REACHABILITY IS NOT ENOUGH: SELECTION FAILURE AND REPAIR IN CONDITIONAL FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional flow matching (CFM) learns a velocity field by regression, but success is judged by the distribution it generates. If a model class contains a correct generator, must population CFM select one? It need not. We construct a finite-dimensional affine class for independent straight-line CFM with Gaussian endpoints in which every admissible positive-definite matrix weight measurable from the regression input excludes the correct generator. For identifiable affine classes, we characterize when reweighting can select a prescribed field under specified information access and derive the exact minimum global spectral contrast required. Near this obstruction, even the best attainable asymptotic variance for the weighted-regression ratio estimator studied here diverges, whereas an endpoint-derived moment equation using the same observations remains regular even at the obstruction. Controlled finite-update experiments show CFM moving correct synthetic and pretrained generators away from their targets. In a structured family, CFM attains lower regression risk than likelihood fitting yet much worse endpoint KL; in a richer version, endpoint fitting substantially reduces CFM's KL under a prespecified 10% empirical regression-risk budget.

1 INTRODUCTION

A generative model can represent the right distribution yet be trained to produce the wrong one. Conditional flow matching (CFM) fits velocities along an interpolation, while generation is judged by the endpoint. A restricted class may contain correct generators without representing the interpolation's marginal velocity.

CFM's conditional-to-marginal regression identity remains valid (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Tong et al., 2024), but its best reference-velocity approximation need not produce the correct endpoint. More accurate regression can therefore move a model away from its generative purpose.

We ask whether population regression can exclude every endpoint-correct member of a class, even after positive reweighting, and what can escape that exclusion.

By *selection* we mean population-risk minimization within a fixed class. A *selecting weight* makes a prescribed correct field the population minimizer. In our constructed classes, regression excludes every correct member, which velocity misspecification alone would not imply.

We establish three results:

1. **Whole-set selection failure.** In a smooth Gaussian squeeze family, every admissible positive-definite matrix weight measurable from time and state excludes the unique correct generator and incurs a common positive endpoint-error floor (Theorem 1). A second construction excludes infinitely many correct generators under its specified weight families.
2. **Exact feasibility and global contrast.** For identifiable affine parameter classes, Theorem 2 gives the minimum global spectral contrast under a specified information restriction. A transverse tilt crosses the squeeze obstruction and reveals the cost of restricting weight information.
3. **Estimator-specific statistical cost.** In the squeeze family, an endpoint-derived moment equation remains regular using the same tuples, while positive-weight regression estimators

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.

have diverging asymptotic variance as the tilt vanishes and no positive selecting weight exists at zero tilt (Theorems 3–4). Importance sampling refines the regression estimator's cost; it does not change its population selection rule.

Reference-velocity approximation and endpoint realizability. Velocity-error bounds transfer reference-field approximation to endpoint control (Benton et al., 2023). Consistency analyses in growing classes (Kumar et al., 2026) control that approximation, which is stronger than containing some correct generator. We preserve the population loss identities of generator matching and weighted flow matching (Holderrieth et al., 2025; Billera et al., 2025) and study their restricted-class minimizers. Han et al. (2026) characterize when CFM quantities equal pointwise endpoint negative log-likelihood; we instead ask whether regression selects an endpoint-correct member of a restricted class, including the limits of positive input-measurable reweighting.

2 Reaching the target is not the regression task

Let $X_0 \sim p_0$ and $Y \sim p_1$ have finite second moments and a specified coupling. Time T is uniform and independent of the endpoints unless stated otherwise. Straight-line CFM observes the tuple (T, X_T, U), where

$$X_t = (1-t)X_0 + tY, \qquad U = Y - X_0, \qquad u(t,x) = \mathbb{E}[U \mid X_t = x]. \tag{1}$$

Write $\mu = \mathrm{Law}(T, X_T)$ and q_1^v for the endpoint law of the well-posed ODE $\dot{Z}_t = v(t, Z_t)$, $Z_0 \sim p_0$. A fixed finite sampler $S_{v,N}$ instead generates $q_{1,N}^v = (S_{v,N})_{\#}p_0$. A field is *endpoint-correct* if the specified generator produces p_1; a class is *endpoint-realizable* if it contains such a field, so the target is reachable within the class. Conditional square completion gives

$$R(v) = \mathbb{E}\,\|v(T, X_T) - U\|^2 = \mathbb{E}\,\|U - u(T, X_T)\|^2 + \|v - u\|^2_{L^2(\mu)}\,. \tag{2}$$

The first term is independent of v. CFM selects a best $L^2(\mu)$ approximation to u within the class, whereas endpoint correctness concerns $\{v : q_1^v = p_1\}$. Recovering u transports the reference marginals when its flow is well posed.

Weights and sampling. We allow weights W fixed before fitting, symmetric positive definite and measurable from $\mathcal{H} \subseteq \sigma(T, X_T)$. They cannot use individual endpoint-label information beyond that input. *State-dependent* permits both time and state; *time-only* means $\mathcal{H} = \sigma(T)$; scalar weights have form wI. Admissibility requires finite weighted risk and, for an affine class $v_\theta = v_* + F\theta$, a finite positive-definite Gram $\mathbb{E}[F^T W F]$. Sampling time with a positive loss density w, without correction, weights the uniform-time objective by w. Sampling instead from a positive proposal π and using w/π preserves that objective, with time independent of the endpoints and the required moments finite.

A Gaussian example and scale saturation. For independent $N(0,1)$ and $N(0,\sigma^2)$ endpoints, $\sigma > 0$, and $v_\theta = \theta x$, put $D_\sigma(t) = (1-t)^2 + \sigma^2 t^2$. Since $\mathbb{E}X_t^2 = D_\sigma(t)$ and $\mathbb{E}[X_t U] = D_\sigma'(t)/2$,

$$\theta_{\mathrm{CFM}} = \frac{3(\sigma^2-1)}{2(\sigma^2+1)}, \qquad \theta_c = \log\sigma. \tag{3}$$

The second parameter generates the correct exact-ODE endpoint at the fixed horizon $t = 1$. Thus $\sigma^2 = 3$ gives $\theta_{\mathrm{CFM}} = 3/4$ and variance $e^{3/2}$ despite the available correct value $\frac{1}{2}\log 3$. Weighting by D_σ^{-1} selects θ_c: its normal-equation denominator is one and its numerator is $\frac{1}{2}[\log D_\sigma]_0^1 = \log\sigma$.

The reference log-scale rate is $D_\sigma'/2D_\sigma$. The endpoint uses its time integral, whereas regression averages it with weight D_σ. Consequently $\theta_{\mathrm{CFM}} = \frac{3}{2}\tanh(\log\sigma)$ saturates for large rescalings. Inverse-spread weighting removes this distortion.

For every target, $|\theta_{\mathrm{CFM}}| < 3/2$. Its exact-ODE standard deviation therefore lies in $(e^{-3/2}, e^{3/2})$; under Euler with $N \geq 2$ steps it lies in $((1-3/(2N))^N, (1+3/(2N))^N)$. The Euler-correct constant on the branch $1 + \theta/N > 0$ is instead $N(\sigma^{1/N} - 1)$, attaining the finite-step drift bound of Claustre et al. (2026, Theorem 1). Correcting exact-ODE selection need not correct numerical deployment.

3 WHY EVERY POSITIVE WEIGHT CAN FAIL

A positive weight can change the lengths and angles used by regression, but it cannot reverse a strictly positive quadratic pairing. Suppose a correct field v_* has residual $e = v_* - u$, and its only fitted direction is $F = k(t)e$ with $k(t) > 0$. At v_*, every positive matrix weight satisfies

$$\mathbb{E}[F^T W e] = \mathbb{E}[k(t) e^T W e] > 0 \tag{4}$$

whenever the residual is nonzero on positive measure. The normal equation cannot hold. The squeeze family below also excludes every other correct parameter.

To make this obstruction concrete, we use independent $N(0, \mathrm{I}_2)$ endpoints and build a correct loop whose fitted direction aligns with its regression residual. Set

$$D = (1-t)^2 + t^2, \quad a = (2t-1)/D, \quad f = t(1-t), \quad k = t + \tfrac{1}{2}, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{5}$$

The reference velocity is $u(t,x) = a(t)x$; here $D = D_1$ from the Gaussian example. Define

$$\begin{gathered} M(t) = e^{fK_0}, \quad K_0 = \mathrm{diag}(1,-1), \quad C_0 = (1-2t)K_0, \quad E = C_0 - a\mathrm{I}, \\ H_\varepsilon = k(\mathrm{I} + \varepsilon J)E, \qquad v_{\theta,\varepsilon}(t,x) = (C_0 + \theta H_\varepsilon)x. \end{gathered} \tag{6}$$

Here $\varepsilon \in [-1,1]$ fixes the class and only $\theta \in \mathbb{R}$ is fitted. Since $M(0) = M(1) = \mathrm{I}$, $v_{0,\varepsilon}$ follows a determinant-one loop and generates the correct endpoint. It squeezes one axis and stretches the other, then reverses both; the reference flow instead shrinks and regrows the cloud radially (Figure 1). At zero tilt the fitted direction $F = \partial_\theta v_{\theta,0} = ke$ is positively collinear with the residual. The tilt breaks this collinearity without changing the correct member.

Liouville's formula gives the terminal covariance determinant:

$$\det \mathrm{Cov}(q_1^{\theta,\varepsilon}) = e^{-4b\theta}, \qquad b = 1 - \pi/4 > 0. \tag{7}$$

Thus $\theta = 0$ is the unique endpoint-correct parameter. The chosen gain $k = t + 1/2$ gives $\int ka = b > 0$, so every nonzero parameter changes terminal volume.

Theorem 1 (An endpoint-realizable class with no positive-weight repair)**.** *At $\varepsilon = 0$ in (6), every fixed positive-definite matrix weight measurable from (T, X_T), with finite weighted risk and a finite positive regression Gram, selects $-2 < \theta_W < -2/3$. Consequently,*

$$W_2^2(q_1^{\theta_W,0}, N(0,\mathrm{I}_2)) > 2\left[e^{2(1-\pi/4)/3} - 1\right]^2 > 0. \tag{8}$$

No uniform condition-number or spectral bound is imposed on these weights.

Proof mechanism. Weighted regression is a scalar quadratic, with minimizer $\theta_W = -\mathbb{E}[ke^T We]/\mathbb{E}[k^2 e^T We]$. Where the residual is nonzero, each observation would cancel it at $-1/k(t)$. Positive weights only change how these parameter preferences are averaged, so their population minimizer remains strictly between -2 and $-2/3$; zero is outside their range. Equation (7) excludes the target throughout that interval, and a second-moment Wasserstein bound gives (8). Appendix B supplies the full calculation. The obstruction persists even with commuting coefficient matrices.

Excluding an infinite correct set. Appendix E constructs infinitely many correct generators. With independent radial endpoints, every admissible time-only positive-definite matrix weight or state-dependent positive scalar weight excludes all of them; full state-dependent matrices lie outside this claim. Uniform-time CFM fails for every endpoint coupling. Under the continuity and nonvanishing conditions in Appendix E.4, each correct generator becomes selectable by a suitable rotational coupling and positive loss density, chosen for that generator.

4 WHEN REWEIGHTING WORKS, AND WHAT IT COSTS

We now ask a fixed-candidate question: given a prescribed field v_*, what information and spectral contrast must a positive metric have to select it? Fix $v_\theta = v_* + F\theta$, where $F(z) \in \mathbb{R}^{d\times m}$, $e = v_* - u$,

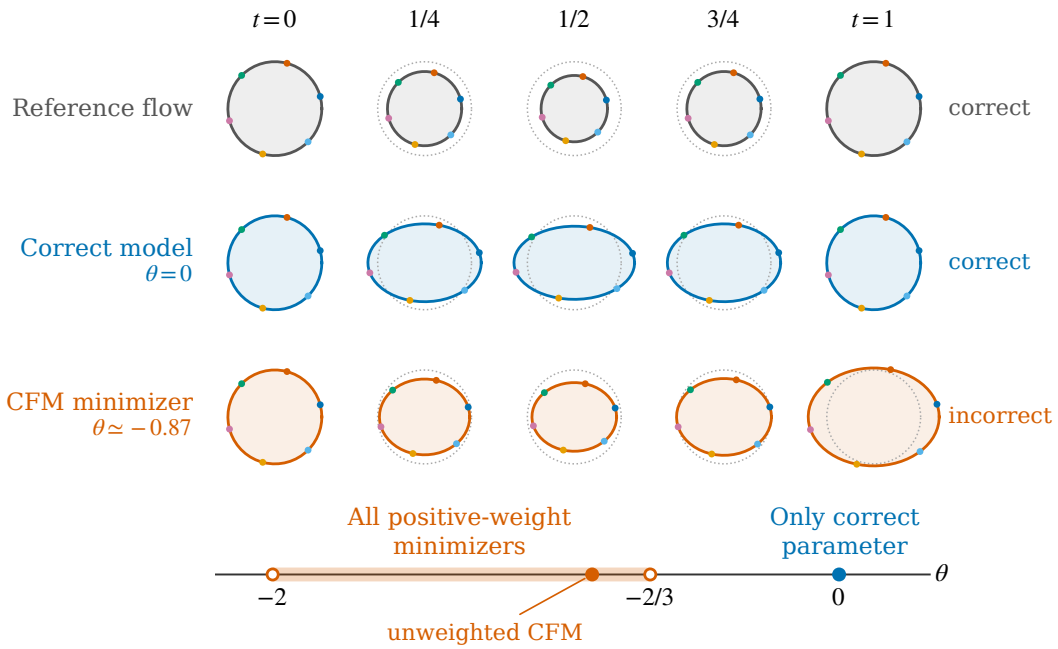


Figure 1: **The correct generator is available, but positive-weight regression cannot select it.** In the zero-tilt squeeze family, the reference flow and the correct model take different routes to the same standard Gaussian. Unweighted CFM ends elsewhere. Ellipses are images of the source unit circle; colored markers follow fixed contour points through each flow, not independently coupled interpolation paths. On the parameter line, every admissible positive-definite matrix weight measurable from time and state selects inside $(-2, -2/3)$, while zero is the only correct value.

$e, F \in L^2(\mu)$, and $\mathbb{E}F^TF \succ 0$. The class is affine in its parameters; features may be nonlinear in state. An admissible $\mathcal{H}$-measurable weight selects v_* exactly when its unique weighted minimizer is zero:

$$\mathbb{E}[F^TWe] = 0. \tag{9}$$

Write $b_{\mathrm{reg}} = \mathbb{E}F^Te$. For a parameter direction c, define

$$S_c = \mathbb{E}[\mathrm{sym}\{e(Fc)^T\} \mid \mathcal{H}], \qquad q(c) = \mathbb{E}\,\|S_c\|_*\,, \qquad \rho_{\mathcal{H}} = \sup_{q(c)>0} \frac{|c^Tb_{\mathrm{reg}}|}{q(c)}. \tag{10}$$

Here $\mathrm{sym}\,A = (A + A^T)/2$, $\|\cdot\|_*$ is the nuclear norm, and $\rho_{\mathcal{H}} = 0$ when $q \equiv 0$. For a fixed candidate the quantities are identified from CFM tuples: with $r_* = v_*(T, X_T) - U$, $b_{\mathrm{reg}} = \mathbb{E}[F^Tr_*]$ and $S_c = \mathbb{E}[\mathrm{sym}\{r_*(Fc)^T\} \mid \mathcal{H}]$, so no separate estimate of u is needed (Appendix A). The ratio compares the regression mismatch to the largest allowed change in its directional pairing; it measures proximity to a sign obstruction and lies in $[0, 1]$.

Theorem 2 (Minimum global contrast for a selecting weight)**.** *Among $\mathcal{H}$-measurable selecting weights with common spectral bounds $m_0\mathrm{I} \preceq W \preceq M_0\mathrm{I}$, $m_0 > 0$, the least global spectral contrast is*

$$\kappa_* := \inf M_0/m_0 = \frac{1 + \rho_{\mathcal{H}}}{1 - \rho_{\mathcal{H}}}. \tag{11}$$

For $\rho_{\mathcal{H}} < 1$, the minimum is attained. At $\rho_{\mathcal{H}} = 1$, no strictly positive allowed weight with a well-defined normal equation selects zero, even without uniform bounds.

Why this formula. Normalize $\mathrm{I} \preceq W \preceq \kappa\mathrm{I}$ and write $W = (\kappa+1)\mathrm{I}/2 + (\kappa-1)A/2$, $-\mathrm{I} \preceq A \preceq \mathrm{I}$. The attainable perturbation set $\{\mathbb{E}F^TAe\}$ is compact, symmetric and convex, with support function q. Cancellation is possible exactly when $\rho_{\mathcal{H}} \le (\kappa - 1)/(\kappa + 1)$. At $\rho_{\mathcal{H}} = 1$, an oriented direction c with $c^Tb_{\mathrm{reg}} = q(c) > 0$ yields a nonzero positive-semidefinite conditional pairing that every positive weight preserves. Appendix A proves attainment and the unbounded-weight exclusion.

Thus $\rho_{\mathcal{H}} = 0$ means ordinary weighting already selects the desired field, $0 < \rho_{\mathcal{H}} < 1$ means repair has a finite price, and $\rho_{\mathcal{H}} = 1$ means positivity itself prevents repair. With full matrix access, $q(c) = \mathbb{E}\,\|e\|\,\|Fc\|$. For scalar weights under information $\mathcal{H}$, it is instead $\mathbb{E}|\mathbb{E}[e^T Fc \mid \mathcal{H}]|$, reducing to $\mathbb{E}|e^T Fc|$ at full input access. More information can only reduce the minimum contrast, by conditional Jensen. The zero-tilt squeeze class realizes the boundary $\rho_{\mathcal{H}} = 1$ at every allowed information level: $S_1 = \mathbb{E}[kee^T \mid \mathcal{H}]$ is positive semidefinite and has positive expected trace.

For a scalar parameter, write $S = \mathbb{E}[\mathrm{sym}(eF^T) \mid \mathcal{H}] = S_+ - S_-$, $A_\pm = \mathbb{E}\,\mathrm{tr}\, S_\pm$, and orient a nonzero mismatch positively. Then

$$\kappa_* = \frac{A_+}{A_-} \quad (A_- > 0). \tag{12}$$

The optimal metric upweights the negative spectral subspace; if $A_- = 0$, selection is impossible. Coarser information averages the pairing before its signs can be separated. Appendix A gives the optimizer and equality conditions.

This conditional metric geometry connects metric-angle inequalities and condition-number design (Lin & Sinnamon, 2012; Lu & Pong, 2011); Appendix A.4 gives the finite-block comparison.

Global spectral bounds control both anisotropy and scalar variation across inputs; pointwise condition numbers control only anisotropy. The theorem selects an already specified field; nonlinear tangent equations generally give only stationarity.

Repair under a transverse perturbation. For every $\varepsilon \neq 0$ in the squeeze family, scalar weights still fail, but full matrices can cancel the transverse pairing. With $h = |\varepsilon|$, the exact global contrasts are

$$\begin{aligned} \kappa_{\mathrm{state}} &= \frac{(\sqrt{1+\varepsilon^2}+1)^2}{\varepsilon^2} \sim 4h^{-2}, \\ \kappa_{\mathrm{time}} &= \frac{N_\varepsilon + G_0}{N_\varepsilon - G_0} \sim \frac{3G_0}{2} h^{-3}, \end{aligned} \tag{13}$$

where $G_0 = \int kD\,\mathrm{tr}(EE^T)\,dt = 68/15 - \pi$ and $N_\varepsilon = \int \left\| kD\,\mathrm{sym}\{EE^T(\mathrm{I} - \varepsilon J)\}\right\|_* dt$. Thus every fixed finite global-contrast budget excludes repair sufficiently near zero tilt in this family. A state-dependent matrix uses each residual direction. A time-only matrix averages the pairing first: its useful negative eigendirection occurs only in endpoint bands of width $O(h)$, with magnitude $O(h^2)$. For the chosen $\varepsilon \neq 0$, let $P_-(t)$ be the negative spectral projector of $\mathbb{E}[\mathrm{sym}(eF^T) \mid T = t]$. The optimal time metric is $W_h(t) = \mathrm{I} + (\kappa_{\mathrm{time}} - 1)P_-(t)$; the notation suppresses the sign of ε. Appendix B derives these family-specific asymptotics.

5 Statistical difficulty depends on the selecting rule

An endpoint-derived moment equation remains regular at zero tilt, where no positive selecting weight exists. We compare its estimator with positive-weight regression on the same observation type; Figure 4 (appendix) distinguishes the interventions.

5.1 Importance sampling changes the price, not the population fit

Fix a positive selecting weight before sampling. Draw one fresh independent endpoint pair per time in the squeeze family, and let $\xi = U - u(T, X_T)$. Gaussian conditioning gives $\xi \mid (T, X_T) \sim N(0, D^{-1}\mathrm{I}_2)$. The empirical minimizer is the ratio

$$\widehat{\theta}_n = -\frac{\sum_i A_i}{\sum_i B_i}, \qquad A_i = F_i^T W_i(e_i - \xi_i), \quad B_i = F_i^T W_i F_i. \tag{14}$$

Under $\mathbb{E}A = 0$, $0 < \mathbb{E}B < \infty$ and $\mathbb{E}A^2 < \infty$, its asymptotic variance is $\mathcal{V}_W = \mathbb{E}A^2/(\mathbb{E}B)^2$. Near zero tilt, increasingly extreme selecting weights amplify label noise through WF.

Theorem 3 (The statistical price of the weighted-regression estimator)**.** *For the ratio estimator (14), over positive selecting weights satisfying the stated moment conditions, the best asymptotic-variance orders as $h = |\varepsilon| \downarrow 0$ are:*

(a) $\Theta(h^{-2})$ *for state-dependent matrix weights.*

(b) $\Theta(h^{-3})$ *for time-only matrix weights under uniform time.*

(c) With the time metric W_h*, sampling independently of the endpoints from*

$$\pi_h(t) = 1 - \eta + \frac{\eta}{2h}\mathbf{1}_{[0,h]\cup[1-h,1]}(t), \quad 0 < \eta < 1, \quad 0 < h < \tfrac{1}{2}, \tag{15}$$

and using W_h/π_h *preserves the population objective and gives asymptotic variance of order* $\Theta(h^{-2})$*. No independent time proposal with a positive selecting weight improves this exponent within the estimator protocol. State-dependent proposals are excluded.*

Label noise contributes $\mathbb{E}\,\|WF\|^2/D$ to $\mathbb{E}A^2$, forcing a quadratic price. Under uniform time, useful endpoint bands are also rare; importance sampling visits them more often. Objective-preserving time sampling also appears in diffusion training (Kingma & Gao, 2023, Section 3.2). Appendix C gives matching bounds and constants. The central limit theorem holds at fixed h, followed by the small-h comparison; it is not a finite-sample risk bound.

5.2 An endpoint equation remains regular at the obstruction

Weighted regression tests the residual against WF. A general moment equation instead uses a test field, or *instrument*, Ψ_{IV}: $\mathbb{E}[\Psi_{\mathrm{IV}}^T(v_\theta - U)] = 0$. Liouville's log-volume identity supplies an instrument whose zero identifies the correct squeeze parameter.

Theorem 4 (Liouville endpoint instrument)**.** *Let straight-line endpoints have finite second moments and positive-definite raw moments* $S_t = \mathbb{E}X_tX_t^T$ *for all* $t \in [0,1]$*. For a homogeneous linear field* $v(t,x) = B(t)x$*,* $\int_0^1 \|B(t)\|\,dt < \infty$*, started from the same source, let* Σ_t *be its generated raw second moment. Then*

$$\int_0^1 \mathbb{E}[(S_t^{-1}X_t)^T(v(t,X_t) - U)]\,dt = \frac{1}{2}\log\frac{\det\Sigma_1}{\det S_1}. \tag{16}$$

In the squeeze class this equals $-2b\theta$*, whose unique zero is endpoint-correct for every* $|\varepsilon| \le 1$*, including* $\varepsilon = 0$*.*

The model term is $\operatorname{tr} B$, the label term is $\frac{1}{2}\partial_t \log\det S_t$, and $\partial_t \log\det\Sigma_t = 2\operatorname{tr} B$; integration proves the result. This is the linear form of the instantaneous change-of-variables identity (Chen et al., 2018), requiring neither Gaussianity nor commutativity. At zero tilt its residual pairing is zero, whereas every allowed WF has positive pairing. Determinant matching identifies the full endpoint only within the squeeze class.

The tuple already reveals $Y = X_T + (1-T)U$. With $\widehat{S} = n^{-1}\sum_i Y_iY_i^T$, use $S_t(\widehat{S}) = (1-t)^2\mathrm{I} + t^2\widehat{S}$. The variance calculation accounts for estimating the raw moments from the same tuples. Appendix D proves a uniform root-n expansion on $|\varepsilon| \le 1$, with asymptotic variance about 51.0.

The direct estimator $-\log\det\widehat{S}/(4b)$ has lower asymptotic variance (Table 1); since the target here is known, the constant-zero rule is exact. The instrument establishes that the same observations support a regular estimating equation at the positive-weight selection obstruction.

Table 1: **Estimator comparisons in the squeeze family.** Variance orders concern the specified estimators as $h = |\varepsilon| \downarrow 0$, with their stated moment conditions.

Procedure	Change from ordinary regression	Asymptotic variance
State-dependent positive weight	Metric, using the state	Best order h^{-2}
Time-only positive weight	Metric, uniform time	Best order h^{-3}
Same time metric + importance correction	Sampling; objective unchanged	Order h^{-2}
Liouville plug-in equation	Selecting equation; same tuples	50.9711, all $\|\varepsilon\| \le 1$
Direct log-determinant calibration	Available endpoint statistic	5.4284, all $\|\varepsilon\| \le 1$

6 EMPIRICAL CONSEQUENCES OF THE SELECTION GAP

We test finite-update departures from correct generators and endpoint-directed fitting under fixed numerical samplers. Appendix F gives the protocols and inference scope.

6.1 CFM UPDATES CAN LEAVE CORRECT SYNTHETIC AND PRETRAINED GENERATORS

A local perspective. Let Q_θ denote either a well-posed flow endpoint or a fixed differentiable sampler's law, and let $M_\phi(\theta) = \mathbb{E}_{Q_\theta}\phi$. At $Q_{\theta_0} = p_1$, twice differentiability gives, with $d_R = \nabla R(\theta_0)$,

$$\begin{aligned} R(\theta_0 - \eta d_R) - R(\theta_0) &= -\eta \, \|d_R\|^2 + O(\eta^2), \\ M_\phi(\theta_0 - \eta d_R) - M_\phi(\theta_0) &= -\eta \, DM_\phi(\theta_0)[d_R] + O(\eta^2). \end{aligned} \tag{17}$$

If the endpoint response $DM_\phi[d_R]$ is nonzero, regression descent leaves correctness. That condition is not automatic, and later training could recover another correct generator. For an actual update direction d, the corresponding conditions are $\nabla R^T d < 0$ and $DM_\phi[d] \neq 0$. This is a local diagnostic, separate from the whole-correct-set exclusion in Theorem 1.

A width-128, two-hidden-layer SiLU network with an affine skip represents

$$v_i(x) = 0.6 \ \mathrm{SiLU}(\mathrm{SiLU}(1.5x_i)) - 0.45x_i + 0.15x_{i-1}, \tag{18}$$

with cyclic coordinates. Its stored coefficients and Heun-8 map define the target, so starting there is correct by construction. Solver suffixes count steps. CFM draws fresh independent source/target endpoints and uniform time.

A pilot selected one Adam update and the coordinate-averaged mean from five prespecified observables, then fixed these choices before two independent confirmations. Risk and endpoint evaluations use separate streams. Both models use the same source draws for endpoint evaluation, giving a paired mean comparison.

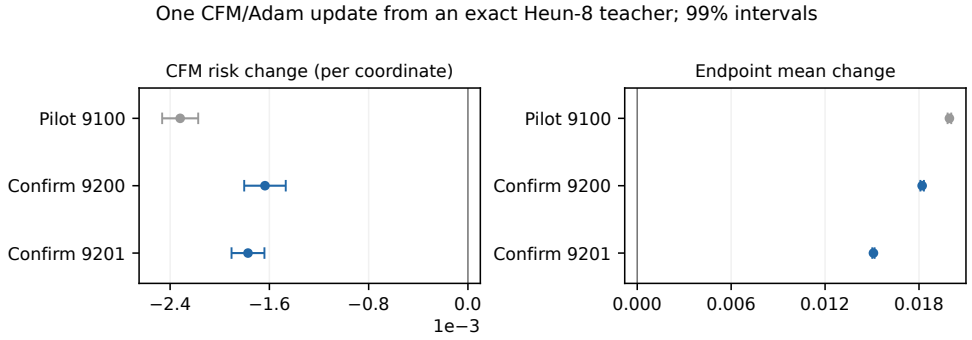


Figure 2: **Improving regression while losing endpoint correctness.** One sampled CFM/Adam update from an exact Heun-8 teacher lowers fresh regression risk and changes the endpoint mean. The pilot selects the update count and observable; both confirmations use fresh training and evaluation. The 99% block intervals condition on the fitted models, not a population of teachers.

The two risk changes are -0.001636 and -0.001772, and the endpoint-mean changes are $+0.018201$ and $+0.015089$; all four approximate 99% block-t intervals exclude zero (Figure 2). The effects exceed measured float32/64 differences. For $\phi(x) = 100^{-1}\sum_i x_i$, $\mathrm{Lip}(\phi) = 0.1$, so the means also imply $W_2(p_1, q) \geq 10|\Delta\mathbb{E}\phi|$, without estimating the full Wasserstein distance.

These are finite sampled Adam updates from one synthetic teacher, distinct from the population-gradient calculation in (17). The longer pilot is nonmonotone, so this diagnostic does not establish failure at convergence.

Pretrained image extension. The pretrained test starts from released TorchCFM CIFAR-10 independent-CFM weights and defines the target as their own Euler-100 law. Fresh Adam refits independent straight-line CFM to teacher-generated targets. In two confirmation streams, eval-mode risk

decreases while pilot-selected Inception observables change at one and 200 updates; all eight nominal 99% intervals have the required signs, also supported by the simultaneous Bonferroni analysis. These comparisons support the existence of a lower-CFM-risk model with a different deployed law. Adam and dropout determine how it was found; they do not identify a population-gradient direction or convergence behavior. Appendix F.2 gives the values, reconstructed baselines and conditional inference.

6.2 Endpoint fitting in an endpoint-realizable structured family

We compare CFM with a two-stage likelihood fit in a structured, endpoint-realizable family. In fixed orthogonal coordinates $x = Q(z, w)$, $z \in \mathbb{R}^8$, $w \in \mathbb{R}^{92}$, consider

$$\dot{z}_i = a_i \tanh(2z_i), \qquad \dot{w} = -kw + C\psi(z), \qquad \psi(z) = (1, z, \sin(Az + \beta)). \tag{19}$$

Both learners know the rotation and features, but not the teacher coefficients. Equation (19) separates the latent flow, contraction $-kw$, and drive $C\psi(z)$. For $0 \le a_i \le 3$, $0 \le k \le 4$, each latent Euler-16 step is an increasing bijection and the complementary multiplier $1 - k/16$ is positive, so the teacher belongs to the actual deployed family. Its endpoint has a multimodal latent law around a curved eight-dimensional graph, with complementary Gaussian noise of standard deviation 0.0938.

CFM minimizes a constrained quadratic in 6,725 coefficients to numerical tolerance. Two-stage likelihood fits the latent inverse density, then the conditional Gaussian density, using the same target samples; a conditional-only control retains the CFM latent fit. Across five datasets of 32,768 observations, CFM has lower fresh regression risk but much worse endpoint Kullback–Leibler (KL) divergence (Table 2).

Table 2: **Lower regression risk need not mean a better endpoint.** Five fitted datasets in the shared structured Euler-16 family. Forward KL is a post-hoc estimate in nats per 100D observation; conditional Monte Carlo error and the decomposition appear in Appendix F.3. Screen rejections count fits rejected by the calibrated moment-based endpoint diagnostic; non-rejection does not establish equality.

Selecting procedure	Risk/coordinate	Forward KL	Screen rejections
Exact teacher	0.96666	0	0/5
Independent CFM	0.63166	57.0616	5/5
Two-stage likelihood	0.96749	0.1053	1/5
Conditional-only likelihood repair	0.96739	0.1736	5/5

The endpoint density factors into a latent density and a conditional Gaussian density. Its KL decomposition separates latent-law, conditional-scale and conditional-mean errors; scale accounts for about 87% of base-family CFM KL. Appendix F.3 isolates this scale effect analytically; it does not test time-weighted fitting in the driven family.

On five further datasets, making every coefficient constant within each of 16 time intervals preserves Euler-16 teacher membership and reduces CFM KL from about 57 to 6.5 nats (88.6%), with fresh risk decreasing from 0.631 to 0.459. Likelihood in the original family reaches 0.104 nats; it was not fitted in the expanded family.

An endpoint gain under a regression-risk budget. Temporal freedom removes most scale error, leaving a gap dominated by the conditional mean. Holding the five temporal fits' per-interval latent and contraction coefficients fixed, we fit the drive to the same noisy target samples. A norm-capped fit lowers KL from 6.52 to 0.68 but raises fresh CFM risk about $142\times$. Under the prespecified empirical 10% risk budget, the risk-constrained fit instead reaches KL 1.96, retaining 78.1% of the norm-capped gain (conditional 95% interval 77.7–78.4%). A coefficient-segment control reaches KL 6.20. Fresh per-fit risk ratios are 1.09964–1.10041, so empirical feasibility is not a strict population cap. These sample-only fits use known coordinates and fixed profiles; KL is evaluated post hoc using the known teacher. The teacher belongs to the full family, generally not this subfamily; Appendix F.3 gives the full evaluation and controls.

6.3 Two endpoint corrections capture most of a longer solve's gain

For two coupled nonlinear targets in $\mathbb{R}^2$, a bounded 28-parameter field contains the continuous teacher. We compare baseline CFM, two damped endpoint corrections fitting 192 Fourier-feature means through RK4-8, and a longer bounded endpoint solve; a 78-parameter CFM class and endpoint-validation-selected CFM are stronger baselines. Four pilot datasets fixed two correction attempts before 24 fresh datasets per target. For energy-distance U-statistic scores $\mathcal{E}_{\mathrm{CFM}}, \mathcal{E}_{\mathrm{long}}, \mathcal{E}_2$, paired 95% bootstrap analyses support

$$\mathbb{E}[\mathcal{E}_{\mathrm{CFM}} - \mathcal{E}_2] \geq 0.8\,\mathbb{E}[\mathcal{E}_{\mathrm{CFM}} - \mathcal{E}_{\mathrm{long}}] > 0.$$

Thus two updates retain at least 80% of the long solve's *mean* gain on both targets at about 26–32 times lower measured total cost (Figure 3). The long comparator uses a different bounded least-squares solver on the same objective, capped at 200 residual calls (Appendix F.4).

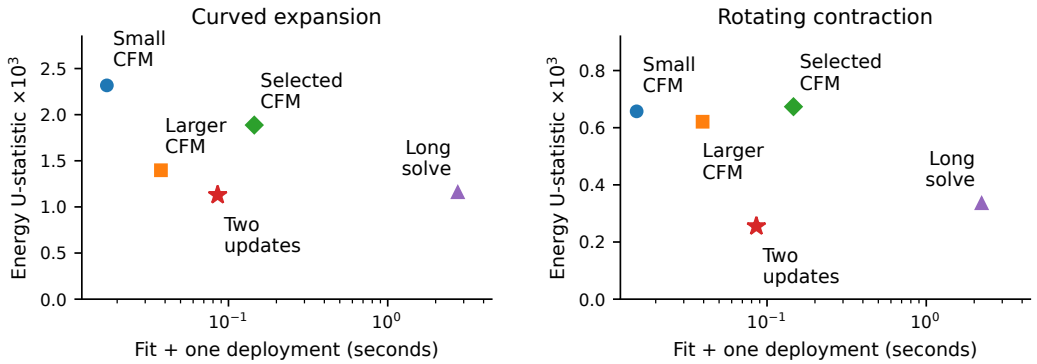


Figure 3: **Most of the endpoint benefit need not require a long solve.** Means over 24 datasets per target; cost is fitting plus one 2,048-point deployment on the same CPU. Lower and left are better. Fixed larger CFM is cheaper; paired comparisons use the mean-gain criterion (Appendix F.4).

Against fixed larger CFM, energy favors correction on contraction and is uncertain on expansion; secondary empirical-Wasserstein comparisons are uncertain on both. Appendix F.4 gives full values and cost accounting. These timings do not establish neural-scale costs or numerical-sampler identity.

7 Discussion

Our constructions separate endpoint realizability from population selection under allowed positive weights. Conditional geometry characterizes reweighting, while same-observation comparisons show that estimator-specific cost need not constrain alternative equations. Input-measurable weights preserve the conditional velocity target; endpoint-dependent weights can change it (Xie et al., 2026, Theorems 5.4 and B.1), and our bounded label-dependent construction selects the excluded squeeze generator (Appendix H). Couplings can likewise change the target. Path changes preserving the marginal path and reference velocity can reduce variance without changing population minimizers (Tyurin, 2026; Claustre et al., 2026); equal-loss discrepancies (Wang et al., 2026a), particular time-weight failures (Wang et al., 2026c), and trajectory-dependent stopping (Lee et al., 2026) address distinct questions.

Other interventions include reflow (Tong et al., 2024; Liu et al., 2023), divergence alignment (Huang et al., 2025) and distribution-matching distillation (Yin et al., 2024). Weighting under misspecification (Shimodaira, 2000; Chow & Kasahara, 2026), moment equations (Hansen, 1982), feature matching (Gretton et al., 2012; Li et al., 2015) and likelihood fitting (Papamakarios et al., 2021) provide related alternatives; Appendix H compares the closest settings.

The experiments show finite-update departures and finite-data repairs in structured models under fixed samplers; neural population-minimizer failure and prevalence remain open. Endpoint objectives still require identification, estimation and optimization: our exploratory neural moment fit loses structural gains, while finite-batch quantile objectives can prefer an incorrect Gaussian scale (Appendix G). Scalable endpoint-valid selection therefore requires more than an arbitrary terminal loss.

AI USE STATEMENT

Generative AI tools were used throughout this work. They contributed substantially to developing theoretical constructions, formulating mathematical claims, turning intuitions and proof sketches into candidate proofs, identifying related work and helping position the paper, designing and implementing experiments, writing code used to generate synthetic datasets, carrying out numerical analysis, interpreting results, critical review, and manuscript drafting and editing. The authors independently developed and refined theoretical formulations, personally checked the proofs and worked through alternative derivations, made the scientific and methodological decisions, and determined which claims, experiments, and interpretations were retained. AI-assisted proofs and numerical results were further checked through the independent derivations, numerical implementations, and reproducibility checks described in the appendices and supplementary material; these checks are not formal verification. The authors take full responsibility for the final text, claims, code, and other artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

Appendices A–E provide the assumptions and proofs for the main theoretical results. Appendix F specifies the models, data-generation procedures, samplers, estimators, experimental protocols, and uncertainty calculations, and distinguishes saved-model reevaluation and targeted refitting from training procedures that were not repeated. Appendix G gives the endpoint-surrogate controls. The supplementary material contains anonymized code and configurations for the main experiments, selected saved numerical outputs, figure data and plotting scripts, and CPU checks for theoretical calculations and reported experimental results. Its README maps these materials to the paper and identifies external dependencies and the limits of the included reproductions.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15571v3.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramér distance as a solution to biased Wasserstein gradients. arXiv:1705.10743, 2017. URL https://arxiv.org/abs/1705.10743.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. arXiv:2305.16860, 2023. URL https://arxiv.org/abs/2305.16860.

Lukas Billera, Hedwig Nora Nordlinder, and Ben Murrell. Time dependent loss reweighting for flow matching and diffusion models is theoretically justified. arXiv:2511.16599, 2025. URL https://arxiv.org/abs/2511.16599.

Antonin Chambolle and Johannes Hertrich. Limit points of reflow with minibatch optimal transport. arXiv:2608.07042v1, 2026. URL https://arxiv.org/abs/2608.07042v1.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Chun Pang Chow and Hiroyuki Kasahara. When is GMM actually LATE? weighting matrices and causal interpretation in overidentified IV. arXiv:2604.07131, version 3, 2026. URL https://arxiv.org/abs/2604.07131v3.

Arsène Claustre, Hugo Negrel, Claire Boyer, Kimia Nadjahi, and Eric Vanden-Eijnden. Gaussian flow-matching schedules: Implications for sampling and training. arXiv:2609.25839v1, 2026. URL https://arxiv.org/abs/2609.25839v1.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Yansen Han, Hongxin Sun, and Tao Lin. When can conditional flow matching replace pointwise negative log-likelihood? arXiv:2608.28010v2, 2026. URL https://arxiv.org/abs/2608.28010v2.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. arXiv:2410.20587, version 3, 2025. URL https://arxiv.org/abs/2410.20587v3.

Yuhao Huang, Taos Transue, Shih-Hsin Wang, William M. Feldman, Hong Zhang, and Bao Wang. Improving flow matching by aligning flow divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25813–25834, 2025. URL https://proceedings.mlr.press/v267/huang25ag.html.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.00848v7.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures. arXiv:2602.22486, version 2, 2026. URL https://arxiv.org/abs/2602.22486v2.

Leo Larsson, Zsolt Páles, and Lars-Erik Persson. Carlson type inequalities for finite sums and integrals on bounded intervals. *Bulletin of the Australian Mathematical Society*, 71(2):275–284, 2005. doi: 10.1017/S0004972700038247.

Cheuk-Kit Lee, Florentin Coeurdoux, Yuyuan Chen, Sophia Tang, Peter Potaptchik, Yilun Du, Michael Samuel Albergo, and Eric Vanden-Eijnden. Beckmann transport models: From autonomous flows to one-step maps. arXiv:2608.01692, version 3, 2026. URL https://arxiv.org/abs/2608.01692v3.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1718–1727, 2015. URL https://proceedings.mlr.press/v37/li15.html.

Minghua Lin and Gord Sinnamon. The generalized Wielandt inequality in inner product spaces. arXiv:1201.6294, 2012. URL https://arxiv.org/abs/1201.6294.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Zhaosong Lu and Ting Kei Pong. Minimizing condition number via convex programming. *SIAM Journal on Matrix Analysis and Applications*, 32(4), 2011. doi: 10.1137/100795097. URL https://zhaosong-lu.github.io/ResearchPapers/cond_opt_rev.pdf.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

Alexander Tyurin. Same flow, different paths: Variance reduction in flow matching. arXiv:2609.17287v1, 2026. URL https://arxiv.org/abs/2609.17287v1.

Chen Wang, Peiran Yun, Pan Xie, and Ke Deng. Beyond trajectory matching: Reflow with marginal distribution alignment. arXiv:2606.29287, version 1, 2026a. URL https://arxiv.org/abs/2606.29287v1.

Leda Wang, Zhehao Xu, Qiang Liu, and Harrison H. Zhou. Computational and statistical guarantees of the c-rectified flow. arXiv:2608.02487v1, 2026b. URL https://arxiv.org/abs/2608.02487v1.

Zimo Wang, Ishit Mehta, Haolin Lu, Chung-En Sun, Ge Yan, Tsui-Wei Weng, and Tzu-Mao Li. Distance marching for generative modeling. arXiv:2602.02928, version 1, 2026c. URL https://arxiv.org/abs/2602.02928v1.

Jize Xie, Haomiao Wu, Qiang Chen, Xiu Su, and Yi Chen. Decision-weighted flow matching for contextual stochastic optimization. arXiv:2606.16790v1, 2026. URL https://arxiv.org/abs/2606.16790v1.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. arXiv:2405.14867, 2024. URL https://arxiv.org/abs/2405.14867.

NOTATION GUIDE

Symbols are scoped to their constructions. The source, target and specified generator determine endpoint correctness. Expectations use $\mathbb{E}$, while E denotes a residual matrix in the squeeze construction.

Shared setup	$e = v_* - u$, parameter features F, loss metric W, allowed information $\mathcal{H}$, regression mismatch $b_{\mathrm{reg}} = \mathbb{E}F^T e$.
Squeeze family	$D = D_1$ from Section 2; loop $M = e^{fK_0}$, $K_0 = \mathrm{diag}(1, -1)$, residual matrix $E = C_0 - a\mathrm{I}$, $Q = EE^T$, and $h = \lvert\varepsilon\rvert$.
Estimator	$A = F^T W(e - \xi)$, $B = F^T WF$, variance $\mathcal{V}_W$; the sampling proposal π carries a reciprocal importance correction.
Moments and endpoints	$S_t = \mathbb{E}X_t X_t^T$ is the reference raw second moment, Σ_t the generated raw second moment, q_1^v an endpoint law, and ϕ an endpoint observable. Solver suffixes such as Euler-16 and Heun-8 denote the deployed step count.
Elliptic family	Profile $g(t)$, $G = \int g$; a positive loss weight $w(t)$ gives $G_w = \int wg$ and weighted moments $I_{0,w}, I_{1,w}, M_w$.

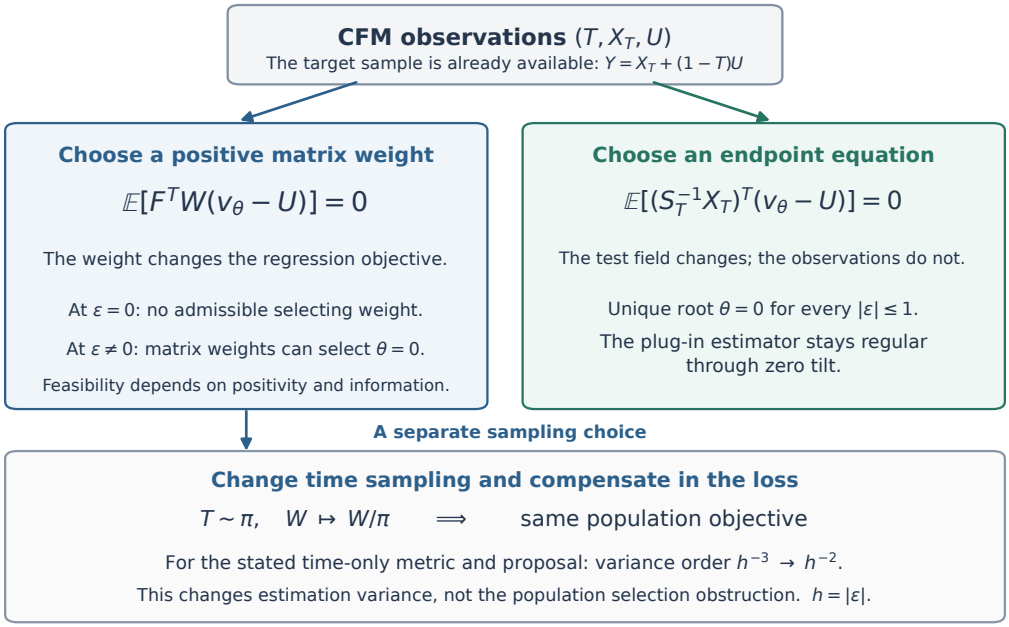


Figure 4: **Selection and sampling are different choices.** The squeeze family separates changing a positive regression metric, changing the estimating equation, and changing the time proposal with reciprocal correction. The two upper branches use the same CFM observations under uniform time. The lower branch changes the time draws, independently of endpoints, while preserving the weighted population objective. The variance orders concern the ratio estimator and proposal in Theorem 3, with their stated moment conditions. The endpoint instrument in Theorem 4 demonstrates regular estimation, not optimality: direct log-determinant calibration has lower asymptotic variance than the instrument, and the known-zero rule is exact.

A SELECTING-WEIGHT GEOMETRY

A.1 NORMAL EQUATIONS AND PROOF OF THEOREM 2

In this section write $\rho = \rho_{\mathcal{H}}$ for the fixed information field. Let $Z = (T, X_T)$, with $u(Z) = \mathbb{E}[U \mid Z]$. For a fixed input-measurable symmetric weight, conditional expectation eliminates the cross term between $U - u$ and every input-measurable residual. Hence

$$R_W(\theta) = \mathbb{E}[(U - u)^T W(U - u)] + \mathbb{E}[(e + F\theta)^T W(e + F\theta)].$$

If $\mathrm{I} \preceq W \preceq \kappa \mathrm{I}$, the parameter Gram is positive definite because $\mathbb{E}F^T F \succ 0$, and the unique minimizer is zero exactly when $\mathbb{E}F^T We = 0$. The same conclusion holds for an unbounded weight whenever the indicated risk and Gram are finite and the Gram is positive definite.

For bounded weights, rescale by the positive lower spectral bound. Define

$$\mathcal{C} = \{\mathbb{E}[F^T Ae] : A \text{ is } \mathcal{H}\text{-measurable and symmetric}, \ -\mathrm{I} \preceq A \preceq \mathrm{I}\}.$$

This set is convex and centrally symmetric. The matrix-valued unit ball is weak-star compact in L^∞; the finite collection of linear functionals defining $\mathcal{C}$ has integrable coefficients because $F, e \in L^2$. The measurable and semidefinite restrictions are weak-star closed. Thus $\mathcal{C}$ is compact in $\mathbb{R}^m$.

For every c, conditioning on $\mathcal{H}$ gives

$$\sup_{z \in \mathcal{C}} c^T z = \sup_A \mathbb{E} \operatorname{tr}(AS_c) = \mathbb{E} \left\| S_c \right\|_* = q(c).$$

The spectral sign of S_c, with value zero on its kernel, is measurable and attains the pointwise maximum. Since $c^T b_{\mathrm{reg}} = \mathbb{E} \operatorname{tr} S_c$, we have $|c^T b_{\mathrm{reg}}| \le q(c)$; in particular b_{reg} annihilates $\ker q$. Work on the finite-dimensional quotient by that kernel. There q is a norm unless the quotient is trivial, and the dual ratio defining ρ is attained on its unit sphere.

The normalized weight can be written

$$W = \alpha \mathrm{I} + \beta A, \qquad \alpha = (\kappa + 1)/2, \quad \beta = (\kappa - 1)/2.$$

For $\kappa > 1$, selection is equivalent to $-b_{\mathrm{reg}} \in (\beta/\alpha)\mathcal{C}$. The supporting-halfspace description of a compact convex set makes this equivalent to $|c^T b_{\mathrm{reg}}| \le (\beta/\alpha) q(c)$ for all c. Solving the resulting inequality proves (11), including attainment for $\rho < 1$. If $b_{\mathrm{reg}} = 0$, $W = \mathrm{I}$ already selects zero and the contrast is one.

For $\rho = 1$, choose and orient an attained quotient direction so that $c^T b_{\mathrm{reg}} = q(c) > 0$. The nonnegative random variable $\left\| S_c \right\|_* - \operatorname{tr} S_c$ has zero expectation, so $S_c \succeq 0$ almost surely and is nonzero on positive measure. Every strictly positive allowed W then has $\operatorname{tr}(WS_c) > 0$ on that set. Admissibility gives finite $\mathbb{E}[e^T We]$ and $\mathbb{E}[(Fc)^T W(Fc)]$; if finite risk is specified first at another affine parameter, the finite weighted Gram transfers finiteness to the candidate. Hence weighted Cauchy–Schwarz gives

$$\mathbb{E}|e^T WFc| \le \{\mathbb{E}[e^T We] \, \mathbb{E}[(Fc)^T W(Fc)]\}^{1/2} < \infty.$$

Put $W_n = W \mathbf{1}_{\{\|W\|_{\mathrm{op}} \le n\}}$. Bounded conditional expectation yields $\mathbb{E}[e^T W_n Fc] = \mathbb{E}[\mathbf{1}_{\{\|W\|_{\mathrm{op}} \le n\}} \operatorname{tr}(WS_c)]$. Dominated convergence on the left and monotone convergence on the right give $c^T \mathbb{E}F^T We = \mathbb{E} \operatorname{tr}(WS_c) > 0$, contradicting selection. This proves the unbounded-weight exclusion as well; the truncations are used only to justify the identity, not as admissible fitting weights.

A.2 Scalar corollary and the two contrast notions

For a positive matrix weight define

$$\kappa_{\mathrm{glob}}(W) = \frac{\operatorname{ess\,sup} \lambda_{\max}(W)}{\operatorname{ess\,inf} \lambda_{\min}(W)}, \qquad \kappa_{\mathrm{pt}}(W) = \operatorname{ess\,sup} \frac{\lambda_{\max}(W)}{\lambda_{\min}(W)}.$$

Finite global contrast requires finite common bounds with a positive lower bound. Always $\kappa_{\mathrm{pt}} \le \kappa_{\mathrm{glob}}$; a scalar multiple varying across inputs can change the latter without changing the former.

For $m = 1$, orient a nonzero mismatch positively and write $S = \mathbb{E}[\operatorname{sym}(eF^T) \mid \mathcal{H}] = S_+ - S_-$, $A_\pm = \mathbb{E} \operatorname{tr} S_\pm$. Then $q(1) = A_+ + A_-$ and $b_{\mathrm{reg}} = A_+ - A_- > 0$. Theorem 2 gives (12). Equivalently, under $\mathrm{I} \preceq W \preceq \kappa \mathrm{I}$,

$$\mathbb{E} \operatorname{tr}(WS) \ge A_+ - \kappa A_-,$$

with equality for $W = \mathrm{I} + (\kappa - 1)P_-$. Spectral projectors are measurable Borel functions of S. If $A_- = 0$, the positive pairing cannot vanish; if $b_{\mathrm{reg}} = 0$, I already selects and the minimum contrast is one.

Positive constant rescaling preserves selection. For $A_- > 0$, normalize an optimal weight to $\mathrm{I} \preceq W \preceq \kappa_* \mathrm{I}$. Equality forces $(W - \mathrm{I})S_+ = 0$ and $(\kappa_* \mathrm{I} - W)S_- = 0$ almost surely, because their traces are nonnegative with zero expectation. Thus all normalized global optimizers equal I on $\operatorname{ran} S_+$, $\kappa_* \mathrm{I}$ on $\operatorname{ran} S_-$, and any symmetric operator with spectrum in $[1, \kappa_*]$ on $\ker S$. This characterization concerns one scalar normal equation.

A.3 Access specializations and construction limits

For full input access, the nuclear norm of $\mathrm{sym}(ab^T)$ is $\|a\|\,\|b\|$: its two possibly nonzero eigenvalues are $(a^Tb \pm \|a\|\,\|b\|)/2$. This proves the full-state formula for q. Hence $\rho_{\mathcal{H}} = 1$ exactly when there is a direction c with $q(c) > 0$ such that, after possibly reversing c, $e^TFc = \|e\|\,\|Fc\|$ almost surely. This characterizes selection of the prescribed field; excluding every correct field still needs a class-specific argument. For scalar weights measurable from $\mathcal{H}$, the admissible perturbation is $A = s(Z)\mathrm{I}$, $|s| \le 1$, and the support is $\mathbb{E}|\mathbb{E}[e^TFc \mid \mathcal{H}]|$. With full input access this reduces to $\mathbb{E}|e^TFc|$. If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, conditional Jensen gives $q_{\mathcal{H}_1}(c) \le q_{\mathcal{H}_2}(c)$, proving monotonicity of the optimal contrast.

For a fixed candidate, let $r_* = v_*(Z) - U$. Since $\mathbb{E}[r_* \mid Z] = e(Z)$ and F is Z-measurable, conditional expectation gives $\mathbb{E}[F^Tr_*] = b_{\mathrm{reg}}$; the tower property for $\mathcal{H} \subseteq \sigma(Z)$ gives $\mathbb{E}[\mathrm{sym}\{r_*(Fc)^T\} \mid \mathcal{H}] = S_c$. The products are integrable by the stated second-moment assumptions. Thus finite-partition conditional moments reduce to cell averages, while continuous information requires conditional-moment estimation. The conditional expectation must be estimated before taking its nuclear norm, since averaging norms of individual noisy residual products estimates a different functional. Population identification is not a finite-sample guarantee, contrast estimation becomes unstable near $\rho_{\mathcal{H}} = 1$, and a data-selected candidate requires appropriate handling of reuse, such as an independent evaluation split.

For fixed candidate and information access, enlarging the affine direction space cannot reduce the minimum selecting contrast. If the enlarged design is $[F, G]$, restricting its support ratio to directions $(c, 0)$ gives $\rho_{\mathcal{H}}([F, G]) \ge \rho_{\mathcal{H}}(F)$ and hence $\kappa_*([F, G]) \ge \kappa_*(F)$. An excluded field therefore remains excluded under exact nesting and the unchanged objective, although the larger class may contain other endpoint-correct fields.

A pointwise bound on $\lambda_{\max}(W(z))/\lambda_{\min}(W(z))$ allows arbitrary scalar variation with z, whereas the theorem bounds the spectrum between two common constants. These constraints are different. Likewise, a spectral sign chosen for one support direction need not solve every coordinate of the normal equation. The compact-set proof establishes an attainable solution to the whole vector equation; it does not establish a universal one-direction recipe or an efficient procedure for finding v_*.

A.4 Classical metric design and the information restriction

For deterministic $e \neq 0$ and feature subspace $V = \mathrm{ran}\,F$, Theorem 2 gives $\rho = \|\,\mathrm{Proj}_V\, e\|/\|e\|$. Writing this as $\cos\phi$ gives $\kappa_* = \cot^2(\phi/2)$, the classical metric-angle threshold associated with the generalized Wielandt inequality (Lin & Sinnamon, 2012).

For finitely many information atoms with probabilities p_j, write $S_{i,j} = \mathbb{E}[\mathrm{sym}(eF_i^T) \mid j]$. A selector satisfies $\sum_j p_j\,\mathrm{tr}(W_jS_{i,j}) = 0$ for each parameter coordinate. With

$$\mathcal{W} = \mathrm{diag}(W_1, \ldots, W_n), \qquad \mathcal{C}_i = \mathrm{diag}(p_1S_{i,1}, \ldots, p_nS_{i,n}),$$

these become linear constraints $\mathrm{tr}(\mathcal{W}\mathcal{C}_i) = 0$, and the condition number of $\mathcal{W}$ is precisely the global contrast. Thus the finite feasible problem is an instance of convex constrained condition-number design (Lu & Pong, 2011). The probabilities belong in the constraint matrices, not in the metric blocks. The measurable-field proof above identifies its conditional nuclear-norm support and establishes the positivity obstruction even for unbounded weights.

For affine generalized method of moments (GMM) equations $m(\theta) = a + B\theta$, selecting a prescribed θ_* with a fixed positive metric A requires $B^TAm(\theta_*) = 0$, the deterministic metric problem. Target-first GMM weight design is established in Chow & Kasahara (2026, Proposition 16). Unrestricted moment-space matrices can mix observations; our loss weights cannot. Conversely, finite-information regression can be lifted to structured minimum distance by factoring the conditional second moments of $[e, F]$; the metric must then repeat within each information block. The distinction is the allowed structure, not merely whether expectation is taken before or after a quadratic loss. Neither this population equivalence nor Theorem 2 supplies the same empirical estimator or an efficiency bound.

B SQUEEZE FAMILY: WHOLE-SET FAILURE AND CONDITIONING

B.1 REACHABILITY AND TERMINAL VOLUME

Since $K_0 = \mathrm{diag}(1, -1)$ has trace zero and $f(0) = f(1) = 0$, the loop $M = e^{fK_0}$ has determinant one and equals the identity at both endpoints. Differentiation gives $C_0 = f'K_0 = (1-2t)K_0$, whose fundamental matrix is M.

Gaussian conditioning of (1) gives $u(t,x) = a(t)x$ and conditional label covariance $D^{-1}\mathrm{I}$. Put $r = 2t - 1$. The residual matrix $E = -r(K_0 + D^{-1}\mathrm{I})$ is symmetric, so $\mathrm{tr}(JE) = 0$. With $D = (1+r^2)/2$ and $k = 1 + r/2$, odd-part cancellation gives

$$\int_0^1 ka\, dt = \int_{-1}^1 \frac{r^2}{2(1+r^2)}\, dr = 1 - \pi/4 = b, \qquad \int_0^1 \mathrm{tr}\, H_\varepsilon\, dt = -2b.$$

Liouville's formula therefore gives determinant $e^{-2b\theta}$ for the terminal fundamental matrix and (7) for the generated covariance, independently of ε. Since $b > 0$, equality with the standard Gaussian requires $\theta = 0$, which is sufficient by the loop construction.

B.2 POSITIVE WEIGHTS CANNOT SELECT THE UNIQUE CORRECT MEMBER AT ZERO

With $e = EX$ and $F = H_\varepsilon X$, weighted regression gives

$$\theta_W = -\frac{\mathbb{E}F^T We}{\mathbb{E}F^T WF}.$$

At $\varepsilon = 0$, $F = ke$, so this is a positive weighted average of $-1/k$ and lies strictly between -2 and $-2/3$. The endpoint times have zero measure, and the effective density is positive on a set of positive measure. No spectral bound is needed.

For any coupling of $X \sim N(0, \mathrm{I}_2)$ and $Y \sim N(0, \Sigma)$, the reverse triangle inequality in L^2 gives

$$\left(\mathbb{E}\,\|X - Y\|^2\right)^{1/2} \geq \left|\sqrt{\mathrm{tr}\,\Sigma} - \sqrt{2}\right|.$$

If $\det \Sigma \geq 1$, then $\mathrm{tr}\,\Sigma \geq 2\sqrt{\det \Sigma} \geq 2$. Taking the infimum over couplings and squaring therefore gives

$$W_2^2\big(N(0,\mathrm{I}_2), N(0,\Sigma)\big) \geq 2\big[(\det\Sigma)^{1/4} - 1\big]^2.$$

Here $\det \Sigma = e^{-4b\theta_W} > 1$. The previously established bound $\theta_W < -2/3$ and monotonicity of $(x-1)^2$ on $x \geq 1$ prove (8).

In the centered Gaussian setting, the determinant bound is attained exactly for isotropic covariance: anisotropy makes the trace–determinant inequality strict, while scalar dilation attains equality.

For $W = w\mathrm{I}$, skew symmetry gives $F^T We = wk\,\|e\|^2$ and $F^T WF = wk^2(1+\varepsilon^2)\,\|e\|^2$. Hence

$$-\frac{2}{1+\varepsilon^2} < \theta_w < -\frac{2}{3(1+\varepsilon^2)},$$

so every positive scalar weight fails for all stated ε.

B.3 THE OPTIMAL STATE-DEPENDENT MATRIX

For $\varepsilon \neq 0$, set $s = \sqrt{1+\varepsilon^2}$, $n = e/\,\|e\|$, and $R = (n, Jn)$ where $e \neq 0$. Define

$$W_\varepsilon = \frac{1}{s(s-1)} R \begin{pmatrix} \varepsilon^2 & -\varepsilon \\ -\varepsilon & \varepsilon^2 + 2 \end{pmatrix} R^T, \tag{20}$$

and set $W = \mathrm{I}$ at $e = 0$. The inner eigenvalues are $s(s-1)$ and $s(s+1)$. In this frame e is parallel to $(1,0)$ and F to $(1,\varepsilon)$, giving $F^T We = 0$. Thus (20) selects zero with contrast $(s+1)/(s-1)$.

For necessity, suppose the essential supremum of the pointwise condition numbers is at most κ, and put $\ell = \lambda_{\min}(W) > 0$. Then

$$\left\|W/\ell - \tfrac{\kappa+1}{2}\mathrm{I}\right\|_{\mathrm{op}} \leq \tfrac{\kappa-1}{2}, \qquad F^T We \geq \frac{\ell k\,\|e\|^2}{2}\big[(\kappa+1) - s(\kappa-1)\big].$$

For $\kappa < (s+1)/(s-1)$, this pairing is nonnegative and strictly positive on $\{e \neq 0\}$, a set of positive measure. Admissibility makes its expectation finite, so selection is impossible. Pointwise contrast is bounded by global contrast, proving both lower bounds. Geometrically, if $\phi = \arctan|\varepsilon|$ is the angle between e and F, the optimal contrast is $\cot^2(\phi/2)$, the classical metric-angle threshold (Lin & Sinnamon, 2012).

B.4 Time-only contrast and its cubic asymptotic

Put $Q = EE^T$ and $S_\varepsilon(t) = kD\,\mathrm{sym}\{Q(\mathrm{I} - \varepsilon J)\}$. The time-only selecting equation is $\int \mathrm{tr}(WS_\varepsilon) = 0$. Its trace integral is

$$G_0 = \int_0^1 kD\,\mathrm{tr}\,Q\,dt = \frac{68}{15} - \pi > 0.$$

Let $S_\varepsilon = S_+ - S_-$, $A_\pm = \int \mathrm{tr}\,S_\pm$, so that $A_+ - A_- = G_0$ and $A_+ + A_- = N_\varepsilon$. The scalar corollary in Appendix A.2 gives $\kappa_* = (N_\varepsilon + G_0)/(N_\varepsilon - G_0)$ when $A_- > 0$. It also gives all normalized global optimizers. For the variance calculation we use

$$W_h = \mathrm{I} + (\kappa_* - 1)P_-,$$

which agrees with the centered spectral-sign choice off the measure-zero spectral boundaries. For state access, the eigenvalues of $\mathrm{sym}(eF^T)$ are $k\,\|e\|^2\,(1 \pm s)/2$, so the same corollary recovers the preceding global contrast.

For completeness, write $r = 2t - 1$, $E = -r(K_0 + D^{-1}\mathrm{I})$. For $\overline{Q} = (K_0 + D^{-1}\mathrm{I})^2$, we have $Q = r^2\overline{Q}$. Define

$$\tau = \mathrm{tr}\,\overline{Q} = 2 + 2D^{-2}, \quad d_0 = \det\overline{Q} = (D^{-2} - 1)^2, \quad \ell = \sqrt{\tau^2 - 4d_0} = 4/D.$$

The eigenvalues of $\mathrm{sym}\{Q(\mathrm{I} - \varepsilon J)\}$ are $(\mathrm{tr}\,Q \pm s(\lambda_1(Q) - \lambda_2(Q)))/2$. Symmetry under $t \mapsto 1 - t$, with $k(t) + k(1-t) = 2$, therefore gives

$$N_\varepsilon - G_0 = 2\int_0^{1/2} Dr^2[s\ell - \tau]_+ dt. \tag{21}$$

At zero, $\tau = \ell = 4$, so the integral is positive for every $\varepsilon \neq 0$. Near zero, $d_0 = 16t^2 + O(t^3)$, $\tau = 4 + O(t)$, and

$$s\ell - \tau = \frac{h^2\ell^2 - 4d_0}{s\ell + \tau}, \qquad h^{-2}(s\ell(hy) - \tau(hy)) \longrightarrow 2 - 8y^2.$$

On any fixed interior interval d_0 is bounded below; near an endpoint $d_0 \geq ct^2$. Thus the positive part is confined to $O(h)$ endpoint bands and has a uniformly bounded rescaled integrand. Dominated convergence in (21) yields

$$(N_\varepsilon - G_0)/h^3 \longrightarrow 2\int_0^{1/2}(2 - 8y^2)dy = 4/3,$$

which proves the cubic asymptotic. The constant G_0 follows by integrating $r^2\{2D + 2/D\}$ over $t \in [0,1]$, using the symmetry of k.

Pointwise contrast has a different infimum. For $\varepsilon \neq 0$, use $Q = r^2\overline{Q}$ and the existing trace τ and eigengap ℓ. On the indefinite part of S_ε, its positive-to-negative spectral mass ratio is

$$\frac{\tau + s\ell}{s\ell - \tau} > \frac{s+1}{s-1}.$$

Indeed, Q is positive definite on $(0,1) \setminus \{1/2\}$, so $\ell/\tau < 1$; at either endpoint $\ell/\tau \to 1$. Every admissible time-only weight with pointwise contrast at most $(s+1)/(s-1)$ has positive integrated pairing, so the threshold is not attained. For any larger κ, $\mathrm{I} + (\kappa - 1)P_-$ gives negative pairing on endpoint bands of positive measure; increasing its positive scalar multiplier there, or interpolating with I if its integral is already nonpositive, produces zero integrated pairing. Each such construction has finite common bounds for fixed ε, κ. Thus the time-only pointwise infimum is $(s+1)/(s-1) \sim 4h^{-2}$, equal to the attained state-dependent minimum. This does not remove the uniform-time h^{-3} variance penalty: Theorem 3 allows arbitrary scalar variation in time-only weights.

C ESTIMATOR-SPECIFIC VARIANCE AND IMPORTANCE SAMPLING

The conditional Gaussian noise formula follows from $\mathrm{Cov}(U, X) = (2t-1)\mathrm{I}$, $\mathrm{Cov}(X) = D\mathrm{I}$, and $\mathrm{Var}(U \mid X, t) = 2\mathrm{I} - (2t-1)^2\mathrm{I}/D = D^{-1}\mathrm{I}$. For fixed h, the CLT and law of large numbers applied to (14) give the stated asymptotic variance. In particular,

$$\mathbb{E}A^2 = \mathbb{E}(F^T We)^2 + \mathbb{E}\frac{\|WF\|^2}{D} \geq V_{\text{noise}} := \mathbb{E}\frac{\|WF\|^2}{D}. \tag{22}$$

C.1 QUADRATIC LOWER BOUND AND AN ATTAINING WEIGHT

Let $m = WF$, $B = m^T F > 0$ where $F \neq 0$, and $\Gamma = \mathbb{E}B$. Since $F = k(e + \varepsilon Je)$, the selecting equation implies

$$0 = \mathbb{E}(m^T e) = \mathbb{E}(B/k) - \varepsilon\mathbb{E}(m^T Je).$$

Using $k \leq 3/2$ and Cauchy–Schwarz weighted by D,

$$\Gamma/(3/2) \leq h\sqrt{V_{\text{noise}}}\sqrt{\mathbb{E}[D\,\|e\|^2]}.$$

The final expectation is $R_0 = \int D^2 \operatorname{tr} Q = 116/105$, so $\mathcal{V}_W \geq 1/[(9/4)R_0 h^2]$ for every selecting state-dependent weight with the stated moments.

For (20), direct multiplication gives

$$WF = \frac{s(s+1)}{\varepsilon}kJe, \qquad F^T WF = s(s+1)k^2\,\|e\|^2\,, \qquad F^T We = 0.$$

Consequently

$$\mathcal{V}_{\text{state}} = \frac{B_0}{A_0^2 h^2}, \quad A_0 = \frac{83}{21} - \frac{3\pi}{4}, \quad B_0 = \frac{\pi}{4} + \frac{19}{15},$$

with $B_0/A_0^2 \approx .8054226358$. This proves the optimal state-dependent order. For the specific weight (20), this large-sample variance equals the Liouville variance at $h \approx 0.1257$ and the direct log-determinant variance at $h \approx 0.3852$; these are not crossings for the unknown variance-optimal state weight. This formula also follows from the exact conditional-Gaussian ratio $\widehat{\theta}^{\text{state}} = \varepsilon^{-1}\sum_i k_i (Je_i)^T \xi_i / \sum_i k_i^2\,\|e_i\|^2$.

C.2 UNIFORM-TIME LOWER BOUND FOR EVERY TIME-ONLY SELECTING WEIGHT

Use $\mathsf{B} = H_\varepsilon H_\varepsilon^T$, $S = D\,\mathrm{sym}(EH_\varepsilon^T)$. Gaussian integration gives

$$\Gamma = \int D\operatorname{tr}(W\mathsf{B})dt, \quad \int \operatorname{tr}(WS)dt = 0, \quad V_{\text{noise}} = \int \operatorname{tr}(W^2\mathsf{B})dt.$$

Since $E = (\mathrm{I} - \varepsilon J)H_\varepsilon/[k(1+h^2)]$, subtract six times the zero equation and set

$$T = D\mathsf{B} - 6S = D[-a_0\mathsf{B} + b_0\varepsilon\,\mathrm{sym}(J\mathsf{B})], \quad b_0 = 6/[k(1+h^2)], \quad a_0 = b_0 - 1.$$

For $h \leq 1$, $2 \leq b_0 \leq 12$ and $b_0/a_0 \leq 2$. Diagonalizing B shows $T \preceq 0$ wherever

$$\chi(t) = \frac{2\sqrt{\det \overline{Q}}}{\sqrt{(\operatorname{tr}\overline{Q})^2 - 4\det\overline{Q}}} \geq 2h.$$

On $\mathcal{A}_h = \{\chi < 2h\}$, matrix Cauchy–Schwarz gives

$$\operatorname{tr}(WT) \leq 12h\sqrt{\operatorname{tr}\mathsf{B}}\sqrt{\operatorname{tr}(W^2\mathsf{B})}, \quad \Gamma \leq 12h\sqrt{\int_{\mathcal{A}_h}\operatorname{tr}\mathsf{B}\,dt}\sqrt{V_{\text{noise}}}.$$

Here the expression simplifies to

$$\chi = \frac{D^{-1} - D}{2} = f(1 + D^{-1}) \geq 2f \geq \min(t, 1-t).$$

Thus $\mathcal{A}_h$ has length at most $4h$. Writing $u = r^2 \in [0,1]$, we have $\operatorname{tr} Q = 2u + 8u/(1+u)^2 \leq 4$, since this expression is increasing in u. Consequently $\operatorname{tr}\mathsf{B} = k^2(1+h^2)\operatorname{tr} Q \leq 18$. For $0 < h \leq 1/4$,

$$\mathcal{V}_W \geq V_{\text{noise}}/\Gamma^2 \geq \frac{1}{10368h^3}.$$

This conservative constant suffices for the lower exponent; it is not a claim of an optimal variance constant.

C.3 Attainment and importance correction

Away from measure-zero spectral boundaries, the optimal time metric is $W_h = \mathrm{I} + (\kappa_h - 1)P_-$, where P_- projects onto the negative eigendirection of S. Its active set consists of the two endpoint bands of width asymptotic to $h/2$. If $\tau_Q = \operatorname{tr} Q$ and $\ell_Q = \lambda_1(Q) - \lambda_2(Q)$, direct two-dimensional spectral algebra gives

$$\operatorname{tr}(P_- H_\varepsilon H_\varepsilon^T) = \tfrac{1}{2}k^2 s(s\tau_Q - \ell_Q).$$

With $t = hy$, $\lambda_1(Q) = 4 + O(h)$ and $\lambda_2(Q) = 4h^2y^2 + O(h^3)$. Hence the last expression divided by h^2 tends to $k(0)^2(1 + 4y^2)$ on $0 \le y < 1/2$, and similarly at the right endpoint. Integration gives

$$\int_{\text{active}} \operatorname{tr}(P_- H_\varepsilon H_\varepsilon^T)dt \sim \tfrac{5}{3}h^3, \qquad \Gamma_h \longrightarrow \Gamma_* = A_0 + \tfrac{5}{2}G_0 = \tfrac{107}{7} - \tfrac{13\pi}{4}.$$

Using $W_h^2 = \mathrm{I} + (\kappa_h^2 - 1)P_-$ and $\kappa_h \sim (3G_0/2)h^{-3}$,

$$V_{\text{noise},h} \sim \tfrac{15G_0^2}{4}h^{-3}, \qquad \mathcal{V}_{\text{time},h} \sim C_T h^{-3}, \quad C_T = \frac{15G_0^2}{4\Gamma_*^2} \approx .2819575421.$$

The remaining score term in (22) is lower order: on the active bands both $\|P_- H_\varepsilon\|_F$ and $\|P_- E\|_F$ are $O(h)$, so $\left\|\operatorname{sym}(H_\varepsilon^T W_h E)\right\|_F = O(h^{-1})$. The Gaussian quadratic-form identity implies $\mathbb{E}(F^T W_h e)^2 = O(h^{-1})$. Outside these bands all coefficients are bounded.

For (15), importance correction leaves $\mathbb{E}_{\pi_h}(\cdot/\pi_h) = \mathbb{E}_{dt}(\cdot)$, including the population objective and Γ_h. The active bands are eventually contained in its endpoint intervals, where $1/\pi_h \sim 2h/\eta$; elsewhere $1/\pi_h$ is bounded. Thus

$$\mathcal{V}_{\text{importance},h} \sim (2/\eta)C_T h^{-2}.$$

For an arbitrary positive time proposal π independent of endpoints, write its effective uniform-measure weight as W. The state-weight lower-bound argument instead yields

$$\Gamma/(3/2) \le h\sqrt{V_{\text{noise},\pi}}\sqrt{\int \pi D^2 \operatorname{tr} Q\, dt}.$$

Since $D^2 \operatorname{tr} Q \le 4$, $\mathcal{V} \ge 1/(9h^2)$. This completes Theorem 3. State-dependent proposals, correlated reuse of endpoint pairs, arbitrary estimating equations, and an oracle returning zero are outside this comparison. The theorem gives fixed-h asymptotic variances and their small-h orders; it does not assert a joint-limit or finite-n risk bound.

The raw ratio can have problematic finite-sample tails. Clipping to a fixed compact parameter interval bounds endpoint moments in this globally well-posed affine family, but does not turn the asymptotic statement into a finite-sample rate.

D Liouville instrumentation and same-sample estimation

D.1 Proof of Theorem 4

Let $\Gamma_t = \mathbb{E}[UX_t^T]$. The raw moment polynomial satisfies $S_t' = \Gamma_t + \Gamma_t^T$. Positive definiteness on the compact interval implies a uniform inverse bound. Therefore

$$\mathbb{E}[(S_t^{-1}X_t)^T B(t) X_t] = \operatorname{tr} B(t), \qquad \mathbb{E}[(S_t^{-1}X_t)^T U] = \operatorname{tr}(S_t^{-1}\Gamma_t) = \tfrac{1}{2}\partial_t \log\det S_t.$$

The generated second moment obeys $\Sigma_t' = B\Sigma_t + \Sigma_t B^T$, so $\partial_t \log\det\Sigma_t = 2\operatorname{tr} B$. Integrating and using $\Sigma_0 = S_0$ proves the identity. Finite endpoint moments and integrable B justify the pairings. In the squeeze class, $S_1 = \mathrm{I}$ and (7) gives $-2b\theta$.

This is a homogeneous-linear-field result: a nonzero affine offset introduces an additional mean term. A determinant does not identify an arbitrary endpoint law. Identification of the squeeze target follows only because its determinant equals the target's precisely at the unique correct parameter.

D.2 Empirical root and its nuisance term

With fresh iid tuples (T_i, X_{0i}, Y_i), define $X_i = (1-T_i)X_{0i}+T_iY_i$, $U_i = Y_i - X_{0i}$, $\widehat{S} = n^{-1}\sum Y_iY_i^T$, and $S_t(S) = (1-t)^2\mathrm{I} + t^2 S$. The empirical instrument root is

$$\widehat{\theta}_{\mathrm{IV}}(\varepsilon) = -\frac{n^{-1}\sum_i X_i^T S_{T_i}(\widehat{S})^{-1}(C_0(T_i)X_i - U_i)}{n^{-1}\sum_i X_i^T S_{T_i}(\widehat{S})^{-1} H_\varepsilon(T_i) X_i}.$$

Define it arbitrarily if the denominator is zero; that event has probability tending to zero uniformly over the stated ε interval. Gaussian $\widehat{S}$ is positive definite almost surely for $n \geq 2$.

The population numerator at $S = \mathrm{I}$ is zero, and its derivative in a symmetric direction Δ is $\mathrm{tr}(K\Delta)$, with

$$K = -\int_0^1 \frac{t^2}{D(t)} \operatorname{sym} E(t) dt = \operatorname{diag}\big(1/2, (\pi-3)/2\big). \tag{23}$$

Set

$$\zeta_i = \frac{X_i^T(C_0(T_i)X_i - U_i)}{D(T_i)} + \mathrm{tr}\{K(Y_iY_i^T - \mathrm{I})\}.$$

Then

$$\sup_{|\varepsilon|\leq 1} \left| \sqrt{n}\widehat{\theta}_{\mathrm{IV}}(\varepsilon) - \frac{1}{2b\sqrt{n}} \sum_{i=1}^n \zeta_i \right| \xrightarrow{p} 0. \tag{24}$$

To prove this, on $\|S - \mathrm{I}\|_{\mathrm{op}} \leq 1/2$ we have $S_t(S) \succeq \mathrm{I}/4$ uniformly in time. First and second inverse-moment derivatives of the numerator have envelopes bounded by a constant times $\|(X_0, Y)\|^2$, with finite Gaussian moments. Since $\widehat{S} - \mathrm{I} = O_p(n^{-1/2})$, the sample Taylor remainder is $O_p(n^{-1})$. Replacing the sample derivative by its expectation also costs $O_p(n^{-1})$. The denominator is $A_n + \varepsilon B_n$, where after the plug-in $A_n \to -2b$ and $B_n \to \int k\,\mathrm{tr}(JE) = 0$ in probability. It therefore converges uniformly on $|\varepsilon| \leq 1$ to $-2b$, bounded away from zero. Uniform inversion proves (24). In particular the nuisance term is correlated with the tuple score and cannot be treated as independent.

For an explicit calculation of its variance, let $L_t = [(1-t)\mathrm{I}\ \ t\mathrm{I}]$, $Q_t = [-\mathrm{I}\ \ \mathrm{I}]$, and

$$B_t = D^{-1}\operatorname{sym}(L_t^T C_0 L_t - L_t^T Q_t), \qquad K_4 = \operatorname{diag}(0_{2\times 2}, K).$$

With $Z = (X_0^T, Y^T)^T \sim N(0, \mathrm{I}_4)$, $\zeta = Z^T(B_T + K_4)Z - \mathrm{tr}\,K$. Its mean is zero, and Gaussian quadratic-form moments give

$$V_\zeta = \int_0^1 \{2\,\mathrm{tr}[(B_t + K_4)^2] + (\mathrm{tr}\,B_t)^2\}\,dt = 7\pi - \frac{\pi^2}{2} - \frac{23}{3} \approx 9.3896797079.$$

The iid CLT therefore gives $V_{\mathrm{IV}} = V_\zeta/(4b^2) = 50.9710897$, independently of $\varepsilon \in [-1, 1]$. The expansion also covers deterministic ε_n sequences after variance standardization.

The direct comparator $\widehat{\theta}_{\mathrm{direct}} = -\log\det\widehat{S}/(4b)$ has first-order influence $-\mathrm{tr}(YY^T - \mathrm{I})/(4b)$. Since $\mathrm{Var}(\|Y\|^2) = 4$ in two dimensions, its asymptotic variance is $1/(4b^2)$, equal to 5.4284162 for every stated tilt. Neither CLT asserts convergence of raw ratio second moments or a finite-sample endpoint-risk guarantee. The target is known in this construction; the available constant-zero rule remains the strongest comparator in this known-target problem.

E Elliptic correct sets and coupling–schedule complementarity

Let the source ν be radial in two dimensions with covariance $m_2\mathrm{I}$, $m_2 > 0$, and target $c_\#\nu$, $c > 0$. Consider

$$v_\theta(t, x) = a(t)x + g(t)(P + \theta J)x, \qquad \int a = \log c, \quad G = \int g \neq 0, \tag{25}$$

where $a \in L^2$, g is bounded and P is fixed symmetric traceless with $P^2 = p^2\mathrm{I}$, $p > 0$.

Weighted regressions below satisfy the finite-risk and positive-Gram admissibility conditions.

Proposition 5 (Whole-correct-set exclusion and complementarity). *In (25), the exact correct set is $G^2(\theta^2 - p^2) = k^2\pi^2$, $k \geq 1$. Under independent radial endpoints, every admissible positive-definite time-only matrix or positive state-dependent scalar weighted fit satisfies $|\theta_W| < p$. Under uniform time, every coupling of the fixed source and target selects $|\theta_{\mathrm{CFM}}| < \max\{p, \pi/|G|\} < \sqrt{p^2 + \pi^2/G^2}$; its exact attained maximum is given in Proposition 7. If g is continuous and nonzero at $t_* = 1/(1+c)$, for each correct parameter there exists a positive time density and rotational coupling that jointly select it.*

E.1 The correct set and its exclusion under independent endpoints

Let $A = P + \theta J$. Since $P^2 = p^2\mathrm{I}$, $J^2 = -\mathrm{I}$, and $PJ + JP = 0$, $A^2 = (p^2 - \theta^2)\mathrm{I}$. The coefficient matrices $a(t)\mathrm{I} + g(t)A$ commute at different times, despite the noncommuting basis matrices. The terminal map is cB, where $B = e^{GA} = C\mathrm{I} + SA$, and

$$S = \begin{cases} \sinh(G\sqrt{p^2 - \theta^2})/\sqrt{p^2 - \theta^2}, & |\theta| < p, \\ G, & |\theta| = p, \\ \sin(G\sqrt{\theta^2 - p^2})/\sqrt{\theta^2 - p^2}, & |\theta| > p. \end{cases}$$

Direct multiplication gives $\det B = 1$, $\operatorname{tr} BB^T = 2 + 4p^2S^2$. Radial invariance implies that every orthogonal B preserves ν; conversely, preservation forces $BB^T = \mathrm{I}$ by covariance equality. Thus the endpoint is correct iff $S = 0$, proving the displayed winding set in Proposition 5.

Independent endpoints give an $O(2)$-equivariant Bayes field, so $u(t,x) = \alpha(t, \|x\|)x$. In the rotational normal equation, the radial residual is orthogonal to Jx pointwise for scalar weights. The remaining term is

$$wg^2x^T(J^TP + \theta\mathrm{I})x.$$

For $\theta \geq p$ it is nonnegative, and strictly positive on positive measure: at equality its zero set outside the origin is one line, which has zero mass under a radial nonzero law. Its integral is therefore positive. The argument for $\theta \leq -p$ is symmetric. For time-matrix weights, isotropy makes the radial pairing proportional to $\operatorname{tr}(J^TW) = 0$. The remaining pairing is proportional to $\operatorname{tr}[W(PJ^T + \theta\mathrm{I})]$, again strictly positive at $\theta \geq p$. Every selected parameter therefore lies in $(-p, p)$. For time-scalar weights the normal equation is exactly a positive multiple of θ, giving zero.

E.2 A sharp weighted finite-interval Carlson inequality

This weighted form of Carlson's integral inequality has strictness and sharpness properties used below. See Larsson et al. (2005) for related bounded-interval inequalities.

Lemma 6. *Fix $c > 0$. For $\lambda > 0$, define*

$$D_\lambda(t) = \lambda((1-t) - ct)^2 + \frac{t(1-t)}{\lambda c}, \qquad I_\lambda = \int_0^1 D_\lambda(t)^{-1}\,dt.$$

Then $I_\lambda < \pi$. For bounded measurable g with $G = \int_0^1 g \neq 0$, put

$$M = \int_0^1 g^2((1-t) - ct)^2\,dt, \qquad I_1 = \int_0^1 g^2 t(1-t)\,dt.$$

Both are positive, and

$$\left(\int_0^1 |g|\right)^2 < 2\pi\sqrt{MI_1/c}, \qquad \frac{G^4}{\pi^2} < \frac{4MI_1}{c}. \tag{26}$$

For each fixed c, both constants are sharp; neither strict inequality is attained.

Under $s = \log(ct/(1-t))$, the transformed profile $f(s) = g(t)t(1-t)$ satisfies $\int |f| = \int |g|$, $\int f^2 = I_1$, and $M/c = \int 4\sinh^2(s/2)f^2$. Comparison with s^2 explains the classical constant.

Proof. The two nonnegative terms in D_λ vanish at disjoint points, so it is strictly positive on $[0,1]$. With $\delta = (\lambda c)^{-1}$, substitute $y = ct/(1-t)$, then $y = e^s$, to obtain

$$I_\lambda = \delta\int_0^\infty \frac{dy}{(y-1)^2 + \delta^2 y} = \delta\int_{\mathbb{R}} \frac{ds}{\delta^2 + 4\sinh^2(s/2)} < \delta\int_{\mathbb{R}} \frac{ds}{\delta^2 + s^2} = \pi.$$

Strictness follows from $4\sinh^2(s/2) > s^2$ for $s \neq 0$. Since g is nonzero on a set of positive measure and the moment multipliers vanish only at finitely many points, $M, I_1 > 0$. Weighted Cauchy–Schwarz gives

$$\left(\int |g|\right)^2 \leq \left(\lambda M + \frac{I_1}{\lambda c}\right) I_\lambda < \pi \left(\lambda M + \frac{I_1}{\lambda c}\right).$$

Choose $\lambda = \sqrt{I_1/(cM)}$ and use $|G| \leq \int |g|$ to prove (26).

For sharpness, take $g_\lambda = D_\lambda^{-1}$, which is positive and bounded for every finite λ. Rescaling $s = \delta u$ gives

$$I_\lambda = \int_{\mathbb{R}} \frac{du}{1 + 4\sinh^2(\delta u/2)/\delta^2} \longrightarrow \int_{\mathbb{R}} \frac{du}{1+u^2} = \pi.$$

Dominated convergence applies because the integrand is bounded by $1/(1+u^2)$. Write $A_\lambda = \lambda M$, $B_\lambda = I_1/(\lambda c)$, with moments evaluated at g_λ. The identity $g_\lambda^2 D_\lambda = g_\lambda$ and (26) give

$$G = I_\lambda, \qquad A_\lambda + B_\lambda = I_\lambda, \qquad I_\lambda^4/\pi^2 < 4A_\lambda B_\lambda \leq I_\lambda^2.$$

Thus $(A_\lambda - B_\lambda)^2 \to 0$, and

$$G \to \pi, \qquad \lambda M \to \pi/2, \qquad \frac{I_1}{\lambda c} \to \pi/2, \qquad \frac{cG^4}{4\pi^2 M I_1} \to 1. \tag{27}$$

Because $g_\lambda > 0$, these limits prove sharpness of both inequalities. □

E.3 The exact interval selected by couplings

For a positive probability density $w(t)$ on $[0,1]$, define

$$G_w = \int wg, \qquad I_{0,w} = \int wg^2((1-t)^2 + c^2t^2), \qquad I_{1,w} = \int wg^2 t(1-t),$$
$$M_w = I_{0,w} - 2cI_{1,w} = \int wg^2((1-t) - ct)^2 > 0, \qquad L_w = 2cI_{1,w}.$$

The unweighted moments are obtained at $w = 1$, with subscripts omitted.

Proposition 7 (Exact coupling–schedule criterion). *In (25), let $w > 0$ almost everywhere, $\int w = 1$, and $\int wa^2 < \infty$. Draw time independently of the endpoints with density w, without importance correction. Every endpoint coupling has a unique population CFM minimizer, and the set of selected parameters is exactly $[-\Theta_w, \Theta_w]$, where*

$$\Theta_w = \max\left\{ \underbrace{\frac{2pcI_{1,w}}{I_{0,w}}}_{\Theta_{\mathrm{refl},w}}, \underbrace{\frac{c|G_w|}{\sqrt{M_w(M_w + 4cI_{1,w})}}}_{\Theta_{\mathrm{rot},w}} \right\}. \tag{28}$$

Every value in this interval is attained by a deterministic orthogonal coupling. A correct parameter $\pm\theta_k$, where $k \geq 1$ and $\theta_k = \sqrt{p^2 + k^2\pi^2/G^2}$, is selectable if and only if

$$\Theta_{\mathrm{rot},w} \geq \theta_k. \tag{29}$$

If it is selectable, a pure rotational coupling suffices.

For uniform time, write $\Theta = \Theta_1$. Then

$$\Theta < \max\{p, \pi/|G|\} < \theta_1, \qquad \theta_1^2 - \Theta^2 > \min\{p^2, \pi^2/G^2\}. \tag{30}$$

For fixed c, p, the simpler uniform-time ceiling is sharp:

$$\sup_{g,\gamma} \frac{|\theta_{\mathrm{CFM}}(\gamma)|}{\max\{p, \pi/|G|\}} = 1,$$

over bounded measurable g with nonzero integral, and the supremum is not attained. Sharpness also holds after fixing any positive sup norm of g, or separately any nonzero value of G.

Proof. Normalize $K = m_2^{-1}\mathbb{E}[X_0 Y^T]$; Cauchy–Schwarz gives $\|K\|_{\mathrm{op}} \le c$. Write $K = \alpha \mathrm{I} + \zeta J + Q$ with Q symmetric traceless, and set $S = J^T P$, $\tau = \mathrm{tr}(SQ) = \mathrm{tr}(SK)$. Direct moments give

$$m_2^{-1}\mathbb{E}[X_t X_t^T] = ((1-t)^2 + c^2 t^2 + 2\alpha t(1-t))\mathrm{I} + 2t(1-t)Q,$$
$$\mathbb{E}[(JX_t)^T U] = -2m_2\zeta, \qquad \mathbb{E}[(JX_t)^T P X_t] = 2m_2 t(1-t)\tau.$$

The $a(t)x$ term is orthogonal to Jx, so the normal equation is

$$\theta_w(K) = -\frac{\mathcal{N}_w(K)}{\mathcal{D}_w(K)}, \qquad \mathcal{N}_w(K) = \zeta G_w + \tau I_{1,w}, \quad \mathcal{D}_w(K) = I_{0,w} + 2\alpha I_{1,w} \ge M_w > 0. \tag{31}$$

The assumptions ensure finite risk; its quadratic coefficient is $2m_2\mathcal{D}_w(K) > 0$. The numerator is linear and denominator affine in K.

Take an SVD $K/c = U\,\mathrm{diag}(s_1, s_2)V^T$, $0 \le s_i \le 1$. The orthogonal matrices $O_\eta = U\,\mathrm{diag}(\eta_1, \eta_2)V^T$, $\eta_i \in \{-1, 1\}$, with probabilities $\omega_\eta = (1 + \eta_1 s_1)(1 + \eta_2 s_2)/4$, satisfy $K = \sum_\eta \omega_\eta c O_\eta$. Each cO_η is realized by $Y = cO_\eta^T X_0$, since ν is radial. Mixing these couplings independently of X_0 realizes every K in the ball. Moreover,

$$\theta_w(K) = \sum_\eta \frac{\omega_\eta \mathcal{D}_w(cO_\eta)}{\mathcal{D}_w(K)} \theta_w(cO_\eta).$$

The coefficients are nonnegative and sum to one, so it suffices to bound rotations and reflections.

For a reflection O, $\alpha = \zeta = 0$ and

$$|\theta_w(cO)| = \frac{cI_{1,w}|\,\mathrm{tr}(SO)|}{I_{0,w}} \le \frac{2pcI_{1,w}}{I_{0,w}}.$$

Frobenius Cauchy–Schwarz gives the trace bound, since $\|S\|_F = \sqrt{2}p$, $\|O\|_F = \sqrt{2}$. The two signs are attained at $O = \pm S/p$. For rotations,

$$\theta_w(cR_\varphi) = -\frac{cG_w \sin\varphi}{I_{0,w} + L_w \cos\varphi}.$$

The identity

$$(I_{0,w} + L_w\cos\varphi)^2 - (I_{0,w}^2 - L_w^2)\sin^2\varphi = (I_{0,w}\cos\varphi + L_w)^2$$

gives maximum magnitude $c|G_w|/\sqrt{I_{0,w}^2 - L_w^2}$, attained with both signs when $G_w \ne 0$ at angles satisfying $\cos\varphi = -L_w/I_{0,w}$. If $G_w = 0$, the rotational image is $\{0\}$. Since $I_{0,w}^2 - L_w^2 = M_w(M_w + 4cI_{1,w})$, this is the second term of (28).

Each component of $O(2)$ is connected. Continuity and both signed extrema show that reflections and rotations each cover their full symmetric interval; their union is $[-\Theta_w, \Theta_w]$. Hence every selected value has a deterministic realization $Y = cO^T X_0$. The reflection maximum is strictly below $p < \theta_k$, so $\Theta_w \ge \theta_k$ is equivalent to $\Theta_{\mathrm{rot},w} \ge \theta_k$. The rotational interval then proves (29). This statement uses the unweighted G for correctness and G_w for selection.

Now take $w = 1$. Lemma 6 gives

$$M(M + 4cI_1) > 4cMI_1 > c^2G^4/\pi^2,$$

so $\Theta_{\mathrm{rot}} < \pi/|G|$, while $\Theta_{\mathrm{refl}} < p$. Equation (30) follows.

Along $g_\lambda = D_\lambda^{-1}$, equation (27) yields

$$\frac{\Theta_{\mathrm{refl}}}{p} = \left(1 + \frac{M}{2cI_1}\right)^{-1} \to 1, \qquad \frac{\Theta_{\mathrm{rot}}}{\pi/|G|} = \sqrt{\frac{cG^4}{4\pi^2 M I_1}}\left(1 + \frac{M}{4cI_1}\right)^{-1/2} \to 1.$$

Thus the ratio to the larger ceiling tends to one, and the strict bound prevents attainment. Both displayed ratios are invariant under $g \mapsto bg$, $b \ne 0$. Rescale each profile to a prescribed positive sup norm, or separately to a prescribed nonzero integral. This also proves sharpness at every fixed positive $p|G|/\pi$, including the balanced case. It does not assert sharpness under arbitrary simultaneous constraints on the norm and integral. □

The fixed-profile separation concerns selected parameter magnitude, not a new endpoint-distance lower bound.

E.4 A JOINT INTERVENTION FOR EACH WINDING

Suppose g is continuous and $g(t_*) \neq 0$, $t_* = 1/(1+c)$. Choose positive bounded approximate-identity densities concentrating at t_*, for example normalized truncated Gaussians. Each satisfies $\int wa^2 < \infty$, and

$$G_w \to g(t_*) \neq 0, \qquad I_{1,w} \to g(t_*)^2 t_*(1-t_*) > 0, \qquad M_w \to 0.$$

Thus $\Theta_{\mathrm{rot},w} \to \infty$, and (29) selects any prescribed signed winding with a suitable density and rotation. The class and its unweighted integral G are fixed. Independent coupling has $K = 0$ and selects zero under every such density. Each fixed density has finite Θ_w, so the statement is per winding, not one density selecting all windings.

F EXPERIMENTAL METHODS AND EVIDENCE SCOPE

Monte Carlo standard errors (MCSEs) quantify numerical integration uncertainty conditional on fitted models unless stated otherwise.

F.1 CORRECT-START NEURAL DIAGNOSTIC

The network is $v(t,x) = W_s^{\mathrm{net}} x + b_s + W_3^{\mathrm{net}}\,\mathrm{SiLU}(W_2^{\mathrm{net}}\,\mathrm{SiLU}(W_1^{\mathrm{net}}[x;t] + b_1) + b_2) + b_3$, with two width-128 hidden layers. The teacher uses only the first 100 units: diagonal weights 1.5, 1, 0.6 in the three layers, all biases and time weights zero, and skip weights -0.45 on the diagonal and 0.15 at cyclic predecessor coordinates. The stored float32 coefficients define the parameter state; evaluation promotes them to float64. Heun-8 defines the target law and deployment map.

Training uses Adam at 0.001, batch 256, norm cap 10, no weight decay, and fresh population target samples each update. Reported CFM risks average squared residuals over coordinates. Training-generated targets use float32 arithmetic; evaluation uses float64 with the same stored weights. Precision comparisons below assess this implementation difference.

The pilot seed is 9100; confirmations are 9200 and 9201. Five observables are coordinate averages of x, x^2, x^3, $\cos(2x)$, and $x_i x_{i-1}$. The pilot selects the earliest checkpoint with a negative risk-change 99% interval and any nonzero observable-change interval, then the largest absolute mean/SE witness. The selected count is one and the observable is the mean, with positive direction; both choices were fixed before confirmation. Each confirmation requires the same direction and risk decrease.

Training generators and NumPy evaluation streams are distinct. Risk and law streams are separately seeded, with 32 blocks of 1,024 examples. For block differences d_b, the interval is $\bar{d} \pm t_{31,0.995} s_d/\sqrt{32}$. Its approximate coverage concerns a fixed fitted model. The reported confirmation intervals are

	Risk change	Mean change
9200	$[-0.00180289, -0.00146812]$	$[0.01808820, 0.01831308]$
9201	$[-0.00190505, -0.00163987]$	$[0.01501735, 0.01516142]$.

Precision controls compare float32/64, while Heun-16/32 and RK4-64 comparisons describe sensitivity to changing the sampler. They use 128 independent inputs and are not uniform error certificates. The fixed-Heun-8 claim does not require that map to equal the exact ODE flow. The pilot at 10, 100 and 300 updates also has lower risk and nonzero mean changes, but the changes are nonmonotone and those later counts lack fresh confirmation. For the coordinate-averaged mean, $\mathrm{Lip}(\phi) = 1/\sqrt{100}$, hence $W_2 \geq W_1 \geq 10|\Delta \mathbb{E}\phi|$ for these finite-second-moment laws. Transforming the positive mean-interval endpoints gives approximate conditional 99% lower bounds 0.18088 and 0.15017 on unnormalized Euclidean W_2, not estimates of the full distance. The four nominal 99% intervals have a Bonferroni joint-coverage lower bound of 96% only if their approximate marginal calibration holds. They do not estimate a success probability over teachers or training seeds.

F.2 CORRECT-START REFITTING OF A PRETRAINED IMAGE GENERATOR

The starting model is the released TorchCFM CIFAR-10 independent-CFM U-Net, using its 400,000-step EMA weights. The pinned architecture has base width 128, two residual blocks per level, channel multipliers $(1, 2, 2, 2)$, attention at resolution 16 and dropout 0.1. The source is standard Gaussian in

$\mathbb{R}^{3\times32\times32}$; 100 left Euler steps on the native uniform grid define the raw target law and deployment. The learner starts at exactly the weights defining that target. This changes the original real-image target to the teacher's deployed law; it is not ordinary continuation on CIFAR-10. The pinned native objective uses zero smoothing and independent straight-line pairs.

Following an independent pilot, the two confirmation streams use fresh Adam with learning rate 2×10^{-5}, betas $(0.9, 0.999)$, batch 128 and gradient-norm cap 1. Training has dropout active and consumes fresh generated targets once; reported risk and deployment use eval mode. The first update is a small fresh-Adam step, approximately sign-normalized, not an exact population-gradient step. Target, training, risk and endpoint seeds are 301–304 and 311–314. The pilot fixes both primary counts, one and 200, and separate unit-norm feature directions for them before confirmation.

Risk averages squared velocity residuals over coordinates on 16,384 fresh triples, paired with the teacher on identical triples. Endpoint evaluation uses 10,000 common Gaussian sources and the observable $\phi_j(x) = d_j^T F(\mathrm{image}(x))$, where F is the pinned 2,048-dimensional Inception feature map and d_j is its direction selected in the pilot and fixed before confirmation. Native image conversion clips $127.5x + 128$ to $[0, 255]$ and truncates to uint8. The expectation difference of this fixed observable witnesses different deployed laws; paired image displacement alone would not. Its units are feature projections, not a calibrated distributional distance.

Table 3: Pretrained correct-start confirmations. Means and nominal 99% paired t intervals, conditional on each fitted model. Risk changes use an explicitly **reconstructed** teacher baseline; endpoint projections are retained observations. Risk entries are multiplied by 10^5.

Stream	Updates	Risk change $\times10^5$	Observable change
001	1	-3.288 [-3.785,-2.790]	0.0884 [0.0821,0.0947]
001	200	-58.074 [-65.353,-50.794]	0.8129 [0.7733,0.8525]
002	1	-4.030 [-4.833,-3.227]	0.0779 [0.0695,0.0864]
002	200	-55.054 [-60.710,-49.397]	0.7639 [0.7310,0.7969]

The confirmation criterion requires all eight primary 99% intervals to have the stated signs. Their Bonferroni joint-coverage lower bound is 92% under the approximate interval model. A supplementary correction to 99.875% individual intervals gives nominal 99% simultaneous coverage across all eight, with every sign unchanged. Independently reproduced 2,000-resample bootstrap intervals agree in sign; these checks do not turn mean inference into a distribution-free guarantee. In particular the one-update changes are small and are not evidence of a practically important image-quality loss.

Reconstruction and evidence scope. The original evaluator did not retain teacher per-triple risk arrays. They were reconstructed from the recorded evaluator and seeds. Combining each reconstructed paired delta with its baseline reproduces the retained student arrays byte for byte for all eight stream/checkpoint combinations at steps 1, 10, 100 and 200; baseline summaries also match. Table 3 uses these reconstructed risks. Independent CPU checks reproduced the reported summaries, simultaneous intervals and bootstrap results from loss/projection vectors and checked their source/checkpoint associations. GPU generation and feature extraction were not rerun. The array agreement supports reconstruction without proving identity of all intermediate input tensors.

Baseline risk includes reference-velocity approximation error and conditional label variance. This experiment concerns finite updates of one pretrained checkpoint under Euler-100, with dropout during training and eval-mode measurement. It does not establish convergence behavior or prevalence across architectures. Recorded total worker time was 12.62 hours, including evaluation-only reconstruction and non-GPU work.

F.3 Reachable near-manifold family and structure-aware likelihood

Family, target and exact deployed membership. Let $d = 100$, $r = 8$, $s = 92$, $N = 16$, and $x = Q(z, w)$ for a fixed orthogonal Q. The source is standard Gaussian, equivalently in these rotated coordinates. Equation (19) has eight coefficients $a_i \in [0, 3]$, one $k \in [0, 4]$, and $C \in \mathbb{R}^{92\times73}$. Its 73 fixed features consist of a constant, eight coordinates and 64 sine features. The architectural rotation, frequencies and phases are known to both fitting procedures. The teacher coefficients and all evaluator diagnostics are withheld from ordinary fitting.

The public basis is generated with seed 240924: QR factorization of a standard-Gaussian 100×100 matrix supplies Q, frequency entries are Gaussian with standard deviation $1/\sqrt{8}$, and phases are uniform on $[0, 2\pi]$. The teacher has

$$a_* = (1.5, 1.5, 0.12, 0.16, 0.20, 0.24, 0.28, 0.32), \qquad k_* = 2.2.$$

Its drive coefficients, generated with seed 91731, are independent Gaussians with standard deviation $1.2/\sqrt{73}$, except that the constant-drive column is zero. The source implementation pins the Torch CPU random streams and stores the generated basis and parameters; these arrays resolve library-specific QR and random-number conventions.

An Euler step updates

$$z_{j+1,i} = z_{j,i} + \frac{a_i}{16}\tanh(2z_{j,i}), \qquad w_{j+1} = \left(1 - \frac{k}{16}\right) w_j + \frac{1}{16} C\psi(z_j).$$

Every latent step is an increasing bijection, with derivative at least one. Every residual multiplier is at least 0.75. Write T_a for the separable latent endpoint map. Solving the residual recursion and expressing the initial latent point as $T_a^{-1}(z)$ gives

$$w = m_\theta(z) + \sigma_\theta E, \qquad E \sim N(0, \mathrm{I}_s), \qquad \sigma_\theta = (1 - k/16)^{16},$$

with E independent of the terminal latent coordinate. This proves membership of the teacher in the *actual Euler-16* family, not only in a continuous ODE class. It also gives a full-dimensional density around a smooth eight-dimensional graph. At the teacher, $\sigma_* = 0.0937876513$, the complementary-noise RMS is 0.89958, and $\mathbb{E}\,\mathrm{dist}(X, \mathcal{M})^2 \le .80924$. Noise lies in a fixed complement, not necessarily the geometric normal bundle.

The first two latent coordinates form four strong sign-defined groups, with possible weaker bimodality in other coordinates. The drive rank is 72, leaving 20 complementary linear combinations with only Gaussian noise; the graph's linear span has dimension at most 80. A finite-difference check projects terminal-coordinate graph acceleration onto the normal space. At 24 fresh points its median section curvature is 0.219189 for step sizes 0.002 and 0.001. This is post-hoc numerical evidence of curvature, not a uniform bound.

Fitting and finite-data design. Confirmation uses seeds 201–205, one fixed teacher and five independent datasets. Each has 32,768 target observations, 262,144 independent-pair CFM tuples, 131,072 fresh risk pairs, and 8,192 fresh generated and predictive-evaluation observations. Two earlier pilots are excluded. CFM solves its constrained quadratic by profiling the shared contraction after linear feature projection and solving the eight scalar latent regressions. The common numerical diagonal jitter is 10^{-10} times the Gram diagonal mean. Thus the fit is to numerical tolerance, not claimed exact real-arithmetic minimization.

For the alternative, the latent endpoint density is evaluated by inverting each of the 16 monotone steps and accumulating its log derivative. Stage one fits a by marginal likelihood, with L-BFGS and a sigmoid parameterization of $[0, 3]$. Stage two profiles C by conditional least squares at each k, then searches $k \in [0, 4]$ by golden-section search, including the endpoints and a 17-point coarse-grid check. This is a *two-stage likelihood estimator*, not a claim of the joint finite-sample MLE or a general global optimizer. The conditional-only control retains the CFM latent coefficients. A smaller linear-drive ablation removes the sine features; nonlinear latent paths mean this need not yield a linear endpoint graph.

Both methods use the same target sample and known model information. Trial inverses in likelihood are computed from the current candidate, not supplied from the teacher. Independent evaluations and a fresh independently implemented same-size refit support the comparison. The latter has negligible CFM first-order residual and again selects $k \approx 1.339$, while likelihood selects $k \approx 2.201$.

Secondary endpoint screen. The secondary endpoint screen uses pulled-back latent Fourier moments, residual means, full residual covariance and mixed degree-one/two latent–residual moments. It takes the maximum statistic divided by its null-calibration median.

The diagnostic was revised after a failed pilot calibration, then fixed before confirmation. Twenty-four additional replicates of each stated alternative were detected, and one of 24 fresh null replicates was rejected. These counts have substantial binomial uncertainty and do not certify nominal size or power against all alternatives. The score and post-hoc KL estimates use different evaluation procedures.

Interpretable endpoint divergence. Every model within the stated bounds factorizes as $q_\theta(z, w) = q_{\theta,Z}(z)N(w; m_\theta(z), \sigma_\theta^2 \mathrm{I}_s)$. The relative-entropy chain rule and the Gaussian conditional formula give exactly

$$D_{\mathrm{KL}}(P_*\|Q_\theta) = D_{\mathrm{KL}}(p_{*,Z}\|q_{\theta,Z}) + \frac{s}{2}\left[\frac{\sigma_*^2}{\sigma_\theta^2} - 1 + \log\frac{\sigma_\theta^2}{\sigma_*^2}\right] + \frac{\mathbb{E}_{Z\sim p_{*,Z}}\|m_*(Z) - m_\theta(Z)\|^2}{2\sigma_\theta^2}. \tag{32}$$

The three terms are latent-law, conditional-scale and conditional-mean errors. This additive divergence accounting is not a causal decomposition of training failure. In particular, the mean term is not a pure curvature contribution.

After confirmation, an independent implementation evaluated (32) on 8,192 new latent probes (seed 24092481), analytically integrating out the 92-dimensional residual noise. A second implementation reproduced the calculation and checked the identity against full 100D log densities using conditional moment-exact cubature. Both use the same probe stream; their agreement is numerical verification, not another independent Monte Carlo replicate.

Table 4: Post-hoc KL accounting for five saved fits; nats per complete 100D observation. Only the scale term is free of latent integration error.

Method	Latent law	Scale	Conditional mean	Total
CFM	0.022451	49.674356	7.364754	57.061560
Two-stage likelihood	0.000087	0.000143	0.105027	0.105257
Conditional-only repair	0.022451	0.000470	0.150723	0.173643
Linear-drive ablation	0.000087	45.093595	38.685714	83.779396

Integration standard errors for the mean totals are 0.05002, 0.000750, 0.001845 and 0.18986, respectively. They average across fits within each common probe before estimating Monte Carlo error, and condition on the saved models. They are not across-training confidence intervals. Scale contributes 87.05% of CFM's KL. The likelihood improvement is consistent with the independent held-out predictive log-density gap, not merely with the calibrated diagnostic-score ratio.

Gaussian scale control and an analytic weighted comparator. Apply (3) with $\theta = -k$ and $\sigma = \sigma_*$. For the graph-free control, $k_{\mathrm{CFM}} = 1.4738417$ and Euler-16 gives standard deviation 0.2130554 rather than 0.09378765, although the correct $k_* = 2.2$ is available. For any thin target $0 < \sigma \le 1$, the selected k lies in $[0, 3/2)$, so its Euler-16 thickness exceeds $(29/32)^{16} = 0.2069990$. This scale mismatch occurs without graph curvature; it does not explain every error of the driven fit.

The population weight $1/D_{\sigma_*}(t)$ instead selects $k = -\log\sigma_* = 2.3667221$. Its exact ODE is correct, but Euler-16 gives 0.0772120, or 0.8233 times the target thickness. This is an analytic control, not a sample-estimated time-weighted driven-family run; no such run is included. In the bounded constant family $k \le 4$, reachable thickness is at least $(3/4)^{16} = 0.0100226$, so a diverging thin-target error ratio cannot be claimed while retaining reachability in that fixed family.

Temporal expansion with the target and Euler-16 fixed. Five further datasets (seeds 401–405), each with 32,768 target samples, compare constant CFM, temporal contraction alone, and all coefficients varying independently over 16 intervals. Each uses the same 524,288 regression pairs stratified equally across the intervals. A separate pilot (seed 301) is excluded. The original two-stage likelihood estimator is refit on the same target samples. The CFM parameter counts are 6,725, 6,740 and 107,600, respectively; the likelihood family remains the original 6,725-parameter family.

The expanded latent bounds are $[-6, 6]$, and contraction bounds are $[-8, 12]$. Every latent Euler derivative is at least $1/4$, and every residual multiplier at least $1/4$. Latent steps are onto and increasing, so the triangular density and KL identity remain valid. Repeating the teacher coefficients embeds exactly the unchanged target. No fitted bounds are active; the constant optimum lies inside its original box. The temporal expansion therefore also widens the coefficient ranges.

Full temporal CFM removes 88.6222% of the paired KL, a reduction of 50.5311 nats with five-fit 95% t interval [50.4970,50.5651] conditional on the common probes. Separate paired integration

Table 5: Temporal confirmation, five further fits. Risk is per coordinate; KL components are nats per complete 100D observation, on 16,384 common latent probes. The small negative latent estimate is retained Monte Carlo error.

Method	Risk	Latent	Scale	Mean	Total KL
Constant CFM	0.631241	0.020577	49.677704	7.320238	57.018519
Temporal contraction	0.597550	0.020577	31.984155	11.402661	43.407393
Full temporal CFM	0.458564	0.009672	0.194741	6.283045	6.487459
Original likelihood	0.965064	-0.000025	0.000097	0.103816	0.103888

MCSE is 0.03943 nats, conditional on the fits; these are not a joint interval. All five fits in each CFM family fail the unchanged secondary screen. The five further likelihood fits are not rejected, which neither proves equality nor removes the original study's rejected likelihood fit.

Independent evaluations gave temporal KL 6.5617 on 8,192 probes and 6.5114 on 32,768 probes, with MCSE 0.02028 and 0.00974; the 6.4875 estimate on the primary stream has MCSE 0.01389. These estimates use different probe streams. Fresh independent risk evaluation also preserved the ordering across the CFM families. The comparison establishes a substantial improvement from temporal freedom in this expanded family, not a population-CFM failure or a comparison of both objectives optimized over the same expanded family.

The Gaussian control also separates representation from deployment: its 16-bin population-CFM fit gives Euler-16 thickness 0.0714240, and left Euler-16 applied to the continuous reference velocity gives 0.0690672. Thus temporal freedom can change under-contraction into over-contraction at the fixed sampler.

Conditional endpoint design. The conditional fits reuse the same five temporal models. Hold their latent and contraction profiles fixed. With terminal latent z, let $z_j(z)$ be its candidate Euler latent path after inversion. The conditional mean is linear in the drive:

$$m_C(z) = \sum_{j=0}^{15} \frac{1}{16} \left[\prod_{\ell>j} (1 - k_\ell/16) \right] C_j \psi(z_j(z)).$$

Sample-only drive fitting with a norm or risk budget. The two sample-only fits reuse fields 401–405, the same 32,768 noisy target observations and known public features, holding all latent and contraction profiles fixed. They do not use teacher means for fitting or selection. Stack the 16 drive blocks into $C \in \mathbb{R}^{92\times 1168}$; the endpoint design above remains linear in C. Conditional least squares on the observed terminal (z, w) pairs is, at these fixed noise scales, conditional Gaussian likelihood fitting.

The norm-constrained fit minimizes this sample endpoint loss subject to $\|C\|_F \le b_C \|C_{\mathrm{CFM}}\|_F$, with $b_C \in \{0.5, 1, 2, 4, 8\}$. A separate 8,192-observation noisy validation set selects the smallest-norm candidate whose paired excess loss over the best is within one standard error; saved CFM is a fallback. This is a selection heuristic, not a coverage guarantee. The grid was fixed after a separate pilot; all five confirmations select cap 8. The mean norm is 443.40 versus about 55.42 for CFM. On its original 65,536-probe stream, KL falls from 6.51155 to 0.68233, a 5.82922-nat reduction with 95% interval across five fits, conditional on the shared evaluation probes, [5.81582,5.84262] and separate shared-probe MCSE 0.006249. The average fractions removed are 89.52% of total KL and 92.47% of raw conditional-mean squared error.

The risk-constrained fit reuses the original endpoint summaries and the CFM quadratic from 524,288 original regression pairs. At fixed profiles, complete that positive-definite drive quadratic about its conditional minimizer. Cholesky whitening turns the constraint into a Euclidean ball; the implementation solves the resulting endpoint least-squares trust-region problem by SVD and scalar dual bisection. The primary budget, set before the pilot, is

$$\widehat{R}(C) \le 1.1 \widehat{R}(C_{\mathrm{CFM}}).$$

No coefficient-norm cap is added. A control optimizes the endpoint loss along the segment from saved CFM to the selected norm-constrained fit under the same empirical risk budget. It reaches only

2.63–2.68% of that segment. This is a comparison to the specified segment, not to every possible interpolation rule. The risk-constrained fit makes no new validation choice and adds no learner target observations; its reference inherits the norm-constrained fit's validation selection. The reused validation set is descriptive only.

Table 6: Primary sample-only drive repairs on the same five temporal fits. All rows use the same 65,536 endpoint probes and fresh disjoint risk streams. KL is per 100D observation; risk is per coordinate. Norms are original-coordinate Frobenius norms; the risk-constrained fit's norm is measured, not constrained.

Method	Forward KL	Fresh risk	Mean drive norm
Saved temporal CFM	6.515038	0.458759	55.42
Selected norm cap 8	0.680174	65.029703	443.40
Empirical +10% risk budget	1.960148	0.504626	86.09
Matched +10% line blend	6.197991	0.504337	55.43

Endpoint evaluation uses 65,536 common latent probes (seed 9862701), conditioning on the saved models. Risk uses 131,072 fresh independent-CFM pairs per field, with target/source/time seed blocks starting at 9871000, 9872000, 9873000, 9874000 and 9875000. These disjoint blocks replace the inferential interpretation of the norm-constrained fit's historical overlapping-stream risk interval; they do not overwrite its point estimates. The unchanged norm-constrained fit has about 141.75 times baseline risk on this new evaluation.

The risk-constrained fit removes 4.55489 nats, 95% interval across five fits, conditional on the shared evaluation probes, [4.52694,4.58285], with shared-probe MCSE 0.005978. It retains 78.063% of the norm-constrained fit's same-stream endpoint gain, interval [77.728%,78.398%], and removes 69.913% of baseline KL; these are different denominators. Its advantage over the blend is 4.23784 nats, interval [4.21133,4.26436], MCSE 0.005710. Fractions average per-field ratios. Endpoint intervals condition on common evaluation probes, while the separate MCSE averages fields within each probe; neither combines all uncertainty or gives simultaneous coverage over comparisons.

The empirical cap is met to about 10^{-13}. Fresh mean risk rises by 0.0458674, five-field interval [0.0456632,0.0460715], with evaluation MCSE 0.00008424. The mean fresh ratio is 1.099981; individual ratios range from 1.099638 to 1.100414. These small overshoots are retained: empirical feasibility is not a strict population-risk guarantee. The fixed latent/scale contribution remains about 0.2075 nats. The teacher is available in the full model, generally not in this conditional subfamily.

Independent checks of the primary repairs. Independent NumPy/SciPy checks reproduced all 30 model/evaluation-stream endpoint combinations on 65,536 probes, the paired risk summaries, all 8,192-observation validation decisions, and empirical-risk feasibility from the original sufficient statistics. Raw-input fixtures checked endpoint, validation and risk calculations; a fresh 32,768-pair risk stream per field preserved the ordering. These are checks of saved fits. Endpoint optimality was not checked independently, so the result establishes an attained feasible gain. Unselected norm-frontier and secondary risk-budget checkpoints were not reevaluated.

Oracle capacity at the fitted temporal profiles. An oracle fits these coefficients to known teacher conditional means under the target latent law, using 65,536 fitting probes and 65,536 separate held-out probes. The primary relative singular-value cutoff is 10^{-8}, chosen after a pilot on field 301; nested 32,768-probe fitting and cutoffs $10^{-10}, 10^{-12}$ are sensitivity controls.

Table 7: Two diagnostics on the same five temporal fits and fixed Euler-16 target. KL is per 100D observation. Raw mean error sums squared errors over 92 conditional coordinates and averages over target latents.

Method	Latent KL	Scale KL	Mean KL	Total KL	Raw mean error
Saved CFM, Euler-16	0.01217	0.19474	6.30937	6.51627	0.121868
Oracle drive, Euler-16	0.01217	0.19474	0.01971	0.22662	0.000381
Saved CFM, refined	0.00565	4.98478	5.37572	10.36615	0.156647

The oracle removes 99.6877% of raw mean-square error on average. Its mean-KL reduction is 6.28965 nats, with 95% interval across five fits, conditional on the shared evaluation probes, [6.27910,6.30021] and separate paired integration MCSE 0.00660. It retains 1,081 of 1,168 scaled design directions; original-parameter Frobenius norms are 231,717–433,844, versus about 55 for saved CFM. Doubling fitting probes changes mean KL from 0.020239 to 0.019710; tighter cutoffs give about 0.019707 while norms rise to millions and hundreds of millions. This demonstrates attained capacity in an unrestricted conditional subfamily, not a population projection optimum or robustness to noisy labels, coefficient perturbations, low precision or untested tails.

Latent and scale terms remain unchanged, contributing an estimated 0.20691 nats even if mean error vanished; the exact scale term alone is 0.19474. With nonzero residual multipliers, changing contraction at fixed latent profiles rescales the drive blocks without changing their unrestricted mean-function span, although it changes noise scale and conditioning.

Refined integration of unchanged fields. A separate deployment diagnostic solves each bin's latent ODE exactly,

$$z(s) = \tfrac{1}{2}\,\mathrm{asinh}\big(e^{2a_j s}\sinh(2z_{\mathrm{left}})\big)\,,$$

uses exact exponential residual contraction, and integrates the drive by boundary-respecting quadrature. Orders 8 and 16 agree to numerical precision on the checked probes; order 16 is primary. Densities use the resulting triangular map's actual Jacobian and residual scale.

Against the unchanged Euler-16 target, refinement increases KL by 3.84987 nats, with conditional five-fit interval [3.82693,3.87281] and separate paired MCSE 0.00358. The original teacher itself changes from zero Euler-16 KL to 2.54877 nats. The smaller normalized mean-KL term accompanies a *larger* raw mean error: its denominator increases with the candidate variance. Refinement neither preserves the original exact-membership argument nor tests refitting at finer temporal resolution. The oracle and integration diagnostics are not an additive causal decomposition.

An independent 16,384-probe stream gives total KL 6.52605 for saved Euler-16 CFM, 0.22818 for the oracle, 10.37421 for refined CFM and 2.54826 for the refined teacher, with MCSE 0.01393, 0.00125, 0.01060 and 0.00054. The five-fit intervals condition on common oracle-fitting and evaluation probes; oracle-fitting randomness was not independently replicated. Integration MCSE averages fits within each shared probe before estimating uncertainty. No inference across targets follows. The 6.48746 and 6.51627 estimates use different probes for the same temporal fits.

Verification scope of the follow-ups. Independent NumPy/SciPy evaluation reproduced the temporal maps, densities, common-probe summaries and 15 empirical CFM quadratic solutions from sufficient statistics, without regenerating original training tuples. For the oracle and integration comparisons, it reproduced all 38 map/solver combinations on 65,536 probes, checked a 1,024-probe reference subset and evaluated a fresh stream. Separate RK4 integration and finite-difference Jacobians checked all six fields, including the teacher. These checks did not repeat oracle fitting or execute the author's Torch test suite.

Screen outcomes and deployment scope. The geometric CFM/likelihood score ratio is 3491.9, with an approximate five-seed 95% Student interval [2813.4,4334.0]; it is *not* a Wasserstein-distance ratio. The mean fresh-risk reduction from teacher to CFM is -0.33500, with interval [-0.33717,-0.33283]. Likelihood seed 203 is rejected, as are all conditional-only fits. Non-rejection of the other four likelihood fits does not establish equality. Their small KL and the conditional-only rejections are compatible with detection of residual discrepancies.

Interpolating 1% and 10% from teacher parameters toward the empirical CFM fit decreases fresh risk. Both change the contraction and therefore the conditional variance analytically, although the screen rejects only the 10% moves. These are parameter-direction diagnostics, not optimizer trajectories. Euler-8/32/64 or refined-ODE evaluation changes the teacher-relevant map and breaks closeness to the Euler-16 target. That limits cross-budget use without invalidating native reachability.

The original study and its temporal follow-ups use fixed known coordinates and feature structure. The temporal comparison expands CFM but retains the original likelihood family; it does not compare both objectives in the expanded family. None of these studies estimates an unknown rotation or isolates low intrinsic dimension or curvature as the causal explanation.

F.4 COUPLED 2D FINITE-DATA EXPERIMENT

Table 8: Full values for the short-correction comparison. Energy-U is the unbiased U-statistic estimator of energy distance, multiplied by 1,000. Cost is fitting plus one 2,048-point deployment. Means use 24 independent datasets per target; these are not simultaneous confidence statements.

	Energy-U $\times 10^3$		Seconds	
Method	Expansion	Contraction	Expansion	Contraction
Small CFM, 8 tuples/target	2.317	0.657	0.0172	0.0152
Larger CFM, 8 tuples/target	1.397	0.621	0.0377	0.0394
Endpoint-selected CFM	1.887	0.674	0.1452	0.1466
CFM + two damped updates	1.129	0.255	0.0855	0.0857
CFM + long endpoint solve	1.159	0.336	2.7543	2.2266

The source is $N(0, \mathrm{I}_2)$. Let $d_+ = 2^{-1/2}(1,1)$, $d_- = 2^{-1/2}(1,-1)$, and

$$\psi(x) = (1, \tanh x_1, \tanh x_2, \tanh(d_+^T x), \tanh(d_-^T x), \tanh(d_+^T x - 1), \tanh(d_-^T x + 1)).$$

The small field has features $\phi(t,x) = (\psi(x), (2t-1)\psi(x))$ and velocity $\phi(t,x)\Theta$, with $\Theta \in [-2,2]^{14\times 2}$. These bounded smooth features give globally Lipschitz fields for every head within the stated bounds. Target generation uses RK4-128 and the following evaluator-only heads; learners receive samples, not heads.

Table 9: Teacher coefficients, row order matching ϕ.

	Curved expansion		Rotating contraction	
Row	Output 1	Output 2	Output 1	Output 2
1	0.10	-0.15	0.15	-0.10
2	0.35	1.10	-0.80	0.80
3	-0.45	-0.20	-0.60	-0.60
4	0.50	-0.25	0.20	-0.45
5	-0.25	0.35	-0.35	0.20
6	0.65	0.25	0.45	0.40
7	-0.30	0.45	-0.25	0.10
8	-0.10	0.15	-0.10	0.15
9	0.15	-0.30	0.25	0.45
10	0.40	0.10	-0.35	0.25
11	-0.15	0.25	0.10	-0.20
12	0.10	-0.20	-0.20	0.30
13	0.20	0.10	0.20	-0.25
14	-0.10	-0.15	0.10	0.20

The large basis adds $\tanh(x_1 \pm 1.5)$, $\tanh(x_2 \pm 1.5)$, $\tanh(d_+^T x + 1)$, and $\tanh(d_-^T x - 1)$, with temporal basis $1, r, (3r^2-1)/2$, $r = 2t-1$. It has 78 parameters, with the same coefficient bounds. CFM uses bounded linear least squares, ridge 10^{-5}, and either eight or 32 independent source/time tuples per observed endpoint. Additional tuples reuse the target dataset; they are not additional target observations. Endpoint-selected CFM holds out one quarter, chooses among two bases and ridges $10^{-5}, 10^{-3}, 0.1, 1$ using empirical energy-V, and refits all observations. Its preparation, candidate fitting, validation and refitting costs are included.

Endpoint fitting uses 96 fixed Gaussian frequency vectors from NumPy seed 28041, with cyclic bandwidths 0.5, 1, 2 and scale $\max\{\sqrt{\mathrm{mean}_j \, \mathrm{Var}_{\mathrm{emp}}(Y_j)}, 0.25\}$. The 192 cosine/sine mean residuals are divided by $\sqrt{96}$. A scrambled Sobol source rule with 1,024 points is transformed through the Gaussian inverse CDF. The objective and its exact forward-sensitivity Jacobian use RK4-8.

Each correction solves $\min_\delta \|r + J\delta\|^2 + \mu \|\delta\|^2$ subject to $\theta + \delta \in [-2,2]^{28}$, with initial relative damping 0.01 and μ equal to relative damping times the mean squared Jacobian-column norm (floored at 10^{-12}). A proposal is accepted only if predicted and actual residual reductions are positive and their

ratio is at least 10^{-4}. The relative damping is divided by three if the ratio exceeds 0.75, multiplied by four below 0.25, and multiplied by ten on rejection, with bounds $[10^{-8}, 10^{8}]$. The fixed budget counts attempts, including rejected proposals. The long comparator is SciPy bounded trust-region reflective (TRF) with at most 200 residual calls and tolerances 10^{-7}. Its implementation can evaluate sensitivities at trials where a value-only calculation would suffice, so measured speed differences are between implementations, not a lower bound on every long solver.

The confirmation uses seeds 4000–4023 for each target, each with 2,048 training observations. Independent evaluation uses 2,048 source and target points, plus 512-point empirical optimal assignment for the secondary Wasserstein metric. The primary paired bootstrap uses 4,000 resamples. All 96 primary correction attempts are accepted; three first-attempt heads touch a bound, although no returned two-attempt head does. Negative unbiased discrepancy values are retained. The energy mean-gain and 80%-retention contrasts, in units of 10^{-3}, are:

	$\mathbb{E}[\mathcal{E}_{\mathrm{CFM}} - \mathcal{E}_{\mathrm{long}}]$ (95% interval)	$\mathbb{E}[\mathcal{E}_2 - 0.8\mathcal{E}_{\mathrm{long}} - 0.2\mathcal{E}_{\mathrm{CFM}}]$ (95% interval)
Expansion	$1.1581\ [0.1737, 2.0196]$	$-0.2620\ [-0.4374, -0.0744]$
Contraction	$0.3215\ [0.1068, 0.5210]$	$-0.1449\ [-0.2366, -0.0649]$.

These are dataset-level pointwise bootstrap statements for two fixed tasks.

F.5 Sample-only neural companion and metrics

Exploratory result. At 1,000 updates, adding the distinct-row moment penalty reduces the observed relative covariance score from 0.9742 to 0.2786 and improves energy and MMD. Its structural feature discrepancy increases from 0.001610 to 0.004208, versus 0.002116 for the strongest tested CFM structural comparator; both prespecified structural-improvement criteria fail. The Gaussian baseline has covariance score 0.2478 and structural score 0.005142. These are one-dataset finite-sample results, not a claim that every penalty weight fails. Moment-score improvement therefore did not extend to the structural criterion.

One fresh dataset (seed 9300) has 8,192 teacher outputs. The same-data controls include CFM128 at 2,000/18,000 updates, CFM256 at 6,000, direct refreshed-projection fitting for 1,000 updates, Gaussian MLE, and the evaluator-only teacher. Warm arms start from CFM128-2000; weights zero and one use matched source/direction streams, Adam 0.001, batch 256, and Heun-8, with 300/1,000-update checkpoints. CFM is also evaluated under Heun-32. All ten saved states and 13 model/sampler scores are retained.

Projected-quantile fitting gives equal total weight to 100 fixed coordinate directions and 64 independently redrawn Gaussian unit directions per update. Training-data projected standard deviations are floored at 0.25. Each term compares sorted generated projections with empirical midpoint quantiles of the fixed training observations. The generated batch is 256. The moment term is described in Appendix G.

Independent evaluation uses 2,048 generated and target points and 32,768 further teacher reference points. Structural features are $\cos(\omega x_i)$ and $\sin(\omega x_i)\sin(\omega x_{i+\ell})$, with $\omega = 1, 2, 3$, $\ell = 1, 7, 31$, and all valid coordinate indices. If $\bar{f}_x, s_x^2$ and $\bar{f}_r, s_r^2$ are sample feature means and unbiased variances, the score is the feature average of $(\bar{f}_x - \bar{f}_r)^2 - s_x^2/n - s_r^2/m$. This is unbiased for the squared discrepancy of the specified feature expectations, not for an unrestricted distance between laws.

Energy-U is $2\overline{\|x-y\|} - \overline{\|x-x'\|}_{i\neq j} - \overline{\|y-y'\|}_{i\neq j}$, divided by $\sqrt{100}$. MMD-U averages Gaussian kernels $\exp[-\|x-y\|^2/(2\cdot 100b^2)]$ for $b = 0.5, 1, 2$, excluding same-sample diagonals. Further measures include 128 fresh projected empirical quantile discrepancies, marginal quantiles, mean RMS, and third/fourth central moments. Relative covariance error is a Frobenius norm ratio against the independent reference covariance. Its nonzero true-versus-true value is sample noise, not a known population error floor.

The pilot-selected confirmation criteria require at least 25% covariance-score reduction versus the unregularized arm, at least 10% structural improvement over that arm and the best tested CFM structural score, and no energy/MMD deterioration beyond the absolute true-versus-true calibration relative to that CFM comparator. Both structural conditions fail; no fresh confirmation follows. The result concerns one dataset and one nonzero penalty weight. It cannot settle whether intermediate weights or a different properly estimated discrepancy would improve the tradeoff.

F.6 Reproduction and numerical scope

The coupled study's prior checks include six CFM and six two-update refits, independent discrete sensitivities, and 20 saved-model joint-metric evaluations. These are targeted reproductions, not reruns of all 48 datasets. The neural study was executed on one NVIDIA RTX PRO 6000 Blackwell device with Torch 2.14.0+cu132, deterministic algorithms and TF32 disabled, according to its execution receipts. Recorded case times sum to 48.75 seconds, excluding startup/runtime checks; no cross-hardware timing comparison is made.

Independent NumPy/SciPy code regenerated every diagnostic block for six neural fitted checkpoints, all 13 companion evaluations and all reference summaries. Maximum discrepancies were below 5×10^{-15} and 10^{-15}, respectively. This reproduces saved-state evaluation, not Adam training. The author's Torch test suite and Adam training were not rerun.

G Endpoint-surrogate controls

G.1 Unbiased empirical-target moment penalty

Training observations determine fixed mean $\widehat{\mu}$, coordinate scales $s_j = \max(\widehat{\sigma}_j, 0.25)$, and standardized raw second moment M. For generated standardized rows $Z_1, \dots, Z_B$, the implemented penalty is

$$\widehat{L} = \frac{1}{d}\left[\frac{\sum_{i\neq j}\{Z_i^T Z_j + (Z_i^T Z_j)^2\}}{B(B-1)} - \frac{2}{B}\sum_i Z_i^T M Z_i + \|M\|_F^2\right].$$

With iid generated rows, its conditional expectation is $d^{-1}(\|\mathbb{E}Z\|^2 + \left\|\mathbb{E}ZZ^T - M\right\|_F^2)$. Independence across distinct draw indices proves the identity even if two realized values coincide. Its Monte Carlo values may be negative. Target moments are still empirically estimated, and no cancellation removes the separate projected-quantile bias.

G.2 Finite-batch Gaussian midpoint bias

Let the target be exactly $N(0,1)$, with $q_i = \Phi^{-1}((i-1/2)/B)$. If generated outputs are sZ_i, with iid standard Gaussians and $s > 0$, the expected sorted-batch criterion is

$$L_B(s) = \mathbb{E}\frac{1}{B}\sum_i (sZ_{(i)} - q_i)^2 = s^2 - 2c_B s + A_B, \quad c_B = \frac{1}{B}\sum_i q_i \mathbb{E}Z_{(i)}, \quad A_B = \frac{1}{B}\sum_i q_i^2.$$

The minimizer is $s = c_B$. For $B \geq 2$, symmetry and the ordering of Gaussian order-statistic means give $c_B > 0$. Cauchy–Schwarz and Jensen give

$$c_B^2 \leq A_B\, B^{-1}\sum_i (\mathbb{E}Z_{(i)})^2 \leq A_B\, B^{-1}\sum_i \mathbb{E}[Z_{(i)}^2] = A_B.$$

The map $u \mapsto [\Phi^{-1}(u)]^2$ has strictly positive second derivative $2(1+z^2)/\varphi(z)^2$, $z = \Phi^{-1}(u)$. Strict midpoint convexity gives $A_B < \int_0^1 [\Phi^{-1}(u)]^2 du = 1$. Therefore $0 < c_B < 1$.

Integrating Gaussian order-statistic densities gives

$$c_{256} = 0.9917592848, \quad L_{256}(1) = 0.0114642690,$$
$$L_{256}(c_{256}) = 0.0113963596, \quad L'_{256}(1) = 0.0164814304.$$

The density used for the ith mean is $\Phi(z)^{i-1}(1-\Phi(z))^{B-i}\varphi(z)/\mathrm{B}(i, B+1-i)$. High-order Gaussian-variable quadrature and independent adaptive integration agree. These numerical values are not interval-certified enclosures. In an isotropic Gaussian scale family every unit projection has the same expected criterion, so direction refreshing cannot eliminate this bias. The phenomenon is related to biased sample Wasserstein gradients (Bellemare et al., 2017); it is not a claim that the small Gaussian contraction causes the entire high-dimensional covariance error.

G.3 Why finitely many projections need not identify covariance

For directions $a_1, \dots, a_m$, the linear map $\mathcal{A} : \mathrm{Sym}(d) \to \mathbb{R}^m$, $\mathcal{A}(\Delta)_j = a_j^T \Delta a_j$, has nontrivial kernel when $m < d(d+1)/2$. Choose $0 \neq \Delta \in \ker \mathcal{A}$ and $C \succ 0$. For sufficiently small $\delta > 0$, $C \pm \delta\Delta$ are distinct positive-definite covariances with identical Gaussian laws in every fitted projection. Any sum of those marginal batch losses therefore has equal expectation under both laws. Its variance or distribution need not agree, since projected losses can be dependent. This is a finite-view identification obstruction, not a general failure of randomized projections.

H Scope of the closest objective comparisons

Endpoint weights versus input weights. For complete regression input $Z = (T, X_T)$, including context when present, let $\widehat{w}$ be a fixed nonnegative endpoint-dependent weight. Where $m(Z) = \mathbb{E}[\widehat{w} \mid Z] > 0$, set $\widetilde{u}(Z) = \mathbb{E}[\widehat{w}U \mid Z]/m(Z)$. Conditional square completion gives

$$\mathbb{E}[\widehat{w}\|v - U\|^2] = \mathbb{E}[m(Z)\|v - \widetilde{u}(Z)\|^2] + \text{constant}$$

under the required finite moments. This metric-plus-target decomposition is Theorem 5.4 of Xie et al. (2026); their Theorem B.1 isolates approximation allocation in a two-candidate example. Our input-measurable positive metrics keep $u = \mathbb{E}[U \mid Z]$ fixed. The two intervention classes therefore need not have the same feasible selectors.

At zero tilt in the squeeze family, label dependence gives a concrete escape outside Theorem 1's information restriction. Let $a = F^T(v_* - U)$, $a_- = \max\{-a, 0\}$, and

$$\gamma = \frac{\mathbb{E}a}{\mathbb{E}a_-}, \qquad w = 1 + \gamma \mathbf{1}_{\{a<0\}}.$$

Here $\mathbb{E}a = G_0 > 0$. Conditional on Z, $a = k(\|e\|^2 - e^T\xi)$ with nondegenerate Gaussian ξ, so $0 < \mathbb{E}a_- < \infty$. The fixed population weight is bounded and strictly positive, and $\mathbb{E}[wa] = 0$ with $0 < \mathbb{E}[w\,\|F\|^2] < \infty$. Thus its weighted quadratic has the unique minimizer $\theta = 0$. The construction uses the fixed candidate, known features and observable labels; estimating γ from data would be a separate statistical procedure.

Autonomous models, stopping, and trajectory matching. The autonomous transport result of Lee et al. (2026, Section 2.1, Proposition 1) uses an unrestricted population drift under regular compact target-manifold assumptions, including codimension at least two, and trajectory-dependent terminal time. These conditions differ from the restricted autonomous Gaussian scale class and fixed horizon in Section 2; autonomy by itself is not our obstruction. Wang et al. (2026c, Section 4.1, Figure 6) examine one unsuccessful time-reweighting choice in time-unconditional learning, which does not assert failure for every positive weight. Wang et al. (2026a, Theorem 1 and Appendix A.2) construct students with equal trajectory-matching loss and different endpoint laws; that is distinct from excluding all correct population minimizers. Their marginal-alignment method uses log-density evolution, providing a direct connection to the Liouville identity rather than a competing identification theorem for our squeeze class.

Gaussian paths and the fixed horizon. For independent Gaussian endpoints and linear interpolation schedules, θx represents the reference velocity exactly only for the log-linear variance path σ^{2t} (Tyurin, 2026, Proposition 2.1, with reversed time); its drift is constant (Claustre et al., 2026, equation (6)). Endpoint reachability is the weaker property present here. The horizon matters: for $\sigma \neq 1$, any nonzero θ of the same sign as $\log \sigma$ reaches the scale at time $\log \sigma/\theta$, which need not equal one.

Path factorization. Equation (2) shows that fixing both the marginal path and reference velocity fixes the model-dependent CFM objective. Conditional-path refactorization can change label/gradient variance and a loss constant, but preserves population minimizers (Tyurin, 2026; Claustre et al., 2026).

Gradient variations at a population minimizer. Suppose a population minimizer v_* admits every additive variation $s\alpha(t)\nabla\phi(x)$ for small s, $\alpha \in C_c^\infty(0,1)$, $\phi \in C_c^\infty(\mathbb{R}^d)$. With integrable pairings,

valid differentiation under the expectation and a time weight $w > 0$ almost everywhere, stationarity gives

$$\int_0^1 w(t)\alpha(t)\mathbb{E}[(v_*(t, X_t) - u(t, X_t))^T \nabla\phi(X_t)]\, dt = 0.$$

Time localization gives $\nabla \cdot [p_t(v_* - u)] = 0$ weakly. The reference marginals therefore satisfy the continuity equation driven by v_*; uniqueness and a well-defined induced flow imply their preservation. These are admissible variations at the minimizer; global approximation capacity alone does not supply them. Finite parameterization alone does not imply selection failure either. Our finite shared-parameter construction lacks the required local variation freedom. State-dependent weights instead yield a weighted-flux equation, without this general implication.

This mechanism underlies quadratic-cost gradient projection for c-rectified flow (Wang et al., 2026b, Section 2.5.1 and Appendix B.5). Gradient restrictions also enter reflow-limit characterization under additional support conditions (Chambolle & Hertrich, 2026). These positive results use different class and iteration assumptions.