses (2): $\rho = 1$, no clipping, $\tau = 0.05$, $B \in \{32, 64, 128\}$.

We use ten trajectory seeds for MNIST and five for Fashion-MNIST. In coordinates $\theta = (m, \mathrm{vech}(\log C)) \in \mathbb{R}^{152}$, the tracking error is

$$\mathrm{TE}_s = \frac{1}{39}\sum_{k=1}^{39}\left(\frac{1}{10}\sum_{c=1}^{10}\|\theta_{s,k,c} - \theta_{k,c}^{\mathrm{ref}}\|^2\right)^{1/2}. \tag{26}$$

TE measures finite-bank reference tracking in a fixed chart; its mean/covariance weighting depends on that chart. Reconstruction from all 45 saved checkpoints reproduces the recorded values.

Table 1: Reference tracking with fixed ridge and no clipping. Gain is relative to plug-in B. Intervals are pointwise paired 95% seed-bootstrap intervals (10 MNIST and 5 Fashion-MNIST seeds) for $\mathrm{TE}_{\mathrm{plug}} - \mathrm{TE}_{\mathrm{RC}}$. Plug-in $3B$ uses more samples and fewer solves than RC.

Dataset	B	Plug-in	RC	Gain	Paired 95% CI	Plug-in $3B$
MNIST	32	1.599	1.420	11.2%	[0.071, 0.288]	0.612
MNIST	64	0.888	0.668	24.7%	[0.169, 0.262]	0.407
MNIST	128	0.512	0.424	17.1%	[0.065, 0.107]	0.276
Fashion	32	1.428	1.089	23.8%	[0.297, 0.383]	0.679
Fashion	64	0.941	0.849	9.8%	[0.029, 0.157]	0.501
Fashion	128	0.627	0.581	7.3%	[0.031, 0.060]	0.354

RC lowers mean TE by 7.3–24.7% in all six settings. The conditional bootstrap intervals are positive, but a paired t interval includes zero for Fashion-MNIST at $B = 64$. RC uses B samples/three solves; plug-in $3B$ uses $3B$ samples/one solve and tracks better throughout. Averaging RC's same three solves is worse; EMA and disjoint-batch ensembles are competitive in parts of the grid (Appendix E.10). These compare resource trade-offs, not equal wall-clock budgets.

Across every reported seed and setting, RC has a higher terminal negative log-likelihood (NLL) and lower classification accuracy than plug-in. At $B = 64$, accuracy decreases by 0.97 percentage points on MNIST and 4.17 percentage points on Fashion-MNIST (Appendix Table 10). The reference itself also has lower accuracy than plug-in. Thus improved reference tracking coexists with worse predictive performance in these experiments. Equation (24) permits this tracking–prediction trade-off, but does not identify the empirical cause of the observed deterioration in predictive performance.

Scope and remaining uncertainty. The coefficient tests are confined to one five-support geometry, and finite-bank tracking measures agreement with a reference trajectory rather than accuracy relative to the population target. Appendix Table 14 distinguishes sampling guarantees from numerical diagnostics and states which uniform assumptions remain unverified. It also records the remaining issues concerning temporal dependence and the unresolved B^{-2} convergence rates for correction.

7 Conclusion

Exact solves and vanishing steps can approach a biased flow under independently refreshed estimated geometry. The full–frozen identity isolates optimal-plan response, and the local transport tensor determines its velocity coefficient. A two-support example proves nonvanishing at fixed entropic regularization. Finite-horizon averaging explains why smaller steps suppress sampling fluctuations while a systematic drift can persist. Thus numerical accuracy alone need not recover the population-geometry flow. Five-support tests compare independent coefficient predictions with endpoint contrasts and distributional observables, showing that plan adaptation can partly cancel the frozen response. On frozen-feature tasks, response correction improves finite-bank tracking while worsening prediction. Reference fidelity and held-out predictive performance should therefore be evaluated separately.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendices A–D. Appendix E specifies the models, numerical checks, aggregation rules, and retained failures. The accompanying records and analysis scripts reproduce the S3/S4 coefficient comparisons, the E3 200-seed trajectory statistics, the S2 definition audit, and the real-task summaries. Supplementary source snapshots include the solvers, experiment configurations, environment specifications, encoder-training code, and saved encoder/PCA assets. The large derived feature banks are not duplicated. The original uncommitted state of the S2 main audit was not separately preserved; its completed source and recorded provenance are provided. Full training and optimization pipelines have not been rerun as part of this record-level verification.

AI USE STATEMENT

Large language models were utilized exclusively for language polishing and text formatting. All mathematical proofs, experimental methodologies, and data analyses are the original work of the authors. The authors bear full responsibility for the scientific validity, empirical claims, and final text of this manuscript.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, second edition, 2008.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418, 2017. doi: 10.1137/15M1050264.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Simone Di Marino, Emanuele Naldi, and Silvia Villa. Inexact JKO and proximal-gradient algorithms in the wasserstein space. *arXiv preprint arXiv:2505.23517*, 2025.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Florian Bernard, and Daniel Cremers. A unified framework for implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: Asymptotic and gradient properties. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2131–2141, 2020.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690, 2019. URL `https://proceedings.mlr.press/v89/feydy19a.html`.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1574–1583, 2019.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *Information and Inference: A Journal of the IMA*, 13(1):iaad056, 2024a. doi: 10.1093/imaiai/iaad056.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041, 2024b. doi: 10.1214/24-EJS2217.

Pratik Jawanpuria, Dai Shi, Bamdev Mishra, and Junbin Gao. A riemannian approach to ground metric learning for optimal transport. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890250.

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.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition, 2003. doi: 10.1007/b97441.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M. Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15243–15256, 2021.

Nikhil Nayak, Julia White, Urchade Zaratiana, Kelton Zhang, Henrijs Princis, Dhruv Atreja, Henry Fawcett, Matthew Thomas, George Hurn-Maloney, and Ash Lewis. Correcting stochastic update bias in preconditioned language model optimizers. *arXiv preprint arXiv:2605.20756*, 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Gabriel Peyré, Clarice Poon, and Oscar Tron. Curvature of optimal transport with respect to the cost and applications to inverse optimal transport. *arXiv preprint arXiv:2604.22670*, 2026.

M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3–4):353–360, 1956. doi: 10.1093/biomet/43.3-4.353.

Panpan Ren, Michael Röckner, Feng-Yu Wang, and Simon Wittmann. Stochastic intrinsic gradient flows on the wasserstein space. *arXiv preprint arXiv:2506.12755*, 2025.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Shang Wu and Yazhen Wang. Computational and statistical asymptotic analysis of the JKO scheme for iterative algorithms to update distributions. *arXiv preprint arXiv:2501.06408*, 2025.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Endpoint Response and Coordinate Changes

A.1 Proof of Theorem 1

Throughout this proof, (τ, t, θ) are fixed and all derivatives without arguments are evaluated at $(u^\star, g^\star)$. The implicit-function theorem applied to $q(\mathsf{R}_\tau(\theta, g); \theta, g) = 0$ yields a C^3 local response. Differentiating once in h gives

$$HD_g\mathsf{R}_\tau[h] + q_g[h] = 0.$$

Differentiating this identity in k, including the dependence of H on the endpoint, gives

$$HD_g^2\mathsf{R}_\tau[h,k] + q_{gg}[h,k] + q_{ug}[D_g\mathsf{R}_\tau[h],k] + q_{ug}[D_g\mathsf{R}_\tau[k],h] + q_{uu}[D_g\mathsf{R}_\tau[h], D_g\mathsf{R}_\tau[k]] = 0.$$

Invertibility of H proves (6)–(7). Uniform derivative bounds and $\|H^{-1}\| \le \kappa^{-1}$ bound the first three response derivatives on the branch. Convexity of U permits Taylor expansion along the segment from $g^\star$ to $\widehat{g}_B$:

$$\mathsf{R}_\tau(\theta, \widehat{g}_B) - u^\star = J\Delta_B + \tfrac{1}{2}D_g^2\mathsf{R}_\tau[\Delta_B, \Delta_B] + \eta_B, \qquad \|\eta_B\| \le C\|\Delta_B\|^3.$$

Taking expectations and using (5) proves the mean formula. For covariance, the first-order Taylor remainder ζ_B satisfies

$$\mathsf{R}_\tau - u^\star = J\Delta_B + \zeta_B, \qquad \mathbb{E}\|\zeta_B\|^2 \le C\mathbb{E}\|\Delta_B\|^4 = O(B^{-2}).$$

The cross term is at most $C(\mathbb{E}\|\Delta_B\|^2)^{1/2}(\mathbb{E}\|\zeta_B\|^2)^{1/2} = O(B^{-3/2})$. The squared mean is $O(B^{-2})$, so the covariance is $J\Sigma J^\top/B + o(B^{-1})$.

Apply the same second-order expansion to the composite map $g \mapsto \varphi(\mathsf{R}_\tau(\theta, g))$. Its Hessian in directions h, k is

$$D\varphi(u^\star)D_g^2\mathsf{R}_\tau[h,k] + D^2\varphi(u^\star)[Jh, Jk].$$

Its first derivative applied to the estimator mean yields $D\varphi J\beta/B$. Contracting the Hessian with Σ proves (9). This derivation also makes clear that the observable curvature term is part of the statistical response, even if the chart mean happens to be centered. □

A.2 A localized version and the role of assumptions

The condition $\widehat{g}_B \in U$ almost surely can be replaced by a localization condition. Let $A_B = \{\widehat{g}_B \in U\}$ and let Y_B denote the selected endpoint also on A_B^c. It suffices that the moment expansions hold, the derivative bounds hold on U, and

$$\mathbb{E}\big[(\|\Delta_B\| + \|\Delta_B\|^2 + \|\Delta_B\|^4 + \|Y_B - u^\star\| + \|Y_B - u^\star\|^2)\mathbf{1}_{A_B^c}\big] = o(B^{-1}).$$

For an observable, additionally require $\mathbb{E}[|\varphi(Y_B) - \varphi(u^\star)|\mathbf{1}_{A_B^c}] = o(B^{-1})$. The preceding proof applies on A_B, and all omitted contributions are negligible at the displayed order. These tail requirements cannot be replaced solely by $\Pr(A_B^c) \to 0$: rare large endpoints can change a mean or covariance.

For smooth functions of averages of bounded independent observations, the stated moments follow on a compact neighborhood by Taylor expansion and concentration. A fixed ridge removes the singularity in covariance inversion. It does not establish a global lower Hessian bound for the outer variational objective. In applications, the theorem therefore describes a stable local regime, not every batch size or every parameter value. When B is small relative to feature dimension, asymptotic coefficients may provide an inaccurate quantitative approximation despite smoothness.

Gaussian covariance estimation with a fixed ridge. Here is an explicit unbounded-data example for the estimator part of Assumption 1. Fix d and $\rho > 0$. Let $Z_1, \dots, Z_B$ be i.i.d. $\mathcal{N}(\mu, C_0)$ observations, with $C_0 \succ 0$, and define

$$\widehat{C}_B = \frac{1}{B}\sum_{i=1}^{B}(Z_i - \bar{Z})(Z_i - \bar{Z})^\top, \qquad s_0 = \mathrm{vech}(C_0).$$

In covariance coordinates define the ridge map

$$h(s) = \mathrm{vech}\left(\frac{d\,(C + \rho I)^{-1}}{\mathrm{tr}((C + \rho I)^{-1})}\right), \quad C = \mathrm{vech}^{-1}(s).$$

Use a fixed unweighted half-vectorization throughout. Put $\widehat{g}_B = h(\mathrm{vech}(\widehat{C}_B))$ and $g^\star = h(s_0)$. If Γ is the covariance of $\mathrm{vech}((Z-\mu)(Z-\mu)^\top)$, then its entries are

$$\Gamma_{(ij),(k\ell)} = (C_0)_{ik}(C_0)_{j\ell} + (C_0)_{i\ell}(C_0)_{jk}.$$

The geometry estimator satisfies (5), with derivatives evaluated at s_0 and

$$\beta = -Dh(s_0)s_0 + \tfrac{1}{2}\sum_{\alpha,\gamma}\Gamma_{\alpha\gamma}D^2h(s_0)[e_\alpha, e_\gamma], \qquad \Sigma = Dh(s_0)\Gamma Dh(s_0)^\top.$$

More precisely, $\mathbb{E}\|\widehat{g}_B - g^\star\|^3 = O(B^{-3/2})$ and $\mathbb{E}\|\widehat{g}_B - g^\star\|^4 = O(B^{-2})$. With the unbiased covariance divisor $B-1$, the term $-Dh(s_0)s_0$ disappears; the leading covariance and curvature term are unchanged. The divisor therefore matters for β, even though it does not affect the leading Σ.

To verify these claims, write $\delta_B = \mathrm{vech}(\widehat{C}_B - C_0)$. The centered Gaussian covariance has the Wishart representation with $B-1$ degrees of freedom, giving

$$\mathbb{E}\delta_B = -s_0/B, \qquad \mathbb{E}[\delta_B\delta_B^\top] = \Gamma/B + O(B^{-2}), \qquad \mathbb{E}\|\delta_B\|^k = O(B^{-k/2}) \quad (k=3,4).$$

These moment orders can also be obtained by expanding $\widehat{C}_B = B^{-1}\sum_i (Z_i-\mu)(Z_i-\mu)^\top - (\bar{Z}-\mu)(\bar{Z}-\mu)^\top$: the centered average has third and fourth absolute moments of the stated orders, while the sample-mean term has kth moment $O(B^{-k})$.

On a fixed covariance neighborhood of C_0, the ridge map h has bounded derivatives through order three. Taylor expansion gives the displayed β and Σ. To remove the neighborhood restriction, note that in fixed dimension Gaussian quadratic-form concentration gives $\Pr(\|\widehat{C}_B - C_0\| > r) \le c_r e^{-c_r' B}$ for every fixed $r > 0$, with positive constants depending on r, d, C_0. This follows entrywise from the moment-generating function of centered Gaussian quadratic forms and a finite union bound. Also, the normalized positive matrix inside h has trace d, hence bounded Frobenius norm on the entire positive-semidefinite covariance cone. Its contribution on the exceptional event is exponentially small. Cauchy–Schwarz and the finite Gaussian moments give the same conclusion for the polynomial Taylor terms. The local Taylor remainder has expectation $O(B^{-3/2})$, proving the mean and second-moment expansions; local Lipschitz continuity and the same tail argument prove the third and fourth moment bounds.

This calculation establishes estimator moments, not stability of the outer minimizer. Continuity of h makes $\Pr(\widehat{g}_B \notin U)$ exponentially small for any fixed neighborhood U of $g^\star$. The localized endpoint theorem still requires control of the endpoint and observable on that event. Uniformly bounded fourth moments of the selected endpoint, together with bounded second moments of the observable, are sufficient by Cauchy–Schwarz. Without such control, concentration of the geometry alone does not justify the endpoint expectation. Constants here depend on fixed d, ρ, C_0; no growing-dimension assertion is made.

A.3 Coordinate changes and observable response

The transition and its observables. For an independently drawn geometry estimate $\widehat{g}_B$ at the current state, define

$$K_{\tau,B}(\theta, A) = \Pr\{\mathsf{R}_\tau(\theta, \widehat{g}_B) \in A\}, \quad L_{\tau,B}\varphi(\theta) = \tau^{-1}\mathbb{E}[\varphi(\mathsf{R}_\tau(\theta, \widehat{g}_B)) - \varphi(\theta)]. \tag{27}$$

Time arguments are suppressed for one-step statements. Chart means are computationally useful but depend on the endpoint chart. Scalar observables φ allow the effect on a fixed quantity to be compared under coordinate changes. Energy noise may be conditioned on or included as an additional estimated argument. Independence removes mixed covariances; it does not automatically center a nonlinear endpoint.

Proposition 7 (Invariance for a fixed observable)**.** *Let $h = \chi(g)$ be a smooth invertible ground reparameterization with bounded derivatives through order three on the estimator neighborhood, or corresponding localized tail control. Put $J_\chi = D\chi(g^\star)$. The transformed estimator has*

$$\Sigma_h = J_\chi \Sigma J_\chi^\top, \qquad \beta_h = J_\chi\beta + \tfrac{1}{2}\sum_{i,j}\Sigma_{ij}D^2\chi[e_i, e_j]. \tag{28}$$

Using these coefficients and $\mathsf{R}_\tau(\theta, \chi^{-1}(h))$ leaves the scalar coefficient in (9) *unchanged. Reexpressing the endpoint and the same observable in another smooth chart leaves the scalar expectation unchanged; any Taylor expansion of the transformed quantities requires the same derivative and moment control.*

Scalar observables describe the same measured quantity across charts; the individual vector b_τ remains chart-dependent. Re-centering an estimate after a nonlinear reparameterization would change the noise model. For jointly estimated energy and geometry, the theorem applies to their joint parameter vector. Conditional cross-fitting removes mixed covariances while retaining the pure geometry response.

A.4 Proof of Proposition 7

Taylor expansion of $\chi(g^\star + \Delta_B)$ gives (28). To check cancellation explicitly, put $f = \varphi \circ \mathsf{R}_\tau$, $\eta = \chi^{-1}$, and $J_\chi = D\chi(g^\star)$. The first two derivatives of $f \circ \eta$ are

$$D(f \circ \eta) = Df\, J_\chi^{-1}, \qquad D^2(f \circ \eta)[a,b] = D^2 f[J_\chi^{-1}a, J_\chi^{-1}b] + Df\, D^2\eta[a,b].$$

Differentiating $\eta \circ \chi = \mathrm{Id}$ twice yields

$$D^2\eta[J_\chi h, J_\chi k] = -J_\chi^{-1} D^2\chi[h,k].$$

Substitution into the coefficient $D(f \circ \eta)\beta_h + \frac{1}{2}\operatorname{tr}_{\Sigma_h} D^2(f \circ \eta)$ cancels the $D^2\chi$ terms and leaves $Df\,\beta + \frac{1}{2}\operatorname{tr}_\Sigma D^2 f$. For an endpoint chart change $y = \zeta(u)$, use the observable $\widetilde{\varphi}(y) = \varphi(\zeta^{-1}(y))$; its composite with the transformed endpoint is exactly the original scalar function. This proves invariance of the scalar coefficient, not of a coordinate mean. □

Joint energy and geometry estimates. If $z = (g, h)$ includes estimated energy parameters, use the joint covariance and bias in Theorem 1. Its second-order coefficient contains pure gg, pure hh, and mixed gh derivatives. Conditional independence of $\widehat{g}$ and $\widehat{h}$ makes the off-diagonal covariance block zero. Neither $D^2_{gg}\mathsf{R} : \Sigma_{gg}$ nor the bias of $\widehat{g}$ is thereby removed. If the energy is held fixed, as in the geometry-only response tests, there is no energy-estimation block to begin with.

B Full and Frozen Transport Objectives

B.1 Proof of Theorem 2

Strictly positive marginals, finite costs, and $\varepsilon > 0$ give a unique positive entropic optimum. Let N be a basis of the zero-marginal tangent space. The constrained Hessian is $N^\top H_\pi N \succ 0$, where $H_\pi = \varepsilon \operatorname{diag}(1/\pi_\star)$. A local feasible parameterization is $a \otimes b + Nz$. The implicit-function theorem applied to its stationarity equation makes the optimal plan smooth in the positive marginals and smooth cost parameters. This justifies envelope differentiation and the mixed derivatives below. It does not supply uniform bounds near degenerate marginals or as $\varepsilon \downarrow 0$.

Select both outer branches through the same reference minimizer. For every candidate u at $g^\star$,

$$S_{\varepsilon,g^\star}(a(u), b(\theta)) = S^{\mathrm{fr}}_{\varepsilon,g^\star}(a(u), b(\theta)).$$

These are identities of functions of u, so their endpoint derivatives of every available order agree. By the envelope theorem, for every u their first derivatives in g agree as well:

$$D_g T_{\varepsilon,g^\star}[h] = \langle D_g C(g^\star)[h], \pi_\star \rangle = D_g T^{\mathrm{fr}}_{\varepsilon,g^\star}[h].$$

This identity applies separately to the cross and self terms. It implies equality of q_g and q_{ug}, while the value identity gives equality of $H = q_u$ and q_{uu}. The common q_g and H also give a common J.

The only term left in the difference of (7) is

$$D^2_g\mathsf{R}_\tau - D^2_g\mathsf{R}^{\mathrm{fr}}_\tau = -\frac{1}{2\tau} H^{-1}\nabla_u D^2_g(S_{\varepsilon,g} - S^{\mathrm{fr}}_{\varepsilon,g}).$$

Contract with $\Sigma/2$. The $J\beta$ terms cancel and yield (11). Since H is invertible and $\tau > 0$, the displayed nonzero criterion is both necessary and sufficient. The first-order covariance equality follows from the common J. □

When C is affine, $D^2_g S^{\mathrm{fr}} = 0$. For a nonlinear cost parameterization, the difference formula retains the required cancellation. Terms containing $D^2_g C$ occur in both objectives and cancel there. An experiment performed in a log-metric chart must use this version, or explicitly transform to affine matrix coordinates with the estimator-bias correction in Proposition 7. A measured full-minus-frozen difference alone is not a numerical evaluation of every derivative in the identity.

B.2 Relation to general nested optimization

The curvature mechanism is not specific to the choice of an OT solver. For a smooth inner problem $V(u, g) = \min_z f(u, z, g)$ with a locally unique interior optimizer and invertible f_{zz}, implicit differentiation gives

$$D_g^2 V = f_{gg} - f_{gz} f_{zz}^{-1} f_{zg},$$

with all terms evaluated at the optimizer. Freezing $z_\star(u, g^\star)$ with respect to g, separately for every u, preserves the value and first ground derivative at $g^\star$ and removes the second term from this Hessian. This is the standard envelope/Schur-complement structure behind (12), using feasible plan coordinates.

The transport-specific work is to construct this comparison without violating marginal feasibility, apply it to the three terms of the Sinkhorn divergence, propagate the surviving curvature through the outer minimizer, and control its small-step limit. The raw inner Hessian difference is negative semidefinite, but self-subtraction and the endpoint derivative do not preserve that sign. Consequently, a moving plan need not amplify total endpoint bias. Neither generic inversion bias nor a nonzero inner value curvature alone establishes the nonzero endpoint and velocity criteria in Theorems 2–3.

The reference geometry is an analysis anchor. Replacing $g^\star$ by an estimated reference introduces an additional statistical approximation that is not bounded by the coefficient identity itself. The decomposition is relative to the declared frozen construction; it is not a unique partition of every possible source of model error. Cost-curvature analysis (Peyré et al., 2026), statistical JKO (Wu & Wang, 2025), and stochastic preconditioning (Nayak et al., 2026) address related objects. We do not claim new envelope calculus or a new generic jackknife algorithm.

B.3 The constrained plan derivative

Let $M\pi = r$ denote a full-row-rank representation of the marginal constraints and let N have full column rank with image $\ker M$. Around the positive optimum, feasible perturbations can be written $\pi_\star + Nz$. Projecting the first-order condition of (3) onto this tangent space gives

$$N^\top \{c(g) + \varepsilon \log(\pi/(a \otimes b))\} = 0.$$

Differentiating at fixed marginals yields

$$N^\top H_\pi N \, D_g z[h] + N^\top c_h = 0.$$

The matrix $N^\top H_\pi N$ is positive definite, so solving and multiplying by N gives the first formula in (12). Envelope differentiation gives $(D_g^2 T - D_g^2 T^{\mathrm{fr}})[h, k] = c_h^\top D_g \pi[k]$, proving the second. In the affine case $D_g^2 T^{\mathrm{fr}} = 0$. The expression is independent of the chosen basis: replacing N by NM for invertible M leaves $N(N^\top H_\pi N)^{-1} N^\top$ unchanged.

Equation (12) is evaluated by linear solves, without forming the inverse or the full response projector. For an $n \times n$ plan the tangent dimension is $(n-1)^2$; a generic dense factorization in this basis costs $O(n^6)$ operations. Marginal-constraint formulations or structured solvers can avoid this large dense system. Conditioning depends on ε, the plan entries, and the chosen basis. Small residuals alone do not certify forward accuracy; the reported derivative sweeps provide a separate sensitivity check.

This Hessian can be nonzero while the endpoint contrast is zero. For example, the contraction with Σ can vanish, the cross and self terms can cancel, or the resulting scalar function can have zero endpoint gradient. A nonzero derivative of a single transport plan is therefore insufficient for a nonzero endpoint-response theorem.

B.4 What solver diagnostics establish

Suppose Ψ is κ-strongly convex on a convex neighborhood containing its minimizer $u^\star$ and an approximate endpoint $\widetilde{u}$. Strong monotonicity gives

$$\kappa \|\widetilde{u} - u^\star\|^2 \le \langle \nabla\Psi(\widetilde{u}) - \nabla\Psi(u^\star), \widetilde{u} - u^\star \rangle \le \|\nabla\Psi(\widetilde{u})\| \|\widetilde{u} - u^\star\|.$$

Hence $\|\widetilde{u} - u^\star\| \le \|\nabla\Psi(\widetilde{u})\|/\kappa$. If only an approximate gradient $\widetilde{q}$ is available, a bound on $\|\widetilde{q} - \nabla\Psi\|$ must be added. Inner marginal residuals are not bounds on this gradient error without a separate stability argument. Agreement of two solvers reduces concern about solver-specific failure but cannot rule out their sharing the same inaccurate inner objective.

B.5 Observable contrasts and endpoint coordinates

The two branches have the same population endpoint and first ground derivative. Their leading endpoint covariances therefore coincide. Applying Theorem 1 to a common C^3 observable gives

$$\mathbb{E}\varphi(\mathsf{R}_\tau(x,\widehat{g}_B)) - \mathbb{E}\varphi(\mathsf{R}^{\mathrm{fr}}_\tau(x,\widehat{g}_B)) = B^{-1}D\varphi(u^\star)b^{\mathrm{act}}_\tau + o(B^{-1}). \tag{29}$$

The observable-Hessian terms cancel. For a smooth invertible endpoint chart $y = \psi(u)$ satisfying the derivative and tail conditions of Theorem 1, the same calculation componentwise yields $b^{\mathrm{act}}_{\tau,y} = D\psi(u^\star)b^{\mathrm{act}}_{\tau,u}$. Thus vanishing of this declared contrast is unchanged by a regular endpoint reparameterization. This does not make either individual chart-mean bias intrinsic. A ground-coordinate change must still describe the same random geometry as in Proposition 7.

B.6 Proof of Theorem 3

Suppress x when no ambiguity arises. Symmetry and diagonal vanishing imply $\nabla_u D_g(x,x) = 0$ for every g. The same holds for D^{fr}_g: transposing a plan preserves the frozen construction, and its three diagonal terms cancel. At $g^\star$, equality of the values and their first ground derivatives holds as an identity in both endpoints. Taking two u derivatives yields

$$A^{\mathrm{fr}}_{g^\star} = A_{g^\star} = A, \qquad D_g A^{\mathrm{fr}}_{g^\star} = D_g A_{g^\star}.$$

These equalities prove (13). With the second endpoint held fixed, the mixed derivative bounds also justify

$$Q_\Sigma(x) = \tfrac{1}{2}\nabla^2_{uu}\left[\mathrm{tr}_\Sigma\, D^2_g(D_g - D^{\mathrm{fr}}_g)(u,x)|_{g^\star}\right]_{u=x} = \tfrac{1}{2}\nabla^2_{uu}\mathcal{L}_\Sigma(u,x)|_{u=x}.$$

The factor $1/2$ comes from the definition of the local tensor; it is distinct from the $1/2$ in the second-order mean expansion. When $C(g)$ is affine, $D^{\mathrm{fr}}_g(u,x)$ and hence $A^{\mathrm{fr}}_g(x)$ are affine in g. Positivity at $g^\star$ also guarantees positivity of the frozen tensor on a sufficiently small neighborhood.

To control the small-step limit, multiply stationarity by τ and define

$$Q(u,x,g,\tau) = \tau\nabla F(u) + \tfrac{1}{2}\nabla_u D_g(u,x).$$

At $(u,\tau) = (x,0)$, $Q = 0$ and $D_u Q = A_g(x)$ is uniformly invertible. The parameter-dependent implicit-function theorem provides a local branch R_τ. Its outer Hessian is $A_g/\tau + O(1)$, so it is a local minimum for positive, sufficiently small τ. The assumed mixed derivative bounds give, after reducing the compact neighborhood if necessary,

$$\mathsf{R}_\tau(x,g) = x + \tau v(x,g) + \tau^2 r_\tau(x,g), \quad v(x,g) = -A_g(x)^{-1}\nabla F(x), \quad \sup_{\tau,x}\|r_\tau(x,\cdot)\|_{C^2} < \infty. \tag{30}$$

One way to check the stated mixed regularity with $F \in C^4$ is to set $u = x + \tau w$ before applying the theorem. The equation for w is

$$\nabla F(x+\tau w) + \left[\int_0^1 \tfrac{1}{2}\nabla^2_{uu}D_g(x+s\tau w, x)\,ds\right] w = 0.$$

At $\tau = 0$ its w derivative is A_g. This equation is C^3 in (w,g,τ) under the stated assumptions, so $w - v = O(\tau)$ in $C^2(g)$. This proves (30). The identical argument applies to the frozen penalty. Thus the ground-derivative control is established before taking the limit of the response coefficient.

Write $f = \nabla F(x)$, $A_h = D_g A[h]$, and $A_{hk} = D^2_g A[h,k]$. Direct differentiation gives

$$\begin{aligned} D_g v[h] &= A^{-1}A_h A^{-1} f, \\ D^2_g v[h,k] &= A^{-1}A_{hk}A^{-1}f - A^{-1}A_h A^{-1}A_k A^{-1}f - A^{-1}A_k A^{-1}A_h A^{-1}f. \end{aligned}$$

For the frozen field, A and A_h agree, while the second derivative is A^{fr}_{hk} (zero in the affine case). The first ground derivatives and the estimator-bias terms therefore cancel. Contracting the difference of the second derivatives with $\Sigma/2$ yields (14) for b^{act}_v. Applying (30) to the fixed-step coefficient yields $b^{\mathrm{act}}_\tau = \tau b^{\mathrm{act}}_v + O(\tau^2)$; the estimator moment coefficients are bounded and independent of τ.

Invertibility of A proves the null condition. For a quantitative lower bound, suppose $\|A\| \le a_1$ and the smallest singular value of Q_Σ is at least $q_0 > 0$. Then

$$\|A^{-1}Q_\Sigma A^{-1}f\| \ge a_1^{-1}\|Q_\Sigma A^{-1}f\| \ge q_0 a_1^{-2}\|f\|,$$

which yields $\|b_v^{\mathrm{act}}\| \geq q_0\|\nabla F\|/(2a_1^2)$. Q_Σ is symmetric but need not be positive semidefinite, because the penalty includes the self-subtraction terms. No such sign assumption was used. In a nonaffine ground chart, the same cancellation uses $Q_\Sigma = \mathrm{tr}_\Sigma\, D_g^2(A_g - A_g^{\mathrm{fr}})$. Affinity is needed only to set the frozen second derivative to zero. □

B.7 Proof of Proposition 4

For marginals $(p, 1-p)$ and $(q, 1-q)$, write the coupling as

$$\pi(z) = \begin{pmatrix} z & p-z \\ q-z & 1-p-q+z \end{pmatrix}.$$

Its first-order condition is

$$\frac{z(1-p-q+z)}{(p-z)(q-z)} = \exp(2g/\varepsilon).$$

At $p = q = 1/2$, put $r = \exp(g/\varepsilon)$. The diagonal and off-diagonal entries are $d = r/[2(1+r)]$ and $o = 1/[2(1+r)]$, respectively. Entropic strict convexity makes the optimum interior and smooth in a neighborhood.

Let $T(p, q)$ denote the optimized entropic value at fixed g, ε. Before eliminating z, the mixed (p, q) derivative of the objective is ε/d, its (p, z) and (q, z) derivatives are both $-\varepsilon(1/d + 1/o)$, and its (z, z) derivative is $2\varepsilon(1/d + 1/o)$. Marginal entropy terms in the KL convention have zero mixed derivative. The Schur complement consequently gives

$$T_{pq} = \frac{\varepsilon}{d} - \frac{\varepsilon}{2}\left(\frac{1}{d} + \frac{1}{o}\right) = -2\varepsilon \sinh(g/\varepsilon).$$

Symmetry implies $\partial_{pp} S(p, q)|_{p=q} = -T_{pq}$, so $A_g(1/2) = \varepsilon \sinh(g/\varepsilon)$. It follows that $Q_\Sigma = s^2 \sinh(g/\varepsilon)/\varepsilon$. Substitution into Theorem 3 proves (15).

Strict positivity of A_g, Q_Σ, and $|F'(1/2)|$ is preserved on a sufficiently small compact neighborhood of a point with $g, \varepsilon, s^2 > 0$ and $F'(1/2) \neq 0$. Continuity gives the stated persistence, and the uniform $O(\tau^2)$ remainder gives the small-step lower bound. For example, $F(p) = \lambda(p-a)^2/2$ with $a \neq 1/2$ satisfies the condition. If desired, the scalar cost can be realized on two points $(0, 0), (1, 0)$ under the trace-fixed ambient metric $\mathrm{diag}(g, 2-g)$, for $0 < g < 2$; the result still concerns the two-support entropic model. No zero-regularization limit is asserted. □

B.8 Separation of the two averaged trajectories

Here the reference geometry $g^\star$ and its noise law may depend on time, as in the dynamics section. Suppose full and frozen averaged fields have uniform expansions on a common invariant region,

$$\bar{v}_{t,B}^{\,j}(x) = v_t^\star(x) + B^{-1} b_{v,t}^j(x) + o(B^{-1}), \qquad j \in \{\mathrm{full}, \mathrm{fr}\},$$

where the remainder is uniform, $v_t^\star$ has a bounded continuous state derivative, and the bias fields are continuous and uniformly locally Lipschitz. Let both averaged flows and the population flow start at x_0. Standard parameter differentiation of the integral equation, or Taylor expansion followed by Gronwall's inequality, gives

$$B(\bar{x}_B^{\mathrm{full}} - \bar{x}_B^{\mathrm{fr}}) \longrightarrow z \quad \text{uniformly on } [0, T], \qquad \dot{z} = D_x v_t^\star(x^\star(t))z + b_{v,t}^{\mathrm{act}}(x^\star(t)), \quad z(0) = 0. \tag{31}$$

For completeness, first apply Gronwall to each perturbed flow to obtain $\sup_t \|\bar{x}_B^j - x^\star\| = O(B^{-1})$. Subtract their integral equations and multiply by B. The linearization remainder of $v^\star$ is $o(1)$ uniformly, and continuity of the bias fields replaces their arguments by $x^\star$. A second Gronwall bound gives (31).

If $b_{v,0}^{\mathrm{act}}(x_0) \neq 0$, then $z(t) = t b_{v,0}^{\mathrm{act}}(x_0) + o(t)$ as $t \downarrow 0$. Thus there is a $T_0 > 0$ such that for each fixed $t \in (0, T_0]$ and sufficiently large B,

$$\|\bar{x}_B^{\mathrm{full}}(t) - \bar{x}_B^{\mathrm{fr}}(t)\| \geq \frac{t\|b_{v,0}^{\mathrm{act}}(x_0)\|}{4B}.$$

Theorem 5 then identifies these distinct averaged paths as the small-step limits for their respective independently refreshed schemes. The conclusion compares full and frozen averaged flows. It is not a universal lower bound between the full flow and the population flow: other bias contributions may cancel, and the accumulated contrast may vanish at a later time.

B.9 What an antithetic response check measures

Let $L(g) = \mathsf{R}_\tau(x, g) - \mathsf{R}^{\mathrm{fr}}_\tau(x, g)$ and fix directions $h_1, \dots, h_m$ in a declared smooth ground chart. Then $L(g^\star) = 0$ and $D_g L(g^\star) = 0$. Define

$$d_\sigma = \frac{1}{2m} \sum_{j=1}^{m} \{L(g^\star + \sigma h_j) + L(g^\star - \sigma h_j) - 2L(g^\star)\}, \qquad \Sigma_m = \frac{1}{m} \sum_{j=1}^{m} h_j h_j^\top.$$

If fourth derivatives of L are bounded on these line segments, Taylor's theorem gives

$$d_\sigma = \sigma^2 b^{\mathrm{act}}_\tau(\Sigma_m) + O(\sigma^4). \tag{32}$$

The coefficient uses the actual finite-direction covariance Σ_m, not an assumed isotropic population covariance. For a nonlinear geometry map applied to the chart, differentiate that same composition. Its second-derivative term multiplied by $D_g L$ vanishes, but the cost in the frozen objective must remain the declared cost. Equivalently, use $D^2_g(S - S^{\mathrm{fr}})$ in the nonaffine version of the identity.

Suppose each approximate endpoint has an error bound $\eta^{\mathrm{full}}_{j,\pm}$ or $\eta^{\mathrm{fr}}_{j,\pm}$, and the population endpoints have bounds $\eta^{\mathrm{full}}_0, \eta^{\mathrm{fr}}_0$. The numerical error in the centered contrast is at most

$$E_{\mathrm{num}} = \frac{1}{2m} \sum_{j=1}^{m} \sum_{s \in \{+,-\}} (\eta^{\mathrm{full}}_{j,s} + \eta^{\mathrm{fr}}_{j,s}) + \eta^{\mathrm{full}}_0 + \eta^{\mathrm{fr}}_0.$$

Consequently the error of the normalized coefficient is bounded by $E_{\mathrm{num}}/\sigma^2 + O(\sigma^2)$. This explains why arbitrarily reducing the noise amplitude at fixed solve accuracy is counterproductive. If the exact outer gradient is approximated with error at most δ_q and local strong convexity is certified by κ, one may use $\eta \le (\|\widetilde{q}\| + \delta_q)/\kappa$. Without bounds on δ_q and κ, precision sweeps and solver agreement remain consistency diagnostics rather than endpoint certificates.

C Effective Dynamics and Fluctuations

C.1 The local field of a diagonal penalty

Assume the penalty is sufficiently smooth near the diagonal, its first endpoint derivative vanishes there, and its diagonal Hessian is uniformly positive definite. Multiplying stationarity by 2τ gives

$$2\tau \nabla F_t(u) + \nabla_u D_g(u, x) = 0.$$

At $\tau = 0$ the selected solution is $u = x$. The implicit-function theorem in τ gives

$$\nabla^2_{uu} D_g(x, x)\, \partial_\tau u\big|_{\tau=0} + 2\nabla F_t(x) = 0.$$

Thus $\partial_\tau u|_0 = -A_g(x)^{-1} \nabla F_t(x)$, where $A_g = \frac{1}{2}\nabla^2_{uu} D_g(x, x)$. Uniform bounds on higher derivatives and inverse Hessians give the uniform $O(\tau^2)$ remainder on a compact stable region. A fixed nonzero entropic parameter stays inside A_g; it cannot be removed by this expansion.

C.2 Proof of Theorem 5

Let $\mathcal{F}_k$ contain the initial state and all draws and numerical errors through step $k - 1$. Fresh refreshment gives a martingale difference

$$\xi_{k+1} = v_{t_k}(X_k, \widehat{g}_{k,B}) - \bar{v}_{t_k,B}(X_k), \quad \mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0, \quad \mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] \le V_B.$$

Set $M_n = \sum_{k<n} \tau \xi_{k+1}$. Orthogonality of increments and Doob's inequality imply

$$\mathbb{E} \max_{n \le T/\tau} \|M_n\|^2 \le 4 \sum_{k<T/\tau} \tau^2 \mathbb{E}\|\xi_{k+1}\|^2 \le 4T\tau V_B.$$

The bounded near-identity remainder contributes at most $CT\tau$ to the sum of increments. Time regularity of $\bar{v}$ and boundedness of its deterministic solution similarly bound the accumulated Euler error by $C_T\tau$. Subtracting the integrated averaged ODE at grid points gives, for $e_n = X_n - \bar{x}_B(t_n)$,

$$\|e_n\| \le \|M_n\| + L\tau \sum_{k<n} \|e_k\| + C_T\tau + \sum_{k<n} \|s_k\|.$$

Discrete Gronwall, squaring, and taking expectations give the grid-point version of (18). Linear interpolation between grid points has no larger interpolated grid-error maximum; the difference between the ODE and its linear interpolation is $O(\tau)$ under the stated regularity. This establishes the claimed continuous-time bound.

For the drift expansion apply the response proof componentwise to the function $g \mapsto v_t(x, g)$. Uniform moments and derivatives give (19) and $V_B = O(B^{-1})$. Subtracting the population and averaged ODEs yields

$$\sup_{t \le T} \|\bar{x}_B(t) - x^\star(t)\| \le C_T \sup_{t,x} \|\bar{v}_{t,B}(x) - v_t(x, g_t^\star)\| = O(B^{-1}) + o(B^{-1}).$$

The triangle inequality in L^2 and (18) prove (20). □

The use of the L^2 norm in (17) matters. Merely requiring $\sum_k \mathbb{E}\|s_k\| = o(1)$ controls an L^1 cumulative error and does not imply the displayed mean-square bound. For example, a rare large error can have a small expectation and a large second moment. Deterministic bounds $\|s_k\| \le a_\tau$ suffice with $e_\tau \le (T/\tau + 1)a_\tau$; averaging then requires $a_\tau = o(\tau)$, while the fluctuation limit below requires $a_\tau = o(\tau^{3/2})$. These are sufficient per-step conditions, not claimed necessities.

Corollary 8 (Fluctuations require a finer error scale)**.** *Under Assumption 2, fix B and define $\Sigma_{v,t,B}(x) = \mathrm{Cov}_{Q_{t,B}}(v_t(x, g))$. Additionally assume a bounded Lipschitz state derivative of $\bar{v}_{t,B}$, continuous $\Sigma_{v,t,B}$, and the conditional Lindeberg condition. If $e_\tau = o(\sqrt{\tau})$, then*

$$\tau^{-1/2}(X^\tau - \bar{x}_B) \Rightarrow Z, \qquad dZ_t = D_x \bar{v}_{t,B}(\bar{x}_B(t)) Z_t \, dt + \Sigma_{v,t,B}(\bar{x}_B(t))^{1/2} dW_t, \quad Z_0 = 0. \tag{33}$$

The diffusion describes fluctuations magnified by $\tau^{-1/2}$. Resolving them requires the finer solver scale in Corollary 8, since an $o(1)$ cumulative error can still exceed $\sqrt{\tau}$. The proof follows stochastic approximation and martingale limit theory (Kushner & Yin, 2003). Appendix E reports the available variance-scaling diagnostics separately from the process-limit theorem.

C.3 Proof of Corollary 8

For fixed B, define the rescaled martingale $\widetilde{M}^\tau(t) = \sqrt{\tau} \sum_{k < t/\tau} \xi_{k+1}$. Its predictable covariance process is

$$\left\langle \widetilde{M}^\tau \right\rangle_t = \tau \sum_{k < t/\tau} \Sigma_{v,t_k,B}(X_k) \;\longrightarrow\; \int_0^t \Sigma_{v,s,B}(\bar{x}_B(s))\, ds$$

in probability, using Theorem 5, covariance continuity, localization, and Riemann-sum convergence. The conditional Lindeberg assumption ensures that large martingale jumps vanish. The martingale functional central limit theorem therefore identifies the limiting noise as the continuous Gaussian martingale with this covariance.

For the drift, the bounded Lipschitz derivative gives

$$\bar{v}_{t,B}(X) - \bar{v}_{t,B}(\bar{x}) = D_x \bar{v}_{t,B}(\bar{x})(X - \bar{x}) + O(\|X - \bar{x}\|^2).$$

The averaging bound with $e_\tau = o(\sqrt{\tau})$ makes the integrated quadratic remainder, after division by $\sqrt{\tau}$, tend to zero in probability. The accumulated near-identity and Euler remainders are $O(\tau)/\sqrt{\tau} = o(1)$; the numerical remainder is $o_{L^2}(1)$ on this scale. Convergence of the resulting linear integral equation gives (33). The $(2 + \delta)$ moment and Lindeberg control remove interpolation jumps. This proves convergence in the continuous-path topology for linear interpolation, or in the usual Skorokhod topology for step interpolation with the same continuous limit. □

Limit order and time-dependent targets. The corollary fixes B. Replacing its covariance by Σ_v/B while allowing B to grow is a triangular-array assertion needing uniform expansions and Lindeberg conditions. It is not part of that statement. Likewise, estimates based on overlapping historical batches, a jointly learned representation, or an adaptive sampling law require a different conditional analysis. The independent-refresh model is appropriate for the controlled setting here; it does not cover every optimizer with an EMA geometry.

C.4 WHY THE GEOMETRY REFRESH PROTOCOL MATTERS

Theorem 5 averages fresh geometry noise over increasingly many steps. It does not describe an estimate learned once and reused throughout the path. For a one-dimensional quadratic energy $F(x) = x^2/2$ and penalty $D_g(u, x) = g(u - x)^2$, the exact update is

$$\mathsf{R}_\tau(x, g) = \frac{g}{g + \tau} x.$$

If one positive estimate $\widehat{g}_B$ is fixed for the entire trajectory, then at fixed t,

$$X^{\tau,B}(t) \longrightarrow x_B(t) = x_0 e^{-t/\widehat{g}_B} \qquad (\tau \downarrow 0).$$

The limit remains random. Suppose $\widehat{g}_B$ stays in a compact positive interval and satisfies Assumption 1's scalar moment conditions with coefficient $\Sigma = s^2 > 0$. Applying the delta method to the explicit solution gives

$$\mathrm{Var}(x_B(t)) = \left(x_0 e^{-t/g^\star} \frac{t}{(g^\star)^2} \right)^2 \frac{s^2}{B} + o(B^{-1}).$$

For $x_0 \neq 0$ and $t > 0$, this variance is of order B^{-1}, rather than τ/B. This example lies outside independent refreshment and illustrates its necessity, not a failure of the theorem. EMA, overlapping historical samples, or a state-dependent sampling law require temporal or conditional dependence analysis. Their memory cannot be discarded merely because each individual geometry estimate has the same marginal moments.

D RESPONSE CORRECTION

The common influence representation referred to in Theorem 6 is

$$u_n - u^\star = n^{-1} \sum_{i=1}^{n} \psi(Z_i) + r_n, \qquad \mathbb{E}\psi = 0, \quad \mathbb{E}\|\psi\|^2 < \infty, \quad \|r_n\|_{L^2} = o(n^{-1/2}). \tag{34}$$

D.1 PROOF OF THEOREM 6

The two halves have the same marginal law, and their union has the full-batch law. Linearity of expectation gives

$$\mathbb{E}u_B^{\mathrm{RC}} = 2m_B - m_{B/2}.$$

Substituting $m_n = u^\star + b_\tau/n + o(n^{-1})$ cancels the B^{-1} term. With the stronger expansion,

$$2\left(u^\star + \frac{b_\tau}{B} + \frac{c_\tau}{B^2}\right) - \left(u^\star + \frac{2b_\tau}{B} + \frac{4c_\tau}{B^2}\right) = u^\star - \frac{2c_\tau}{B^2},$$

up to $o(B^{-2})$.

No independence of the three endpoints is used for the covariance order. The triangle inequality for centered random vectors in L^2 and the assumed variance orders give

$$\|u_B^{\mathrm{RC}} - \mathbb{E}u_B^{\mathrm{RC}}\|_{L^2} \leq 2\|u_B - m_B\|_{L^2} + \tfrac{1}{2}\sum_{i=1}^{2}\|u_i - m_{B/2}\|_{L^2} = O(B^{-1/2}).$$

For the leading covariance, let $A_i = (2/B)\sum_{j \in I_i} \psi(Z_j)$ for the two halves. The full-batch first-order term is $(A_1 + A_2)/2$. The corrected first-order term is $2(A_1 + A_2)/2 - (A_1 + A_2)/2 = (A_1 + A_2)/2$ as well. The remainder is $o_{L^2}(B^{-1/2})$ by (34). This proves the covariance limit. □

D.2 RECURSIVE CORRECTION REQUIRES UNIFORMITY

Here is a sufficient formulation of the additional condition in Section 5.1. Suppose, uniformly on the invariant state region,

$$\mathbb{E}v_t(x, \widehat{g}_n) = v_t(x, g_t^\star) + b_{v,t}(x)/n + c_{v,t}(x)/n^2 + o(n^{-2}),$$

with the derivatives, moments, and Lipschitz bounds required in Assumption 2. Construct the corrected random field from the same full/half sample pattern. It has mean $v^\star - 2c_v/B^2 + o(B^{-2})$, variance $O(B^{-1})$, and uniformly bounded near-identity remainder if the endpoint remainders are uniformly bounded. Apply the averaging proof to this field to obtain the corrected version of (20). Feasibility and invariance of the corrected chart path remain assumptions. An expansion with a remainder merely $o(B^{-1})$ at each fixed τ supplies none of these uniform conclusions.

D.3 Observable and risk qualifications

Let $Y_B = u_B^{\mathrm{RC}} - u^\star$ satisfy $\mathbb{E}Y_B = o(B^{-1})$ and $\mathbb{E}[Y_B Y_B^\top] = \mathcal{V}_\tau/B + o(B^{-1})$. If it also has a negligible third moment at scale B^{-1}, Taylor expansion gives

$$\mathbb{E}\varphi(u_B^{\mathrm{RC}}) - \varphi(u^\star) = \frac{1}{2B}\operatorname{tr}(\mathcal{V}_\tau D^2\varphi(u^\star)) + o(B^{-1}).$$

This contrasts with the scalar jackknife $2\varphi(u_B) - \frac{1}{2}(\varphi(u_1) + \varphi(u_2))$, whose expectation cancels the entire leading scalar-response coefficient. They are different estimators with different feasibility and utility meanings.

For a chart-valued estimator Y, the identity $\mathbb{E}\|Y - u^\star\|^2 = \|\mathbb{E}Y - u^\star\|^2 + \operatorname{tr}\operatorname{Cov}(Y)$ is exact. For held-out smooth loss, (24) follows from Theorem 1; neither the linear term nor the change induced by removing it has a prescribed sign. Furthermore, the stored trajectory metric in (26) takes a square root before averaging over time and seeds. It cannot be substituted into this squared-error decomposition. These distinctions are necessary to interpret E4–E6 without turning a bias theorem into an unsupported claim about all metrics.

D.4 An exact location-family control

Consider a Gaussian location family with scalar mean x, fixed covariance, energy $F(x) = \lambda(x-a)^2/2$ with $\lambda > 0$, and squared transport penalty $D_g(u, x) = g(u - x)^2$. For $g \in [m, M]$ with $m > 0$, the exact update and its local field are

$$\mathsf{R}_\tau(x, g) - a = \frac{g}{g + \tau\lambda}(x - a), \qquad v(x, g) = -\lambda g^{-1}(x - a). \tag{35}$$

Every realized step decreases the energy. Nevertheless, if $\mathbb{E}\widehat{g} = g^\star$ and the estimate is nonconstant, Jensen's inequality gives $\mathbb{E}[\widehat{g}^{-1}] > (g^\star)^{-1}$, so the averaged and population flows contract at different rates. They share the minimizer a but disagree at finite time. For example, equal-probability values $g^\star \pm s$, with $0 < s < g^\star$, give $\mathbb{E}[\widehat{g}^{-1}] = g^\star/((g^\star)^2 - s^2)$. Thus metric-induced trajectory bias does not require an energy-descent failure or energy–metric correlation.

There is also an exact finite-step fluctuation calculation. For independent refreshes, let $\alpha_\tau = \widehat{g}/(\widehat{g} + \tau\lambda)$ and $N = T/\tau$ be an integer. Then

$$\mathbb{E}(X_N - a) = (x_0 - a)(\mathbb{E}\alpha_\tau)^N, \tag{36}$$

$$\begin{aligned}\operatorname{Var}(X_N) &= (x_0 - a)^2\{(\mathbb{E}\alpha_\tau^2)^N - (\mathbb{E}\alpha_\tau)^{2N}\} \\ &= \tau T\lambda^2 \operatorname{Var}(\widehat{g}^{-1})(x_0 - a)^2 e^{-2\lambda T\mathbb{E}[\widehat{g}^{-1}]} + O(\tau^2).\end{aligned} \tag{37}$$

Bounded positive g makes the Taylor remainder uniform. If its estimation variance is $O(B^{-1})$ under the smooth moment conditions, this is the τ/B fluctuation scale of Theorem 5. The mean instead tends to $a + (x_0 - a)e^{-\lambda T\mathbb{E}[\widehat{g}^{-1}]}$. This elementary example is a negative control for novelty: it verifies the separation of drift and variance without any transport-plan response. The additional identity in Theorem 2, rather than the mere existence of bias, is needed to distinguish the transport contribution.

E Experimental Protocol, Evidence, and Limitations

Implementation environment. All experiments were executed under Python 3.11.16 using PyTorch 2.8.0 with CUDA 12.8 support. The software stack included NumPy 2.4.6, SciPy 1.17.1, scikit-learn 1.9.0, POT 0.9.7.post1, and GeomLoss 0.3.1. GPU-accelerated computations were carried out on an NVIDIA GeForce RTX 4090. Exact dependency versions are recorded in the released code environment.

E.1 What each experiment can establish

Table 2 connects the theoretical statements to the available experiments. Analytic results, independently computed coefficient comparisons, and numerical diagnostics have separate roles. Appendix F summarizes the assumptions and the evidence terminology.

Table 2: Theory and empirical evidence. The targeted S3 and S4 tests complete the local coefficient comparisons.

Evidence	Supported conclusion	Boundary of the evidence
E1	Theorem 1: local second-order endpoint response	Also occurs under inverse preconditioning; this experiment alone cannot isolate coupling response.
E2 audit	Theorem 2: quadratic contrast and outer-solver agreement in the five-point case	S3 supplies independent coefficient agreement; other audited configurations remain unresolved.
Two-support model	Proposition 4: analytic nonzero example for Theorem 3	S1 checks the analytic example; S4 checks the five-point coefficient bridge, not a trajectory limit.
E3, 200 seeds	Theorem 5 and Corollary 8: finite-horizon averaging diagnostics and variance scaling	Complete 200-seed records permit recomputation; finite-step mean discrepancies remain in harder moving settings. Variance scaling does not establish a process limit.
E4	Theorem 6: reduction of signed chart-mean bias	The corrected B^{-2} rate is unresolved; variance and MSE need not improve.
E5–E6	Fidelity–risk distinction: lower empirical-reference TE in six settings	Gaussian states on frozen features; prediction worsens and larger-batch plug-in tracks better.

E.2 Data lineage and statistical aggregation

The analysis bundle contains the selected E1/E4 tables, the E2 solver audit, the complete E3 trajectories and summaries, seed-level E5/E6 metrics, and the targeted S0/S1/S3/S4 records and derivative intermediates. Its source hashes record the copied inputs. All 45 real-task trajectory checkpoints were checked before extracting the seed-level table: the chart reconstruction agrees with the stored tracking arrays within $1.8 \cdot 10^{-15}$. The typesetting bundle contains the derived real-task metrics and checkpoint hashes. The supplementary source bundle adds solvers, training code, saved encoders, PCA transforms, and split definitions; it does not duplicate the large derived feature banks. The script regenerates the figures, tables, and paired intervals, and independently evaluates the S3/S4 formulas from stored derivatives. It does not rerun an optimizer or generate new experimental observations.

The unit of replication is a complete seeded trajectory. For each real-task seed, TE is computed over steps 1–39, excluding the identical initial state. Methods are paired within that seed. We resample the seed indices with replacement 20,000 times, using bootstrap seed 20260914, and report the 2.5th and 97.5th percentiles of the paired mean difference. Time steps and classes are not counted as independent replicates. Intervals are pointwise, not simultaneous confidence regions over the entire configuration grid. With five Fashion-MNIST seeds, uncertainty estimates remain limited.

All real-task intervals are conditional on the shared encoder, PCA transform, and reference bank. They do not include variability from retraining the encoder, refitting the projection, or resampling the reference bank. The disjoint bank avoids direct metric/reference sample reuse but does not turn a finite reference into the population geometry.

For E3, the complete archive contains 1,600 nonempty trajectory CSVs and a merged table with 48,600 rows, covering eight configurations with 200 seeds each. Every trajectory has its prescribed consecutive steps. The merged rows agree with the individual records. We recompute terminal means, variance traces, Monte Carlo standard errors, and bootstrap intervals from these trajectories. This recovered export supersedes the incomplete earlier copy; the experiments themselves have not been rerun. A norm is nonnegative, so an ordinary bootstrap interval for a norm near zero can be misleading. Statements about agreement of a mean with a reference use signed components and Monte Carlo standard errors, rather than interpreting a positive lower bound for an error norm as a

test against zero. The small-step variance fit is explicitly restricted to $\tau \leq 0.1$; the full-range fit is also reported.

E.3 One-step and transport experiments

E1 contains a binary response example and Gaussian location/covariance families in dimensions 2 and 10. The binary amplitude and batch sweeps use 20 seeds. Their descriptive log-slopes are 2.034 in amplitude and -1.073 in batch size; Gaussian response-direction cosines exceed 0.997. The response prediction is a second-order expansion in the specified perturbation chart. A slope near two in noise amplitude does not independently test coordinate invariance or prove the assumptions of a uniform batch-size expansion. Nonlinear metric estimators require the β term of Theorem 1.

E2 freezes all three entropic plans with respect to the metric for each candidate marginal pair, as in (10). Freezing only the cross plan would leave self-plan motion in the control. Freezing a plan at one candidate and reusing it with incompatible marginals would define a different problem. The supplied source implements this candidate-dependent construction. The S2 value, derivative, and marginal checks provide independent record-level checks of the declared objective.

The audit includes five seeds, four perturbation directions, and amplitudes $\sigma \in \{0.025, 0.05, 0.075, 0.1\}$ for each configuration. Its original profile uses BFGS with 1,000 Sinkhorn iterations, inner threshold 10^{-11}, outer tolerance 10^{-8}, and 300 outer iterations. The higher-accuracy profiles use BFGS or L-BFGS-B, a dual-Newton inner solver with at most 200 iterations and threshold 10^{-12}, finite-difference step 10^{-7}, outer tolerance 10^{-10}, and 1,000 outer iterations. These tolerances are settings, not achieved error certificates.

Table 3: The complete audited configuration grid. The main text retains the most stable contrast but reports the unresolved cases.

Configuration	(n, ε, τ)	Outcome and interpretation
Larger support	$(5, 0.1, 0.05)$	Full/frozen endpoints agree across the two dual-solver profiles within $5.26 \cdot 10^{-7}$ in L^1; gap disagreement is at most $8.78 \cdot 10^{-7}$. Consistent positive evidence.
Three-point baseline	$(3, 0.1, 0.05)$	Endpoint differences reach $2.08 \cdot 10^{-3}$ and some gap directions reverse. Earlier blanket robustness claims are unsupported.
Low regularization	$(3, 0.03, 0.05)$	Near-zero gaps occur with failed solves and large gradient residuals, reaching approximately 0.69. Neither a true zero nor a fully diagnosed numerical artifact is established.
Smaller step	$(3, 0.1, 0.025)$	Gap differences reach approximately $2.0 \cdot 10^{-4}$ with direction reversals. The tested precision does not settle the response.

Additional analysis of stored five-point endpoints. For each seed, we reconstruct log-ratio coordinates with the last probability as denominator, average the eight signed perturbations (four directions), subtract that method's population endpoint, and then take the full-minus-frozen difference. Taking a norm is the final operation. This reproduces the supplied contrast vectors. Reconstruction S0 yields descriptive log–log slopes $2.085, 2.084, 2.094, 2.062, 2.084$ for direction sets 0–4 under the original solver profile. These are descriptive fits without a claimed population confidence interval. The finite set of directions defines an empirical noise covariance; it is not replaced by an identity covariance.

Between the two smallest amplitudes, the change of the vector d_σ/σ^2, relative to its value at $\sigma = 0.025$, ranges from 2.0% to 3.2%. The largest discrepancy between the two dual-solver profiles, divided by the alternative profile's contrast norm, is 0.003663. Shared inner computations can still yield shared errors. S3 completes the stronger test: it compares d_σ/σ^2 with an independently evaluated coefficient, using the same chart, actual directions, and candidate-dependent frozen objective. Appendix B.9 gives its truncation and numerical-error requirements.

The Gaussian-location negative control gives a zero full-minus-frozen gap, as expected when the coupling is a translation independent of a constant Mahalanobis geometry. It remains a useful control even though both endpoints can have inverse-metric bias. The five-point coefficient comparison supports additional coupling response; it does not claim that all stochastic metric effects are transport-specific.

E.4 Targeted coefficient validation: S0–S4

Model and coordinates. The five support points are $(0,0)$, $(1,0)$, $(0.2,1.1)$, $(1.15,1)$, and $(-0.35,0.75)$. The current marginal is $(0.30,0.25,0.20,0.15,0.10)$ and the energy target is its reverse. The reference matrix and fixed cost scale are

$$G_\star = \begin{pmatrix} 0.8 & 0.25 \\ 0.25 & 1.2 \end{pmatrix}, \qquad c_0 = 1.4085.$$

The saved configuration uses $F(u) = \frac{1}{2}\|a(u) - a_{\mathrm{target}}\|^2$ and the three-term Sinkhorn penalty at $\varepsilon = 0.1$. The endpoint chart is $u_i = \log(a_i/a_5)$ for $i = 1,\dots,4$. The ground perturbation is mapped through a matrix exponential and trace normalization before forming the quadratic support costs, divided by the fixed scale c_0. The saved direction archive includes the actual symmetric matrices, their empirical covariance, and probe costs. We retain those directions without renormalization. The supplied source and coordinate manifest specify the map explicitly:

$$G_j(h) = \frac{2\exp(\log G_\star + h\xi_j)}{\operatorname{tr}\exp(\log G_\star + h\xi_j)}, \qquad C_{kl,j}(h) = \frac{(x_k - x_l)^\top G_j(h)(x_k - x_l)}{c_0}.$$

The regularization $\varepsilon = 0.1$ and scale c_0 remain fixed as h varies. The map is nonlinear, so all comparisons use the general full–frozen curvature difference.

There are five seeded sets of four directions. They share the support, current state, energy, and reference matrix. A "seed" in S3/S4 therefore indexes a noise-direction set, not an independent dataset, support geometry, or trained model. These deterministic conditional comparisons have no population confidence interval. The contraction uses $\Sigma_m = m^{-1}\sum_j h_j h_j^\top$ for each set.

S0: reconstruction of the earlier contrast. The stored perturbed endpoints are first transformed to log-ratio coordinates. Signed perturbations are averaged, each method's reference endpoint is subtracted, and the full-minus-frozen vector is formed. The norm is taken only after these operations. The reconstructed vectors agree with the supplied historical contrasts within $1.7 \cdot 10^{-16}$. This establishes that the new coefficient comparison targets the same measured quantity as the earlier five-point experiment.

S1: analytic two-support control. The high-precision two-support records give $A = 1.8134302039$ and $b_v^{\mathrm{act}} = 0.2757205648$ in their probability coordinate. At $\tau = 0.05$, the predicted fixed-step coefficient is 0.0130258493; the measured normalized contrast at $\sigma = 0.05$ is 0.0132809503, a 1.96% difference. The values of $b_\tau^{\mathrm{act}}/\tau$ are $0.260517, 0.268114, 0.271918, 0.273820$ over the four decreasing step sizes, with relative errors $5.51\%, 2.76\%, 1.38\%, 0.69\%$. These records check the exact example and normalization, independently of the five-point problem.

S2: definition audit. The main audit checks nine candidate marginals: the reference endpoint and positive/negative perturbations in each log-ratio coordinate. It records zero full/frozen value difference at the reference geometry, first-ground-derivative differences at most $1.32 \cdot 10^{-11}$, and marginal residuals at most $9.33 \cdot 10^{-12}$. The reference outer Hessian has minimum eigenvalue 0.16057; ten initializations have maximum endpoint difference $3.32 \cdot 10^{-7}$, and the zero-energy identity error is $2.88 \cdot 10^{-9}$.

The raw main-audit tables and intermediate arrays are available. A supplementary audit supplies 1,350 plan-level index entries, 108 derivative comparisons, 18 neighborhood Hessians, and four diagonal tensors. Every index resolves to a stored array. Reconstructing costs from the support and saved metrics, checking the plan marginals, and recomputing Hessian eigenvalues reproduces the diagnostics: the maximum marginal residual is $9.35 \cdot 10^{-12}$, the minimum neighborhood-Hessian eigenvalue is 0.15976, and the minimum diagonal-tensor eigenvalue is 0.007844. The larger

derivative grid has maximum full/frozen difference $4.43 \cdot 10^{-10}$. These finite checks support local numerical consistency; they do not prove branch uniqueness or uniform strong convexity.

Source provenance. The main S2 preparation manifest records uncommitted modifications and no separate patch was retained. Its completed source snapshot is supplied, but the exact uncommitted state cannot be recovered from the preparation commit alone. The supplementary S2 audit records a clean source revision. E3 includes both its trajectory-generation and final-analysis source snapshots. These version distinctions are retained in the supplementary manifest.

S3: an independent endpoint coefficient. Index the cross, candidate-self, and current-self terms by $s = 1, 2, 3$, with weights $w_1 = 1$ and $w_2 = w_3 = -1/2$. All plans below are vectorized and evaluated at $g^\star$. Let $\pi_s \in \mathbb{R}^{n^2}$ denote a plan, $z_{s,j} = D_g \pi_s[h_j] \in \mathbb{R}^{n^2}$ its ground-direction derivative from (12), and $P_s = D_u \pi_s \in \mathbb{R}^{n^2 \times p}$ its endpoint Jacobian. Thus $z_{s,j}$ is a derivative, not a solver residual. For the candidate-self term, P_2 differentiates both marginals; $P_3 = 0$ because the current state is fixed. Here $n = 5$, $p = 4$, and $m = 4$ directions.

Write $\mathcal{L}_m = \mathcal{L}_{\Sigma_m}$ for the curvature contraction defined after Theorem 3. At fixed support, c_{h_j} is independent of u, so

$$\nabla_u \mathcal{L}_m(u, x) = -\frac{\varepsilon}{m} \sum_{j=1}^{m} \sum_{s=1}^{3} w_s P_s^\top \left[(z_{s,j} \oslash \pi_s)^{\odot 2} \right] \in \mathbb{R}^p. \tag{38}$$

Here $\oslash$ and $\odot 2$ mean elementwise division and squaring. To derive this expression, the directional curvature is $c_{h_j}^\top z_{s,j} = -c_{h_j}^\top N (N^\top H_{\pi_s} N)^{-1} N^\top c_{h_j}$. Differentiate it using $D_u H_{\pi_s}[v] = -\varepsilon \operatorname{diag}((P_s v) \oslash \pi_s^{\odot 2})$. The S3 prediction is $b_{\mathrm{pred}} = -(4\tau)^{-1} H^{-1} \nabla_u \mathcal{L}_m(u^\star, x)$. The prediction uses the outer endpoint $u^\star$, whereas S4 below uses the diagonal $u = x$. The analysis code recomputes the plan response, this gradient, and the outer solve from saved intermediates. Its largest vector discrepancy from the stored predictions is $3.5 \cdot 10^{-13}$; this is an algebraic reproduction check, not a bound on the underlying derivative approximation.

Table 4: S3 coefficient comparison, original solver profile. Errors compare the full vectors, not only their norms. The five rows vary the perturbation directions on one support geometry.

Direction set	$\|b_{\mathrm{pred}}\|$	Error, $\sigma = .025$	Error, $\sigma = .05$	Min. cosine
1	0.29801	1.01%	3.80%	0.999999
2	0.22593	0.85%	3.79%	0.999945
3	0.33025	1.12%	4.27%	0.999987
4	0.21035	0.64%	2.65%	0.999995
5	0.20125	0.83%	3.91%	0.999873

The small-amplitude numerical criteria used in this analysis are cosine at least 0.98, relative vector error at most 20%, and change across the two finer derivative steps at most 5%. All five sets satisfy these descriptive thresholds. The independent endpoint profiles give worst errors of 4.27% (original), 4.30% (higher-precision dual/BFGS), and 4.35% (dual/L-BFGS-B) over the two smaller amplitudes. The independent plan finite difference has relative discrepancy $1.89 \cdot 10^{-8}$ at its finest step. The constrained curvature systems can be ill-conditioned: the maximum reported condition number is approximately $1.22 \cdot 10^7$. Residual and derivative-step checks are therefore retained alongside the agreement. These criteria describe numerical agreement, not a population hypothesis test.

The supplied "numerical error budget" reaches 6.02%. Inspection shows that this quantity is a coefficient sensitivity across Hessian steps, not a proved sum of endpoint-error bounds. We report it as a sensitivity diagnostic. Passing the thresholds supports the local numerical prediction; it does not turn the computation into a certified error bound or establish robustness over arbitrary support geometries.

S4: independent velocity coefficient. At fixed current state x, define the saved $p \times p$ matrices

$$A = \tfrac{1}{2} \nabla_{uu}^2 S_{\varepsilon, g^\star}(a(u), a(x))\big|_{u=x}, \qquad K = \nabla_{uu}^2 \mathcal{L}_m(u, x)\big|_{u=x}.$$

Differentiating the half-Hessian definition of $A_g - A_g^{\mathrm{fr}}$ twice in g gives $Q_{\Sigma_m} = K/2$. Hence Theorem 3 is evaluated as

$$b_v^{\mathrm{act}} = \tfrac{1}{4} A^{-1} K A^{-1} \nabla F(x).$$

Neither Hessian is a total derivative along $(u, x) = (x, x)$. The continuous coefficient is evaluated at $u = x$; each fixed-step coefficient is evaluated at its separately computed reference endpoint. Recalculation from the saved matrices reproduces both. Table 5 gives every direction set and step size.

Table 5: S4: relative vector errors (%) between $b_\tau^{\mathrm{act}}/\tau$ and the independently computed b_v^{act}. All rows decrease as τ decreases.

Direction set	$\|b_v^{\mathrm{act}}\|$	$\tau = .05$	.025	.0125	.00625
1	6.84817	12.989	4.525	1.641	0.659
2	5.22373	13.582	4.726	1.719	0.696
3	7.62775	13.448	4.653	1.679	0.673
4	4.81308	12.618	4.448	1.633	0.662
5	4.66331	13.818	4.808	1.753	0.713

The finer derivative-step variations are at most 1.094% for the velocity and 0.571% for fixed-step coefficients. The minimum diagonal-tensor eigenvalue is 0.007844 and the minimum outer-Hessian eigenvalue is 0.160569. The maximum recorded zero-energy error is $2.88 \cdot 10^{-9}$ and the multi-initialization endpoint difference is $4.16 \cdot 10^{-7}$. These measurements concern tested points, not uniform regularity throughout a neighborhood.

At $\tau = 0.00625$, an initial reference solve had gradient norm about $1.34 \cdot 10^{-5}$, exceeding the 10^{-5} threshold. A refinement reduced the checked gradient to $6.73 \cdot 10^{-10}$ and the objective by $3.39 \cdot 10^{-11}$, with logit displacement $7.3 \cdot 10^{-6}$. Its optimizer flag still reported precision loss. We retain both records and use the achieved residual and derivative diagnostics, rather than replacing an unsuccessful flag with a success label. The refined coefficient errors remain monotone. Their smallest values are close to the derivative sensitivity scale, so the experiment supports the direction and limiting coefficient without establishing sub-percent numerical certification.

E.5 Magnitude and cancellation of the coupling response

The coefficient tests establish agreement with a prediction, but do not by themselves report the contribution's size relative to the full response. The same archived probability endpoints permit that comparison without new solves. Transform every endpoint to the declared four-dimensional log-ratio chart before averaging. For $j \in \{\mathrm{full}, \mathrm{fr}\}$, define

$$r_\sigma^j = \frac{1}{\sigma^2}\left[\frac{1}{2m}\sum_{\ell=1}^{m}\sum_{s\in\{-1,1\}} \mathsf{R}_\tau^j(\theta, g^\star + s\sigma h_\ell) - \mathsf{R}_\tau^j(\theta, g^\star)\right], \qquad r_\sigma^{\mathrm{act}} = r_\sigma^{\mathrm{full}} - r_\sigma^{\mathrm{fr}}.$$

For the centered finite-direction law, r_σ^j approaches the corresponding second-order mean-response coefficient. At finite σ it retains truncation and solver error. We subtract each solver profile's own reference endpoint; the resulting active vectors reproduce the original S0 normalized contrasts within $3.6 \cdot 10^{-13}$. The source contains all 60 combinations of three solver profiles, five direction sets, and four amplitudes.

Table 6: Response cancellation at $\sigma = 0.025$, $\tau = 0.05$. Each interval is the range over the five direction sets, not a confidence interval. All quantities use the specified log-ratio chart.

Profile	$\|r^{\mathrm{act}}\|/\|r^{\mathrm{full}}\|$	$\cos(r^{\mathrm{act}}, r^{\mathrm{full}})$	$\|r^{\mathrm{full}}\|/\|r^{\mathrm{fr}}\|$
Original BFGS	[0.908, 0.923]	[-0.99998, -0.99935]	[0.520, 0.524]
Dual/BFGS	[0.913, 0.927]	[-0.99998, -0.99936]	[0.519, 0.523]
Dual/L-BFGS-B	[0.913, 0.923]	[-0.99998, -0.99937]	[0.520, 0.523]

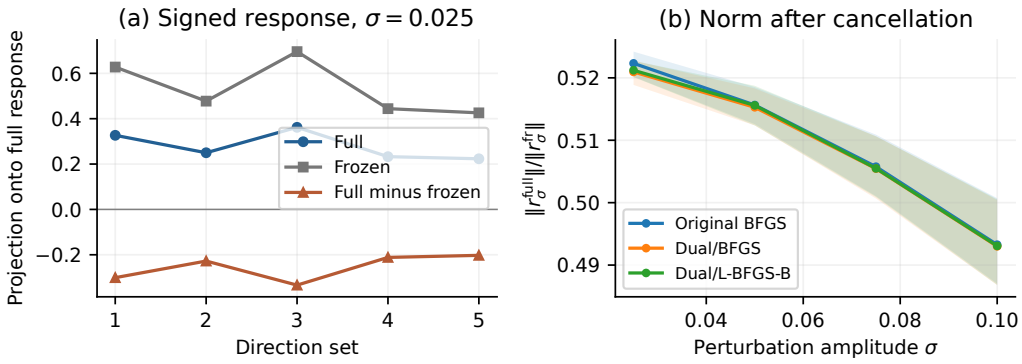

Figure 3: Plan response can offset the frozen response. Left: signed projections of the normalized responses onto the full-response unit vector, using the original solver at $\sigma = 0.025$. The active term points against the frozen term. Right: the full/frozen norm ratio across amplitudes; shading spans the five direction sets for each solver profile and is not statistical uncertainty.

At the smallest amplitude, the active/full norm ratio is 0.908–0.927, while their cosine is between -0.99998 and -0.99935. The full response is only 0.519–0.524 times the frozen-response norm. Thus the contribution is substantial in this example and partially cancels the remaining response. These norm ratios are not additive percentages of population bias: the terms are vectors, and their individual means depend on the endpoint chart. The result concerns this fixed geometry and finite perturbation law; it does not establish how often cancellation occurs in learned transport models.

E.6 Observable responses from the archived endpoints

A log-ratio coefficient is useful for computation, but readers may ask whether it changes a fixed property of the distribution. We therefore reanalyze all five-support endpoint records using three observables: the full probability vector $a(u)$, the declared energy $F(u) = \frac{1}{2}\|a(u) - a_{\text{target}}\|^2$, and Shannon entropy $\mathcal{H}(u) = -\sum_{i=1}^{5} a_i(u) \log a_i(u)$. This is a post-hoc analysis of existing solves, not an additional experiment on independently sampled data.

For a scalar observable φ, or componentwise for a, set $L_\varphi(g) = \varphi(\mathsf{R}_\tau(x, g)) - \varphi(\mathsf{R}_\tau^{\text{fr}}(x, g))$ and define

$$d_{\sigma,\varphi} = \frac{1}{2m} \sum_{j=1}^{m} \sum_{s\in\{-1,1\}} \{L_\varphi(g^\star + s\sigma h_j) - L_\varphi(g^\star)\}, \qquad \frac{d_{\sigma,\varphi}}{\sigma^2} \longrightarrow b_\varphi := D\varphi(u^\star) b_\tau^{\text{act}}(\Sigma_m). \tag{39}$$

The observable is evaluated before averaging. Applying it to an averaged endpoint would test a different quantity. The full and frozen branches share their first ground derivative, so the observable-Hessian terms cancel (Appendix B.5). With bounded fourth derivatives of the composite responses, the normalized truncation error is $O(\sigma^2)$; finite solver error is amplified by σ^{-2} as in Appendix B.9.

For the four-coordinate softmax chart, write $a^\star = a(u^\star)$ and $J_a = [\operatorname{diag}(a^\star) - a^\star (a^\star)^\top]_{:,1:4}$. The three predictions are

$$b_a = J_a b_\tau^{\text{act}}, \qquad b_F = (a^\star - a_{\text{target}})^\top b_a, \qquad b_{\mathcal{H}} = -(\log a^\star + \mathbf{1})^\top b_a.$$

The mass prediction sums to zero. The predictions are reconstructed from the S3 plan derivatives and outer Hessian; they are not fitted to observable contrasts. The same independent S3 prediction is compared with every solver profile. Each profile's own computed reference values are subtracted from its perturbed values, removing its numerical constant offset.

Table 7: Observable coefficient errors (%) in the five-support model. Each cell is the maximum across five direction sets and three solver profiles. Probability error is the relative Euclidean error of the full five-dimensional vector; scalar error is absolute error divided by the magnitude of the predicted coefficient. All denominators are nonzero. These are deterministic comparisons, not confidence intervals.

Observable	$\sigma = .025$	.05	.075	.1
Probability vector	1.96	7.55	18.06	35.75
Energy	1.50	5.44	12.59	23.39
Shannon entropy	1.44	5.00	11.47	21.04

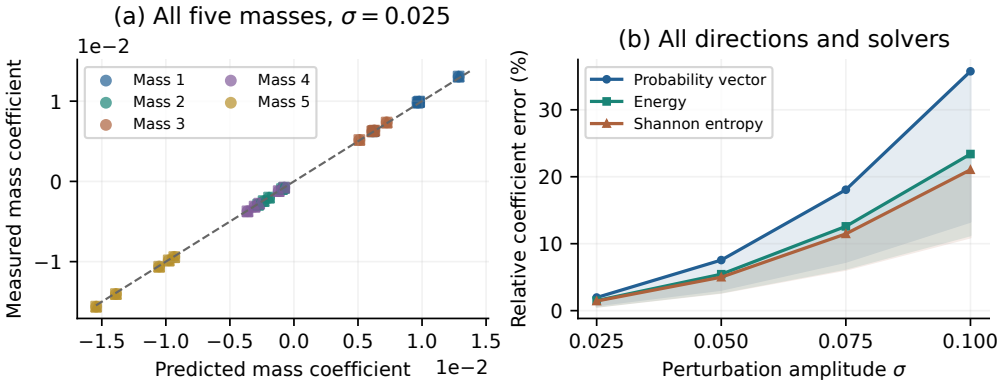

Figure 4: Observable checks on the same five-support geometry. Left: all 75 mass-component comparisons at $\sigma = .025$; color identifies the mass, while circles, squares, and triangles identify the original, dual/BFGS, and dual/L-BFGS-B profiles. Many markers overlap. The dashed line is equality. Right: lines show the worst error across the 15 direction/profile combinations; shading spans their minimum and maximum. It is not a sampling interval.

Across all four amplitudes, the analysis includes 180 observable comparisons (420 scalar components). At $\sigma = .025$, the largest relative errors are 1.96% for probability, 1.50% for energy, and 1.44% for entropy. At $\sigma = .1$, they reach 35.75%, 23.39%, and 21.04%, respectively. Observable agreement therefore also depends on the perturbation regime. Predicted and observed scalar signs agree throughout; mass-vector inner products are positive throughout. These checks connect the contrast to fixed distributional quantities, while retaining S3's limitations on support geometry, numerical accuracy, and statistical replication.

E.7 An additional Gaussian covariance correction control

The retained E4 summaries also contain a two-dimensional Gaussian covariance experiment, separate from the binary and real-feature tasks. It uses zero-mean Gaussian states, energy $F(C) = \mathrm{KL}(\mathcal{N}(0, C)\|\mathcal{N}(0, C_{\mathrm{target}}))$, and the Mahalanobis Gaussian Wasserstein penalty. The step is $\tau = 0.05$, with current covariance $\mathrm{diag}(1, 0.4)$, and target covariance $R(0.55)\,\mathrm{diag}(0.3, 1.4)R(0.55)^\top$, where $R(\alpha)$ is planar rotation by α radians. The reference metric is the trace-normalized matrix $R(-0.35)\,\mathrm{diag}(0.6, 1.4)R(-0.35)^\top$. For independent $z_i \sim \mathcal{N}(0, I_2)$, let $(a, b) = B^{-1}\sum_i z_i/\sqrt{2}$ and $H = \left(\begin{smallmatrix} a & b \\ b & -a \end{smallmatrix}\right)$. The batch metric is the trace-normalized exponential of $\log G^\star + 0.6H$. Full/half-batch correction is applied in the three-coordinate matrix-log covariance chart. This controlled estimator is distinct from the sample-covariance estimator in the feature tasks.

Each of 20 independent seeds contains eight antithetic pairs, at each $B \in \{16, 32, 64, 128, 256\}$. Write e_{sr} for a replicate's chart error, $\bar{e}_s$ for its seed mean, and $\bar{e}$ for the mean across seeds. The archive separately reports $\|\bar{e}\|$, the mean of $\|\bar{e}_s\|$, and the mean squared replicate error. The first is

labeled `population_mean_bias_norm` in the historical CSV, but it is a Monte Carlo estimate, not an exact population norm. Its sampling uncertainty must not be inferred from the interval for the second quantity. We use the retained aggregate CSV and checked aggregation source; per-replicate endpoints were not reconstructed for this additional display.

Table 8: Archived Gaussian covariance correction at $B = 64$. The three error columns are multiplied by 10^4. Norms of pooled and seedwise mean errors are distinct statistics; no interval for one is assigned to the other. Sample/solve counts are per replicate.

Method	$\|\bar{e}\|$	Mean $\|\bar{e}_s\|$	MSE	Samples/solves
Plug-in B	9.064	9.096	1.757	64/1
RC	0.336	4.028	1.743	64/3
Shared-batch ensemble	15.291	15.314	1.784	64/3
Plug-in $3B$	3.132	3.143	0.549	192/1
Independent ensemble	9.766	9.777	0.564	192/3

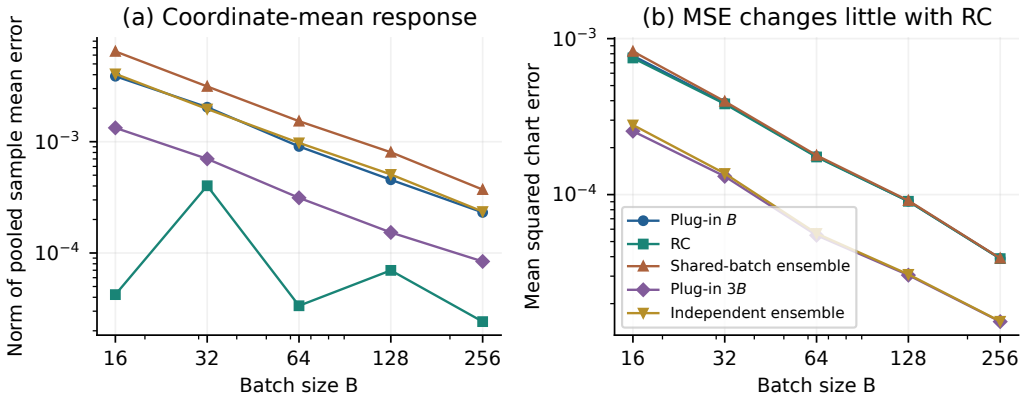

Figure 5: All five archived batch sizes in the Gaussian covariance control. RC reduces the observed pooled mean-error norm, but leaves MSE close to plug-in B. Its nonmonotone small residual does not establish a B^{-2} rate. Larger-batch methods use $3B$ observations. These curves are descriptive summaries of the retained runs.

At $B = 64$, RC's pooled mean-error norm is about 3.7% of plug-in's, whereas its MSE is about 99.2% of plug-in's. This supports the distinction between mean correction and overall error in a covariance model; it does not test the entropic full–frozen identity. Across all methods and batch sizes, the archive's optimizer flags report success, but the recorded gradient checks remain finite-precision diagnostics. The nonmonotone residual mean response and the difference between the two norm statistics preclude a claim of resolved second-order bias scaling.

E.8 Dynamics and correction diagnostics

Table 9 reports all eight recovered configurations. Each row uses 200 independent complete trajectories. The bootstrap resamples trajectory indices 10,000 times with RNG seed 20260912; slope resampling uses the same seed indices across the step-size grid. Signed component intervals, rather than intervals for a nonnegative norm, assess compatibility of the mean with its averaged reference.

Table 9: E3 from complete raw trajectories. Mean error is the norm of the terminal sample mean minus the computed averaged path. MCSE is $\sqrt{\mathrm{tr}(\widehat{\mathrm{Cov}})/200}$. Its ratio with a vector error is a descriptive diagnostic, not a universal significance test. Every row has 200 seeds.

Configuration	τ	Mean error	Variance trace	MCSE	Error/MCSE
Stationary	0.2	0.0123	0.00016	0.000894	13.71
Stationary	0.1	0.00259	6.07e-05	0.000551	4.70
Stationary	0.05	0.000476	2.33e-05	0.000341	1.40
Stationary	0.025	0.000187	1.27e-05	0.000252	0.74
Stationary	0.01	0.000129	5.57e-06	0.000167	0.77
Rotation 90°	0.05	9.34e-05	9.68e-05	0.000696	0.13
Anisotropy 8	0.05	0.00152	7.81e-05	0.000625	2.44
Moving mixture	0.05	0.000696	1.29e-05	0.000254	2.74

The E3 binary experiment fixes $B = 64$, $T = 1$ and tests $\tau \in \{0.2, 0.1, 0.05, 0.025, 0.01\}$ with 200 seeds. The variance log-slope is 1.017 with interval $[0.905, 1.133]$ over $\tau \leq 0.1$. Including $\tau = 0.2$ gives 1.116 with interval $[1.032, 1.203]$, which excludes one. The endpoint variance traces are approximately $1.60 \cdot 10^{-4}$, $6.07 \cdot 10^{-5}$, $2.33 \cdot 10^{-5}$, $1.27 \cdot 10^{-5}$, and $5.57 \cdot 10^{-6}$, respectively. Mean-to-averaged-path errors relative to their Monte Carlo standard errors are approximately 13.71, 4.70, 1.40, 0.74, and 0.77. This comparison uses a computed averaged path, which itself has discretization and integration error.

The rotation example has a corresponding ratio of approximately 0.13. In the stronger anisotropy case it is 2.44, with a signed component interval excluding zero; the moving-mixture case has ratio 2.74 and nonzero signed discrepancies. These findings limit finite-step agreement. Distinguishing discretization, nonuniform response, and reference-integration error would require additional diagnostics; this revision does not assert that such diagnostics were run.

E4 reports both mean and variability. In the binary $B = 64$ example, the plug-in signed chart-mean bias interval is approximately $[2.64, 3.27] \cdot 10^{-4}$, and the corrected interval is approximately $[-4.45, 5.44] \cdot 10^{-5}$, with point estimate $6.68 \cdot 10^{-6}$. These percentile bootstrap intervals need not be centered at the point estimates. In a recursive binary comparison at $B = 64$, reported MSE values are approximately $1.49 \cdot 10^{-4}$ for plug-in, $1.29 \cdot 10^{-4}$ for RC, $4.8 \cdot 10^{-5}$ for EMA, and $4.7 \cdot 10^{-5}$ for plug-in $3B$. Thus a successful mean correction can coexist with stronger variance-reducing baselines.

The observed corrected-bias slopes do not establish B^{-2} convergence. Taking absolute values of noisy mean estimates creates a Monte Carlo floor; higher-order asymptotics also require stronger assumptions than the leading expansion. The paper reports cancellation evidence separately from the conditional higher-order theorem.

Cross-fitting is a different intervention. In the joint-noise binary control at $B = 64$, same-batch estimation has signed chart bias $1.91 \cdot 10^{-4}$. An independent cross-fit using $2B$ observations has bias $2.69 \cdot 10^{-4}$, while splitting the original B observations gives $4.69 \cdot 10^{-4}$. Joint response correction gives about $8 \cdot 10^{-6}$. Thus eliminating dependence between energy and geometry does not necessarily reduce total endpoint bias: it removes mixed terms, while changing sample size can increase the remaining terms. These results rule out an interpretation of this paper as a universal guarantee for lagging or cross-fitting a metric.

E.9 Real-data feature tracking

Both datasets use their standard training/test separation. The supplied asset reports allocate 35,000 training images to the encoder, 5,000 to validation, 10,000 to the reference bank, and 10,000 to metric sampling. The encoder is trained at the unrotated domain and then frozen. Its reported validation accuracy is 98.98% on MNIST and 89.46% on Fashion-MNIST. The fixed PCA projection retains approximately 94.15% and 97.08% of feature variance, respectively. All methods in a dataset share the same assets.

The schedule has 40 stored states, indexed 0 through 39, with angles linearly spaced from 0° to 90° and 39 updates at $\tau = 0.05$. There are ten class-conditional Gaussian states, $\mu_{k,c} = \mathcal{N}(m_{k,c}, C_{k,c})$,

and angle-specific target Gaussians $\nu_{k,c}$ from the reference bank. Write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{10})$ for the labeled state. The declared energy is

$$F_k(\boldsymbol{\mu}) = \frac{1}{10} \sum_{c=1}^{10} \mathrm{KL}\big(\mathcal{N}(m_c, C_c) \| \mathcal{N}(m^\star_{k,c}, C^\star_{k,c})\big) .$$

The class-preserving transport penalty is

$$\mathcal{D}_G(\boldsymbol{\mu}, \boldsymbol{\eta}) = \frac{1}{10} \sum_{c=1}^{10} W_G^2(\mu_c, \eta_c).$$

Both energy and penalty use the same factor $1/10$, so the proximal objective separates into ten classwise solves at the declared τ. This is transport on a labeled product model, not Wasserstein distance between the unlabeled Gaussian mixtures. All geometry estimators in the final unclipped run use the covariance divisor $B - 1$, a fixed ridge $\rho = 1$, and trace normalization. The covariance is estimated from the separate angle-specific metric bank. The large reference geometry uses the reference bank with the same ridge rule and its corresponding sample-size divisor. Trajectory seed indices are 0–9 for MNIST and 0–4 for Fashion-MNIST.

For a positive definite Mahalanobis matrix G, the Gaussian transport penalty can be evaluated using transformed means $\widetilde{m} = G^{1/2} m$ and covariances $\widetilde{C} = G^{1/2} C G^{1/2}$:

$$W_G^2(\mathcal{N}(m, C), \mathcal{N}(n, D)) = \|\widetilde{m} - \widetilde{n}\|^2 + \mathrm{tr}\Big(\widetilde{C} + \widetilde{D} - 2(\widetilde{C}^{1/2} \widetilde{D} \widetilde{C}^{1/2})^{1/2}\Big) .$$

This closed-form Gaussian transport quantity is distinct from the recorded tracking metric. The checkpoints store a 152-dimensional chart per class: 16 mean coordinates and the 136 upper-triangular entries of $\log C$. Equation (26) uses ordinary Euclidean norm on that vector, with no inferred off-diagonal $\sqrt{2}$ weighting. Reconstruction from the stored float32 charts agrees with the tracking arrays to numerical precision; this validates the recorded chart quantity, not the accuracy of the original transport solves.

RC extrapolates in this chart, so exponentiating the corrected covariance coordinate preserves positive definiteness. It can nevertheless change prediction and need not preserve an energy descent inequality. The final fixed-ridge runs supersede earlier clipped runs for the paper's real-task claims. We do not pool these different estimator definitions into a single confidence interval.

E.10 All real-task baselines and resource reporting

Table 10: Terminal prediction at 90° and accumulated solve time at $B = 64$. Lower NLL and higher accuracy are better. Costs exclude encoder training and evaluation; sample/solve counts are per update.

Dataset	Method	NLL	Acc. (%)	Solve-path sec.	Samples/solves
MNIST	Plug-in B	62.75	40.51	0.17	64/1
MNIST	RC	67.04	33.55	0.69	64/3
MNIST	Plug-in $3B$	64.75	36.85	0.17	192/1
MNIST	Reference	65.78	35.45	0.17	0/1
Fashion	Plug-in B	49.34	47.81	0.17	64/1
Fashion	RC	51.21	43.64	0.69	64/3
Fashion	Plug-in $3B$	50.41	45.01	0.17	192/1
Fashion	Reference	50.98	43.56	0.17	0/1

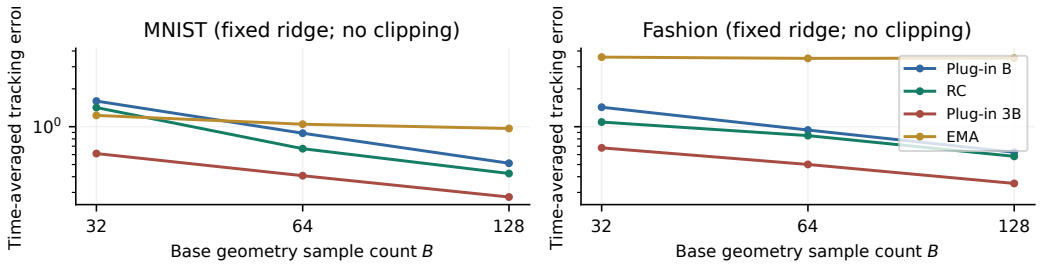

Figure 6: Tracking comparisons on a shared logarithmic vertical scale. RC improves over plug-in B in both datasets, but plug-in $3B$ is stronger throughout. EMA is competitive in part of the grid. The curves compare sample/solve trade-offs and are not equal-wall-clock frontiers. Prediction is reported separately in Table 10.

Table 11: All recorded methods: mean TE at the three batch settings. Labels follow the supplied run configuration. The reference row is zero by definition of TE.

Dataset	Method	$B = 32$	$B = 64$	$B = 128$
MNIST	Plug-in B	1.599	0.888	0.512
MNIST	Cross-fit	1.587	0.841	0.518
MNIST	Lagged	1.611	0.922	0.568
MNIST	EMA	1.230	1.047	0.968
MNIST	Shared-batch ensemble	2.686	1.291	0.654
MNIST	Plug-in $3B$	0.612	0.407	0.276
MNIST	Independent ensemble	1.416	0.709	0.388
MNIST	RC	1.420	0.668	0.424
MNIST	Reference	0.000	0.000	0.000
Fashion	Plug-in B	1.428	0.941	0.627
Fashion	Cross-fit	1.321	0.900	0.634
Fashion	Lagged	1.869	1.352	1.278
Fashion	EMA	3.578	3.497	3.512
Fashion	Shared-batch ensemble	1.844	1.116	0.704
Fashion	Plug-in $3B$	0.679	0.501	0.354
Fashion	Independent ensemble	1.065	0.659	0.416
Fashion	RC	1.089	0.849	0.581
Fashion	Reference	0.000	0.000	0.000

The shared-batch ensemble uses the same full/half-batch geometry construction as RC, but averages its three endpoint charts with weights $(1/3, 1/3, 1/3)$ instead of $(2, -1/2, -1/2)$. Both methods use B distinct observations and three solves per update, each along its own recursive path. This control performs worse than RC in all six TE settings; it compares the weighting rules, not merely the number of solves. The method labeled independent ensemble averages three size-B disjoint batches and uses $3B$ observations. Plug-in $3B$ pools them into one geometry and uses one solve.

At each update, the implementation samples $3B$ indices without replacement from the finite metric bank and partitions them into three groups; the primary group supplies the RC halves. Fresh random indices are drawn at each step; there is no across-step depletion of the bank. Conditional on the fixed feature assets, this refreshment differs from within-step independence: disjoint groups are not independent conditional on the finite bank. EMA and lagged baselines further retain historical information, so Theorem 5 does not apply to them as stated. Thus these tasks illustrate the correction's behavior rather than verify the i.i.d. jackknife asymptotics exactly. The baseline labeled cross-fit uses a separate geometry group; the target energy is fixed from the reference bank for all methods, so this comparison does not isolate shared energy–geometry estimation bias. Sample and solve counts do not establish equal wall-clock cost.

Table 10 reports the recorded accumulated `method_wall_time` over the 39 updates at $B = 64$. These are solve-path measurements, excluding encoder training, feature extraction, and held-out

evaluation. A different field, `wall_time_sec`, records the whole multi-method experiment; it must not be assigned as the cost of each method. No claim of overall runtime superiority or a fully optimized compute frontier is made.

Table 12: Paired terminal prediction differences, RC minus plug-in B. Positive NLL differences and negative accuracy differences indicate degradation. Accuracy is in percentage points. All intervals resample complete paired seeds.

Dataset	B	Metric	Difference	95% CI
MNIST	32	NLL	12.351	[11.812, 12.913]
MNIST	32	Accuracy (pp)	-18.941	[-19.385, -18.507]
MNIST	64	NLL	4.287	[4.130, 4.459]
MNIST	64	Accuracy (pp)	-6.966	[-7.133, -6.806]
MNIST	128	NLL	1.801	[1.754, 1.851]
MNIST	128	Accuracy (pp)	-2.833	[-2.927, -2.742]
Fashion	32	NLL	3.183	[2.969, 3.385]
Fashion	32	Accuracy (pp)	-7.048	[-7.742, -6.380]
Fashion	64	NLL	1.871	[1.846, 1.900]
Fashion	64	Accuracy (pp)	-4.172	[-4.308, -4.038]
Fashion	128	NLL	0.972	[0.908, 1.029]
Fashion	128	Accuracy (pp)	-2.142	[-2.266, -2.032]

The terminal Gaussian-classifier accuracy is not the unrotated encoder's validation accuracy. It measures the tracked class models at 90°, where substantial domain shift remains. Better reference fidelity and worse classification coexist in the recorded runs. Calling the reference a prediction oracle, or describing all task changes as negligible, would misstate these data.

E.11 Scope of the present result

The theory is finite-dimensional and local. It does not construct random metric-measure spaces with changing topology, a random EVI theory, or an intrinsic Itô geometry on unrestricted Wasserstein space. The stochasticity comes from estimated ground geometry; the limiting diffusion is for rescaled fluctuations under independent refreshment. The real-data experiment is a controlled feature-distribution problem, not end-to-end representation learning or image generation.

The two-support result is analytic; the strongest empirical transport-specific evidence is the audited five-support case. Real-feature Gaussian tasks test fidelity, not the entropic identity. S3 and S4 quantitatively support the local endpoint and velocity coefficients in this one support geometry. Higher-order corrected rates and hard-regime averaging remain unresolved. The analysis package reproduces the stated aggregations and intermediate-array checks. Supplementary source and encoder assets support rerunning the experiments, subject to the S2 source-version limitation above and regeneration of the large feature banks. Full end-to-end reruns were not performed during manuscript assembly.

F Notation and Assumptions

F.1 Quantities used in the theory and experiments

The distinction between an endpoint coefficient and a velocity coefficient is essential when the step size varies. The endpoint covariance coefficient $\mathcal{V}_\tau$ and the local transport tensor $A_g(x)$ are different objects; H is the Hessian of the outer proximal objective.

Table 13: Principal notation. All coefficients refer to the stated state and ground coordinates. State dimension is p and ground-parameter dimension is d_g.

Symbol	Meaning
$g_t^\star = \mathcal{G}(P_t)$	Population ground parameter for the specified distribution and estimator functional.
g_t^{ref}	Estimate from the finite reference bank; the real-task reference path uses this parameter.
$\widehat{g}_{t,B}$	Geometry estimate based on B observations.
$C(P)$	Covariance of the feature distribution; distinct from the estimator coefficient Σ.
β, Σ	Leading estimator-mean and second-moment coefficients in (5).
$\mathsf{R}_\tau, \mathsf{R}_\tau^{\mathrm{fr}}$	Full update and candidate-dependent frozen-plan update.
H	Hessian of the full outer objective at the common reference endpoint.
$b_\tau, \mathcal{V}_\tau$	Coefficients of endpoint mean bias and covariance in (8).
b_τ^{act}	Difference of the full and frozen endpoint mean-bias coefficients.
$A_g(x), A_g^{\mathrm{fr}}(x)$	Symmetric $p \times p$ local tensors: half the first-endpoint Hessian, evaluated at $u = x$ with the second endpoint fixed.
$Q_\Sigma(x)$	Symmetric $p \times p$ matrix $\mathrm{tr}_\Sigma\, D_g^2(A_g - A_g^{\mathrm{fr}})(x)\vert_{g^\star}$ in Theorem 3.
$\mathcal{L}_\Sigma(u,x)$	Scalar contraction $\mathrm{tr}_\Sigma\, D_g^2(D_g - D_g^{\mathrm{fr}})(u,x)\vert_{g^\star}$; $\mathcal{L}_m = \mathcal{L}_{\Sigma_m}$.
K	S4 matrix $\nabla_{uu}^2 \mathcal{L}_m(u,x)\vert_{u=x} = 2Q_{\Sigma_m}(x)$; the second endpoint is fixed.
b_v^{act}	Coupling contribution to the velocity bias; $b_\tau^{\mathrm{act}} = \tau b_v^{\mathrm{act}} + O(\tau^2)$.
Σ_m	Covariance of the actual finite perturbation directions in (32).
b_{pred}	Independent numerical evaluation of $b_\tau^{\mathrm{act}}(\Sigma_m)$ in S3.
d_σ	Centered antithetic full-minus-frozen endpoint contrast; its normalized form is d_σ/σ^2.
$z_{s,j}, P_s$	Vectorized plan-direction derivative in $\mathbb{R}^{n^2}$ and endpoint Jacobian in $\mathbb{R}^{n^2 \times p}$, respectively; defined before (38).
w_1, w_2, w_3	Cross/self weights $1, -1/2, -1/2$; s indexes transport terms and j perturbation directions.
TE	Mean trajectory error in Gaussian state coordinates, defined in (26).

F.2 What the assumptions provide

Local response. Assumption 1 makes the selected minimizer differentiable and controls the Taylor remainder in expectation. It requires local stability and estimator moments, rather than global convexity of the energy. A positive Hessian at a computed solution is useful diagnostic information, but it does not establish a uniform lower bound on a neighborhood or rule out other branches. Tail localization must be justified by the sampling model.

Coupling identity. Theorem 2 compares objectives that agree at the reference geometry for every candidate marginal pair. The construction, the three entropic terms, and the choice of ground coordinates are therefore part of the model. Value and derivative checks at several candidates can expose implementation errors; they do not prove an identity over the whole neighborhood. In an affine cost chart, the frozen second ground derivative vanishes. In a nonlinear chart, the corresponding formula uses $D_g^2(S - S^{\mathrm{fr}})$.

Small-step limit. Theorem 3 needs uniformly positive local tensors and mixed derivative bounds to control a remainder in $C^2(g)$. Positive finite marginals and fixed $\varepsilon > 0$ give smooth entropic plans locally. Uniform quantitative bounds still depend on how far the state and parameter sets remain from degeneracy. Zero-energy identity, eigenvalue diagnostics, and derivative sweeps are checks of the implementation and tested points; finite sweeps cannot establish the uniform theorem hypotheses.

Averaging and fluctuations. Assumption 2 specifies independent refreshment, finite-horizon stability, regularity, and moment bounds. Independent draws can be guaranteed by the sampling design. The remaining conditions concern the field on the relevant state region. Corollary 8 additionally requires the finer accumulated solver-error scale. A variance slope close to one supports one consequence of this result, without establishing process convergence. Uniform $(2+\delta)$ moment control can be used to verify the required Lindeberg bound in the standard martingale argument.

Correction. Theorem 6 separates three conclusions. A first-order mean expansion gives leading-bias cancellation; a second-order mean expansion gives the B^{-2} rate; an asymptotically linear representation gives equality of the leading covariance limits. The first conclusion does not require either of the latter two conditions. Finite-batch numerical agreement is assessed separately from these asymptotic assumptions.

F.3 Which assumptions are checked by which records?

Table 14 separates sampling properties fixed by design from numerical diagnostics and conditions that remain unverified. An unverified condition limits the application of a theorem; it does not contradict a conditional theorem. Conversely, an empirically plausible conclusion does not establish all of that theorem's hypotheses.

Table 14: Assumption coverage and remaining gaps. A finite diagnostic is not a certificate of a uniform bound. All real-feature statements are conditional on the shared assets and reference bank.

Result	Requirement and available check	Remaining boundary
Endpoint response (Theorem 1)	Stable interior branch, smoothness, moments and tail control. Gaussian estimator moments are proved in Appendix A.2; S2 checks selected Hessians and marginals.	A fixed ridge does not prove outer stability. Finite Hessian checks do not establish a uniform bound or endpoint tail control.
Plan response (Theorems 2–3)	Candidate-dependent freezing and common first derivatives; positive local tensors and uniform mixed derivatives for the small-step limit. S2–S4 check definitions and coefficients numerically.	Direction sets share one five-support geometry. Fixed-ε smoothness does not give uniform control near degenerate marginals or as $\varepsilon \to 0$.
Averaging (Theorem 5)	Fresh geometry draws independent across steps, an invariant region, uniform field bounds and accumulated solver accuracy. E3 uses independent refreshment.	Harder E3 cases retain finite-step mean discrepancies. Reused geometry and EMA/lagged estimates fall outside the stated refreshment model.
Fluctuation limit (Corollary 8)	Martingale covariance and Lindeberg conditions plus $e_\tau = o(\sqrt{\tau})$. E3 supports a small-step variance slope near one.	A terminal variance slope alone does not verify a diffusion-process limit or the required solver-error scale.
Response correction (Theorem 6)	Compatible mean expansions at B and $B/2$; a stronger expansion for B^{-2} and shared influence for equal leading covariance. E4 measures mean response and MSE separately.	The corrected B^{-2} rate is unresolved. Independent-half sampling is not an exact description of disjoint samples drawn without replacement from a fixed finite bank.
Feature tracking (E5–E6)	At each step, fresh random indices are drawn conditional on the fixed angle-specific bank. All methods share reference, encoder and PCA assets.	Within-step groups drawn without replacement are dependent. TE measures finite-bank chart tracking, not population accuracy; no entropic-coefficient or end-to-end training claim follows.

Within-step dependence and temporal dependence are different. Fresh independent index draws at successive steps can satisfy temporal independence conditional on the bank, even though disjoint groups within one draw are dependent. This sampling property alone does not establish the other uniform assumptions of Theorem 5. Carrying an estimate or an EMA across steps creates temporal dependence; Appendix C.4 gives an exact example where reusing one estimate leaves $O(B^{-1})$ trajectory variance as $\tau \to 0$.

The retained S2 source-history gap and the absence of an end-to-end rerun concern provenance, not additional theorem assumptions. They are disclosed in Appendix E.4 and the reproducibility statement. Numerical reanalysis checks the supplied observations and formula implementations; it cannot recover an unpreserved source state or replace new experiments in a broader model class.

F.4 READING THE NUMERICAL EVIDENCE

We use *descriptive fit* for an observed scaling over tested amplitudes or step sizes, *consistency check* for agreement across numerical methods or precision settings, and *coefficient comparison* for a prediction computed independently of the endpoint contrasts it is compared with. A *certified error bound* additionally needs justified bounds on the exact gradient error and local strong convexity, as in Appendix B.9. The five-point records now provide independent endpoint and velocity coefficient comparisons (S3/S4), supported by consistency checks. Neither the derivative sweeps nor cross-solver agreement provide a certified endpoint-error bound.

A mismatch in that comparison has several possible sources: a different frozen objective, a coordinate convention, finite-amplitude truncation, or numerical error. A discrepancy that persists after those sources are controlled under the theorem's assumptions requires revisiting the derivation or implementation. It cannot be resolved by selecting favorable directions or by changing the empirical claim alone.