BEYOND TRAINING LOSS: FULL-SPACE CERTIFICATES FOR NEURAL JKO AND WASSERSTEIN EVOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A small neural training loss does not certify an inexact Jordan–Kinderlehrer–Otto (JKO) step. We develop a purely theoretical certification framework for entropy–potential–interaction energies in arbitrary finite dimension. A uniformly convex neural transport exposes an ambient residual, including the transported source score and the interaction force. We prove a supporting inequality against every finite-energy competitor, including nonsmooth Brenier maps; an audited residual norm consequently bounds both the unrestricted JKO gap and the one-step Wasserstein error. This is not a guarantee only within a network family. A second, explicitly differentiated transport-interpolation residual controls physical-time error without assuming that a general Wasserstein proximal map is contractive. We separate parameter-visible and missing directions, give a global integration procedure with Gaussian tail bounds, and prove an obstruction to certification from finitely many pointwise jets alone. Graph-consistent approximation yields conditional finite certification, while a separate regularity hypothesis is needed to exclude time-step accumulation. Exact noncommuting Gaussian and genuinely nonlinear manufactured models establish that the assumptions are nonempty. The results concern certifiability, not dimension-independent audit cost or convergence of arbitrary neural optimizers.

1 INTRODUCTION

Implicit distributional time stepping replaces a difficult evolution equation by a sequence of difficult optimization problems. In the JKO construction (Jordan et al., 1998), the next law minimizes

$$\mathcal{J}_{h,\mu}(\nu) = \mathcal{F}(\nu) + \frac{W_2^2(\nu,\mu)}{2h}. \tag{1}$$

Neural transports can propose approximate minimizers efficiently, but an optimizer's exit condition answers a parameter-space question. A small parameter gradient can coexist with a large component of the distributional residual in directions the network does not represent. A finite quadrature loss introduces a second uncertainty. Even an exact minimizer of (1) introduces a third error: discrete JKO time is not continuous physical time.

We ask a narrower and mathematically testable question: *which finite, independently checkable objects suffice to guarantee that a proposed neural evolution has the requested distributional accuracy?* Our answer uses a Brenier map on each time step, but does not identify the entire multidimensional Wasserstein space with a Hilbert space. The distinction is crucial. Convexity holds for the lifted single-step problem with a common source, whereas physical-time stability is proved at the level of probability measures.

The principal contribution is an integrated certificate for coupled entropy–interaction flows. Its central analytic ingredient is a full-space residual inequality that remains valid when the comparison transport has a singular distributional Hessian. Its computational ingredient is a convergent global audit using an explicitly represented source law, derivative enclosures, and tails. Its dynamical ingredient is a second residual along the actual transport interpolation. Together these yield a stopping rule and a physical-time error account that do not rely on neural optimality, empirical residuals, or an unproved Wasserstein resolvent contraction.

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.

The remaining results delimit what this mechanism can and cannot do. A tangent-space decomposition locates the missing directions; a certified feature-insertion direction gives a lower bound on the true gap. A finite-jet oracle obstruction explains why point sampling alone is insufficient. A graph-density theorem and fair enumeration provide a conditional termination result, but no statement that gradient descent finds the required network. All examples are analytic. There is no empirical accuracy or speed claim in this paper.

Related work. Variational neural Wasserstein flows (Fan et al., 2022), JKO normalizing flows (Xu et al., 2023), Deep JKO (Lee et al., 2024), importance-corrected JKO sampling (Hertrich & Gruhlke, 2025), and learned JKO operators (Feng et al., 2026) address neural realization and computational efficiency. Convex-network architectures are established tools (Amos et al., 2017). Primal–dual Wasserstein solvers already provide optimization certificates in their discretized formulations (Carrillo et al., 2022). Inexact JKO convergence under prescribed errors is studied by Di Marino et al. (2025); we address how an admissible error can be audited for a neural proposal. Kernel Gradient Discrepancy already supplies computable variational diagnostics and convergence control (Chazal et al., 2026); neither the existence of a computable discrepancy nor the elementary strong-convexity residual bound is our novelty claim. Residual-based Wasserstein trajectory estimates also predate this construction (Gonnella et al., 2026). Our contribution is the full-space, neural-JKO-specific certificate and its interaction with verified evaluation, representation, and physical time. The underlying transport, displacement-convexity, and gradient-flow theory is classical (Brenier, 1991; McCann, 1997; Ambrosio et al., 2008; Villani, 2009).

2 Setting: a common source, not a global flat geometry

On $\mathcal{P}_2(\mathbb{R}^d)$ consider

$$\mathcal{F}(\nu) = \mathrm{Ent}(\nu) + \int V \,\mathrm{d}\nu + \frac{1}{2}\iint W(x-y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y). \tag{2}$$

Here $\mathrm{Ent}(\rho\,\mathrm{d}x) = \int \rho\log\rho\,\mathrm{d}x$, and it is $+\infty$ for singular laws. We assume $V, W \in C^2$, $\nabla^2 V \succeq \lambda I$ with $\lambda \geq 0$, and W even and convex. Derivatives have at most polynomial growth. Finite-energy comparisons are understood throughout. These assumptions allow nonseparable potentials and interactions; they do not require a product law.

For certification the source has a positive C^1 density ρ_μ, finite moments of every needed order, finite absolute entropy, and score $s_\mu = \nabla \log \rho_\mu \in L^2(\mu)$. A candidate is $T = \nabla u$, where $u \in C^3$ and

$$mI \preceq K(x) := DT(x) \preceq MI, \qquad \sup_x \left\|D^2 T(x)\right\| < \infty, \qquad m > 0. \tag{3}$$

These are restrictions on the *candidate*, not on the unrestricted minimizer. They make T a global diffeomorphism and an optimal quadratic transport from μ to $\nu = T_{\#}\mu$. Its exact lifted objective is

$$\begin{aligned} P_h(T) = \mathrm{Ent}(\mu) &+ \int \left[V(T) - \log\det K + \frac{|T-x|^2}{2h}\right] \mathrm{d}\mu(x) \\ &+ \frac{1}{2}\iint W(T(x)-T(y))\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = \mathcal{J}_{h,\mu}(T_{\#}\mu). \end{aligned} \tag{4}$$

The unknown comparison map need not satisfy (3).

The source can itself be represented by $\mu = G_{\#}\gamma_d$, with a known orientation-preserving diffeomorphism G and $\gamma_d = N(0, I_d)$. Its score is evaluated without an inverse solve:

$$s_\mu(G(z)) = DG(z)^{-T}\{-z - \nabla_z \log\det DG(z)\}. \tag{5}$$

After an accepted step, G becomes $T \circ G$. This composition need not be a gradient map. Only the current relative transport T must be Brenier. Affine-plus-softplus convex potentials and curvature-audited smooth bump networks satisfy the candidate requirements; their precise definitions and derivative bounds appear in Appendix E.

3 A GLOBAL CERTIFICATE AGAINST UNRESTRICTED COMPETITORS

Define the interaction force and the ambient residual by

$$I_\mu[T](x) = \int \nabla W(T(x) - T(y))\,\mathrm{d}\mu(y),$$

$$r_h[T](x) = \nabla V(T(x)) + I_\mu[T](x) + \frac{T(x) - x}{h} + K(x)^{-1}\{s_\mu(x) - \nabla \log\det K(x)\}. \quad (6)$$

The last term is the score of $T_{\#}\mu$ evaluated at $T(x)$, not a score fitted on particles. Under the preceding assumptions $r_h[T] \in L^2(\mu;\mathbb{R}^d)$. Set $\alpha_h = \lambda + h^{-1}$.

Theorem 1 (Full-space residual certificate)**.** *The JKO problem has a unique minimizer $\nu^\star$, with Brenier map $S^\star$ from μ. For every finite-energy law η, with Brenier map S from μ,*

$$\mathcal{J}_{h,\mu}(\eta) \geq P_h(T) + \langle r_h[T], S - T\rangle_{L^2(\mu)} + \frac{\alpha_h}{2}\,\|S - T\|^2_{L^2(\mu)}\,. \quad (7)$$

Consequently, any independently valid bound $R \geq \|r_h[T]\|_{L^2(\mu)}$ gives

$$0 \leq P_h(T) - \mathcal{J}_{h,\mu}(\nu^\star) \leq \frac{R^2}{2\alpha_h}, \qquad W_2(T_{\#}\mu, \nu^\star) \leq \|T - S^\star\|_{L^2(\mu)} \leq \frac{R}{\alpha_h}. \quad (8)$$

Both constants are attained by translations of a Gaussian equilibrium.

Proof mechanism. Convexity of $-\log\det$ on positive definite matrices, λ-convexity of V, and convexity of W give a supporting inequality on source maps. The entropy term requires care: DS is a matrix-valued measure, and its singular part is positive semidefinite. Pairing that part with the positive matrix $\rho_\mu K^{-1}$ only improves the lower bound. Distributional integration by parts produces $\rho_\mu^{-1}\,\mathrm{div}(\rho_\mu K^{-1}) = K^{-1}(s_\mu - \nabla \log\det K)$. Cutoffs remove the boundary terms. This proves (7) without smoothness of S; full details are in Appendix C.

Strong convexity along generalized geodesics based at μ gives $P_h(T) - \mathcal{J}^\star \geq \alpha_h\,\|T - S^\star\|^2_{L^2(\mu)}/2$. Combining this lower bound with (7) at $S^\star$ gives the radius in (8). Completing the square gives its gap bound. No regularity or neural representation of $S^\star$ is used. □

The direct answer to "train until what?" is therefore $R \leq \alpha_h \eta$ for a requested single-step radius η. A poor network, optimizer, or auditor may fail to meet the condition, but cannot make a valid upper certificate falsely optimistic. One also obtains

$$\mathcal{F}(T_{\#}\mu) + \frac{\|T - \mathrm{Id}\|^2_{L^2(\mu)}}{2h} \leq \mathcal{F}(\mu) + g, \qquad g = R^2/(2\alpha_h). \quad (9)$$

An optional relative test $g \leq \theta\,\|T - \mathrm{Id}\|^2_{L^2(\mu)}/(2h)$ guarantees strict energy decrease for a nonzero step and $\theta < 1$. It is distinct from the physical-time accuracy test below.

3.1 USING INTERACTION CURVATURE WITHOUT OVERDAMPING THE MEAN

Suppose the stronger bounds $\nabla^2 V \succeq K_0 \succeq 0$ and $\nabla^2 W \succeq L_0 \succeq 0$ are known. Put $A_0 = K_0 + h^{-1}I$, $A_1 = A_0 + L_0$, and $\bar{r} = \mathbb{E}_\mu r_h[T]$. Then the scalar gap in (8) can be sharpened to

$$g_{\mathrm{mat}} = \frac{1}{2}\left\{\mathbb{E}_\mu[r_h^T A_1^{-1} r_h] + \bar{r}^T(A_0^{-1} - A_1^{-1})\bar{r}\right\}. \quad (10)$$

Every term is nonnegative, including for noncommuting K_0, L_0. The interaction supplies curvature only to mean-zero transport variations, because its second-order remainder is $\frac{1}{4}\mathbb{E}[(D(X) - D(Y))^T L_0 (D(X) - D(Y))]$. Completing the square in that mean/fluctuation metric proves (10); the full theorem and a subtraction-free audit are in Appendix T. For a centered residual and $L_0 = \ell I$, this replaces the denominator $\lambda + h^{-1}$ by $\lambda + \ell + h^{-1}$, while a pure translation receives no such improvement. The proximal curvature $h^{-1}I$ is not a physical flow contraction rate.

3.2 Representation error and a certified enrichment action

For $A = D_\theta T_\theta : \mathbb{R}^p \to L^2(\mu)$, write $b = A^* r_h$ and $H = A^* A$. Then

$$\|r_h\|^2 = b^T H^\dagger b + \|(I - \Pi_{\text{range } A}) r_h\|^2 . \tag{11}$$

The first term is invariant under nonsingular changes of coordinates; the second is invisible to parameter gradients. For the affine family $T(z) = sz$, source $N(0,1)$, and $V(y) = \lambda y^2/2 + \kappa y^4/4$, its restricted minimizer satisfies $r_h = \kappa s^3(z^3 - 3z)$ and $\|r_h\|^2 = 6\kappa^2 s^6 > 0$ when $\kappa > 0$. A compactly supported admissible descent direction proves a strictly positive unrestricted gap (Appendix F). A large residual alone is not a general lower gap bound.

This distinction supports an actionable test, rather than treating every unsuccessful upper certificate as evidence of insufficient network width. Let $v = \nabla\psi$ be compactly supported and let $\bar{t}\,\|Dv\|_{\infty,\text{op}} \le m/2$, so all maps $T + tv$, $0 \le t \le \bar{t}$, remain admissible. If audited constants satisfy

$$\langle r_h[T], v\rangle_{L^2(\mu)} \le -a < 0, \qquad \sup_{0\le t\le \bar{t}} \frac{\mathrm{d}^2}{\mathrm{d}t^2} P_h(T + tv) \le L, \tag{12}$$

then a feasible trial supplies a *lower* bound on the unrestricted gap.

Proposition 2 (A two-sided diagnosis)**.** *For $L > 0$, set $t_0 = \min\{\bar{t}, a/L\}$. Under* (12)*,*

$$0 < at_0 - Lt_0^2/2 \le P_h(T) - \mathcal{J}^\star \le R^2/(2\alpha_h). \tag{13}$$

The map $T + t_0 v$ is an admissible improved proposal. For $L = 0$, use $t_0 = \bar{t}$.

Taylor's theorem proves the lower bound; the global certificate proves the upper one. The second derivative is explicit, including $\mathrm{tr}[(DT + tDv)^{-1} Dv (DT + tDv)^{-1} Dv]$ for entropy and a nonnegative interaction Hessian integral. Appendix F gives the formula and its moment bounds. The signed bump architecture permits such insertions while retaining an independent curvature check. A positive lower improvement proves actual suboptimality; a large upper residual alone does not.

3.3 An alternative audit for a non-Brenier proposal

When $W = 0$, the residual theorem has an independent primal–dual companion. For any diffeomorphism H, the coupling-based value

$$P_h^{\text{coup}}(H) = \mathrm{Ent}(\mu) + \mathbb{E}_\mu \left[V(H) - \log\det DH + \frac{|H - x|^2}{2h} \right] \tag{14}$$

is an upper bound on the candidate's true JKO objective. For a witness f define

$$f^c(x) = \inf_y \left\{ \frac{|x - y|^2}{2h} - f(y) \right\}, \qquad D_h(f) = \mathbb{E}_\mu f^c - \log \int e^{-V - f}. \tag{15}$$

Transport weak duality and the Gibbs variational identity give $D_h(f) \le \mathcal{J}^\star$. Thus a valid primal upper value minus a valid dual lower value bounds the unrestricted gap even when H is not an optimal transport map. A witness of the form $f(y) = |y|^2/(2h) - \psi(y)$ reduces its c-transform to the conjugate of a strongly convex neural potential. Appendix O gives the residual correction for an inexact conjugate solve and a log-partition lower bound with the necessary approximate-mode correction. This is a weak-duality certificate, not a stationarity certificate for an arbitrary generator. Its particular witness class is not claimed to attain zero gap for every problem. For a Brenier output, the smaller of the two independently valid gap bounds may be used.

4 An audit model and an information obstruction

Training integrals and certification integrals have different contracts. Training may use arbitrary samples. Certification freezes the exported map, checks (3), and encloses the norm of (6) over the entire source. For $\mu = G_\# \gamma_d$, set $f(z) = r_h[T](G(z))$. On a box B centered at z_B, suppose $|f(z_B)| \le a_B$, $\mathrm{Lip}(f|_B) \le L_B$, and its radius is δ_B. Then

$$\int_B |f|^2 \,\mathrm{d}\gamma_d \le \gamma_d(B)(a_B + L_B \delta_B)^2. \tag{16}$$

Interval enclosures of f can replace this Lipschitz construction. A polynomial envelope outside a growing cube yields a tail bound from one-dimensional Gaussian moment recurrences. Polynomial interactions reduce to finitely many audited moments; more general interactions require a separately convergent inner integral audit.

Theorem 3 (Soundness and strict-slack completeness of evaluation)**.** *Assume sound convergent interval extensions on compact boxes, a computable integrable polynomial envelope, and an exactly specified frozen model. There is a finite-data procedure returning $R_n \geq \|r_h[T]\|_{L^2(\mu)}$ with $R_n \to \|r_h[T]\|_{L^2(\mu)}$. Thus every threshold strictly larger than the true norm is eventually certified. The same conclusion holds for the interpolation derivative norm used below, uniformly in its interpolation parameter. No finite-stage decision is guaranteed at equality.*

The proof constructs upper and lower cell sums, truncates tails, and sends precision, truncation radius, and mesh resolution to their limits in a specified order (Appendix G). It is a computability result. In dimension d, a generic box audit has exponential dependence on d; neither the norm estimate nor the algorithm removes this cost.

Theorem 4 (Finite-jet obstruction)**.** *A deterministic oracle algorithm that, on every smooth increasing affine-tail map on $\mathbb{R}$, terminates after finitely many adaptive evaluations of finitely many derivatives cannot always return a finite valid upper bound on the unrestricted JKO gap using only those evaluations. This already holds for $\mu = N(0,1)$ and $V(y) = y^2/2$.*

On the identity transcript, put a smooth positive step after the largest queried point. All queried jets are unchanged, but its amplitude makes the JKO gap arbitrarily large. The full proof and a one-neuron fixed-quadrature counterpart are in Appendix H. The theorem does not apply to an auditor that reads network coefficients and certified global derivative bounds.

A sharper analytic example applies even to a single smooth monotone neuron. Put $g(u) = u/\sqrt{1+u^2}$ and $T_w(z) = z + ag((z-c)/w)$, where c avoids every node of a fixed finite quadrature rule with nonnegative weights summing to one. For source $N(0,1)$, $V(y) = y^2/2$, and $\alpha = 1 + h^{-1}$, the rule's residual-derived gap estimate converges, as $w \downarrow 0$, to $\alpha a^2/2$, while the true gap converges to

$$\frac{\alpha a^2}{2} + 2a\phi(c). \tag{17}$$

Their ratio becomes arbitrarily large as $a \downarrow 0$. The proof controls the log-Jacobian integral and uses exact Gaussian integration by parts, not a second approximate optimizer. This is an obstruction to unvalidated gap estimation; it does not state that every Wasserstein-distance diagnostic fails on the same example. Optional fresh-sample bounds with a summable allocation of failure probabilities are given in Appendix I; reusing an adaptively searched holdout is not justified.

5 A SECOND CERTIFICATE FOR PHYSICAL TIME

On a step of duration h, let $\Delta = T - \mathrm{Id}$, $Q_\tau = \mathrm{Id} + \tau\Delta$, and $B_\tau = I + \tau D\Delta$, $0 \leq \tau \leq 1$. These maps remain invertible. Let $\rho_\tau = (Q_\tau)_\# \mu$. The negative Wasserstein gradient of $\mathcal{F}$ is $-a_\rho$, where $a_\rho = \nabla \log \rho + \nabla V + \nabla W * \rho$. The Eulerian velocity of the interpolation is $v_\tau(Q_\tau x) = \Delta(x)/h$. Its defect, in source coordinates, is

$$e_\tau(Q_\tau x) = \frac{\Delta(x)}{h} + A_\tau(x), \qquad A_\tau = a_{\rho_\tau} \circ Q_\tau. \tag{18}$$

The endpoint equals $r_h[T]$. The derivative throughout the step is explicit:

$$\begin{aligned} \partial_\tau A_\tau &= \nabla^2 V(Q_\tau)\Delta + \int \nabla^2 W(Q_\tau(x) - Q_\tau(y))(\Delta(x) - \Delta(y))\,\mathrm{d}\mu(y) \\ &\quad - B_\tau^{-1} D\Delta B_\tau^{-1}(s_\mu - c_\tau) - B_\tau^{-1}\dot{c}_\tau, \\ c_\tau &= \nabla \log\det B_\tau, \qquad (\dot{c}_\tau)_i = \mathrm{tr}(B_\tau^{-2}\partial_i D\Delta). \end{aligned} \tag{19}$$

An audit $C \geq \sup_{\tau\in[0,1]} \|\partial_\tau A_\tau\|_{L^2(\mu)}$ gives $\|e_\tau\|_{L^2(\rho_\tau)} \leq R + (1-\tau)C$.

Theorem 5 (Physical-time error account)**.** *Let $\bar{\rho}(t)$ be the EVI$_\lambda$ gradient flow of* (2)*, and let a finite sequence of smooth transport interpolations satisfy the stated integrability conditions. Suppose*

Algorithm 1 CERTJKO-B: propose, audit, and advance only on certified acceptance

Require: Known G_0, target horizon, error budget, and a proposal mechanism
$G \leftarrow G_0$, $\mu \leftarrow G_{\#}\gamma_d$, $t \leftarrow 0$, $E \leftarrow E_0$
while $t < T_{\mathrm{fin}}$ **do**
  Propose $h \le T_{\mathrm{fin}} - t$ and a neural relative map $T = \nabla u$
  Freeze T; verify curvature and integrability; audit R from (6)
  if the optimization test in (21) fails **then**
    Refine the audit or improve/enrich the proposal; do not advance
  else
    Audit C from (19)
    if the physical-time test in (21) holds **then**
      $E \leftarrow q_h E + a_h R + b_h C$; $\mu \leftarrow T_{\#}\mu$; $G \leftarrow T \circ G$; $t \leftarrow t + h$
    else
      Revise the step proposal; do not advance
    end if
  end if
  A resource-limited exit returns "not certified", never an accepted step
end while
return G, E, and the finite certificate records

$W_2(\mu_0, \bar{\rho}(0)) \le E_0$. *For audited* R_k, C_k *and* $t_{k+1} = t_k + h_k$, *define*

$$E_{k+1} = q_k E_k + a_k R_k + b_k C_k, \qquad q_k = e^{-\lambda h_k},$$
$$a_k = \frac{1 - e^{-\lambda h_k}}{\lambda}, \qquad b_k = \frac{1 - (1 + \lambda h_k) e^{-\lambda h_k}}{h_k \lambda^2}. \tag{20}$$

Then $W_2(\mu_k, \bar{\rho}(t_k)) \le E_k$. *At* $\lambda = 0$ *the continuous extensions are* $a_k = h_k$ *and* $b_k = h_k/2$. *An analogous bound holds inside every step.*

Proof mechanism. The EVI and the displacement subgradient inequality yield $\frac{\mathrm{d}}{\mathrm{d}t} W_2(\rho, \bar{\rho}) \le -\lambda W_2(\rho, \bar{\rho}) + \|e\|_{L^2(\rho)}$. Integrating the endpoint and clock terms of (18) against the exponential kernel gives (20). The measure-level argument in Appendix K does not assume contractivity of the JKO resolvent, and does not require the global composite generator to be a gradient map. □

An arbitrarily accurate inner minimization only reduces R_k. It cannot remove the C_k term for a nonzero time step. Conversely, with a fixed approximate source score satisfying a separately certified L^2 error σ, the endpoint bound must be enlarged by σ/m; the clock bound must also account for that uncertainty. These and other perturbations are detailed in Appendix R.

6 A CERTIFICATE-DRIVEN ALGORITHM AND ITS GUARANTEES

For a horizon T_{fin}, a zero initial error, and target ε, put $\zeta = \varepsilon / T_{\mathrm{fin}}$ and choose $\theta \in (0, 1)$. A step is accepted only when

$$a_h R \le \theta \zeta h, \qquad b_h C \le (1 - \theta) \zeta h. \tag{21}$$

The first failure calls for more optimization, model enrichment, or a tighter audit; the second calls for a better step proposal, typically a shorter step. Neither authorizes advancing time. Induction in (20) gives $E_k \le \zeta t_k$. Nonzero initial error is propagated by its exponential factor rather than silently dropped. Weighted tests certify the grid endpoints; the explicit within-step envelope remains valid everywhere. To enforce the same ε-tube throughout the interpolation, the stronger tests $R \le \theta\zeta$ and $C/2 \le (1 - \theta)\zeta$ suffice (Appendix L).

Theorem 6 (What termination and accuracy do require)**.** *For a fixed source and step, suppose the exact Brenier minimizer lies in the* C^2*-graph closure of curvature-admissible, auditable neural maps, with residuals converging in* $L^2(\mu)$ *to zero. Fairly interleaving those proposals with convergent audits eventually certifies every positive one-step tolerance. Smooth quadratic-tail uniformly convex transport potentials admit such graph approximation by the bump architecture in Appendix M. To guarantee finitely many accepted steps up to a prescribed horizon, a separate uniform small-step*

bound $C_h^\star \le Kh$ and uniform graph-approximation control suffice. Without those hypotheses, no no-accumulation or arbitrary-optimizer convergence claim is made.

The proof is constructive but does not claim efficient enumeration. Under $R_k \le c_R h_k$ and $C_k \le c_C h_k$, Theorem 5 gives

$$W_2(\mu_N, \bar{\rho}(T_{\mathrm{fin}})) \le E_0 + (c_R + c_C/2) T_{\mathrm{fin}} \max_k h_k. \tag{22}$$

The corresponding JKO objective budget is $g_k \le c_R^2 h_k^3/[2(1+\lambda h_k)]$. Thus the familiar cubic inner gap scaling is derived from a checkable full residual and a separate clock condition, rather than assumed to control all errors by itself.

Analytic nonempty model classes. For $V(x) = x^T K x/2$, $W(x) = x^T L x/2$, $K \succ 0$, $L \succeq 0$, and $\mu = N(m_0, \Sigma_0)$, the exact minimizer is Gaussian even when K, L, Σ_0 do not commute. Its relative map is $T(x) = m + A(x - m_0)$ with

$$m = (I + hK)^{-1} m_0, \quad A = B^{-1/2} \frac{h^{-1} I + (h^{-2} I + 4B^{1/2} \Sigma_0^{-1} B^{1/2})^{1/2}}{2} B^{-1/2}, \tag{23}$$

where $B = K + L + h^{-1} I$. Its full residual is exactly zero, proving unrestricted optimality by Theorem 1. Appendix P supplies the derivation. Appendix Q constructs nonaffine, nonseparable, uniformly convex examples with exact neural minimizers, rather than extrapolating from Gaussian closure.

7 Scope and conclusion

The result is a theoretical interface between neural proposals and certified evolution for smooth density-based convex interaction flows. It extends beyond scalar quantiles and known products, but not to arbitrary energies, empirical atomic sources, or unconstrained generators. The audits can be prohibitively expensive. Graph approximation, finite audit termination under strict slack, and optimization success are different statements. We prove the first two under explicit conditions and do not assume the third. An optional primal–dual certificate covers non-Brenier diffeomorphisms when $W = 0$ (Appendix O); it does not silently extend the residual theorem to those maps. Within its stated scope, the framework answers how far training must proceed by requiring auditable full-space and physical-time budgets, not by assigning a distributional meaning to parameter convergence.

AI USE STATEMENT

The authors declare that no generative AI tools or LLMs were used to generate text, code, or ideas for this paper. Standard spell-checking and basic grammar-checking tools were used solely for proofreading.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, second edition, 2008. doi: 10.1007/978-3-7643-8722-8.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017. URL https://proceedings.mlr.press/v70/amos17b.html.

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

José A. Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, 22:389–443, 2022. doi: 10.1007/s10208-021-09503-1.

Clémentine Chazal, Heishiro Kanagawa, Zheyang Shen, Anna Korba, and Chris J. Oates. A computable measure of suboptimality for entropy-regularised variational objectives. arXiv:2509.10393v4, 2026. URL https://arxiv.org/abs/2509.10393v4. First posted in 2025; version 4 dated July 9, 2026.

Simone Di Marino, Emanuele Naldi, and Silvia Villa. Inexact JKO and proximal-gradient algorithms in the Wasserstein space. arXiv:2505.23517, 2025. URL https://arxiv.org/abs/2505.23517.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6185–6215. PMLR, 2022. URL https://proceedings.mlr.press/v162/fan22d.html.

Xue Feng, Li Wang, Deanna Needell, and Rongjie Lai. Learn to evolve: Self-supervised neural JKO operator for Wasserstein gradient flow. arXiv:2601.05583, 2026. URL https://arxiv.org/abs/2601.05583.

Isabella Carla Gonnella, Olga Mula, Federico Pichi, and Gianluigi Rozza. A dynamical approximation scheme on the Stiefel manifold for Wasserstein gradient flows. arXiv:2607.17235v1, 2026. URL https://arxiv.org/abs/2607.17235v1.

Johannes Hertrich and Robert Gruhlke. Importance corrected neural JKO sampling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23083–23119. PMLR, 2025. URL https://proceedings.mlr.press/v267/hertrich25a.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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

Wonjun Lee, Li Wang, and Wuchen Li. Deep JKO: Time-implicit particle methods for general nonlinear gradient flows. *Journal of Computational Physics*, 514:113187, 2024. doi: 10.1016/j.jcp.2024.113187.

Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1): 153–179, 1997. doi: 10.1006/aima.1997.1634.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015. doi: 10.1007/978-3-319-20828-2.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47379–47405, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/93fce71def4e3cf418918805455d436f-Abstract-Conference.html.

SUPPLEMENTARY THEORY: READING MAP

The supplement gives proofs rather than additional experiments. The single-step theorem, numerical audit theorem, and physical-time theorem have different assumptions; the table below indicates where those assumptions are used. All dimensions are finite. A statement that is conditional on representability, computable envelopes, or an EVI solution is labeled as such. No claim of a general polynomial-time algorithm is implicit in the word "computable".

Section	Content	Page
A	Domains, conventions, and assumption hierarchy	10
B	Existence, Brenier coordinates, and generalized convexity	12
C	Full-space support with nonsmooth comparison maps	14
D	Sharp gap/radius constants and energy inequalities	16
E	Neural maps, density scores, and composition bounds	18
F	Invisible directions and certified enrichment	20
G	Global interval integration and Gaussian tails	22
H	Adaptive finite-jet and fixed-quadrature obstructions	25
I	Fresh-sample and repeated-audit probability bounds	27
J	Matrix-valued interpolation derivative calculations	28
K	Measure-level stability and exact clock coefficients	30
L	Finite-horizon, long-time, and vanishing-step guarantees	32
M	Graph approximation by curvature-audited neural maps	34
N	Fair search, strict slack, and nonaccumulation	37
O	An independent primal–dual certificate when $W = 0$	38
P	Exact noncommuting Gaussian interaction model	40
Q	Exact nonaffine coupled examples	43
R	Score, force, initial-law, and export uncertainty	45
S	Optimal tolerance allocation under explicit work models	47
T	Matrix curvature and mean/fluctuation certificates	48
U	Proof dependencies and the finite certificate contract	50

The external foundational inputs are Brenier's transport theorem and its almost-everywhere Jacobian identity, the Wasserstein velocity/distance calculus, and the existence of an EVI gradient flow in the standard displacement-convex setting. These are used in their usual finite-second-moment domains (Brenier, 1991; Ambrosio et al., 2008; Villani, 2009; Santambrogio, 2015). We prove the problem-specific inequalities, all approximation statements asserted here, the numerical information obstructions, and the certificate bookkeeping.

A DOMAINS, ASSUMPTIONS, AND NOTATION

A.1 EXTENDED ENERGIES AND INTEGRABILITY

For a Borel probability measure ν on $\mathbb{R}^d$, write $m_p(\nu) = \int |x|^p \, \mathrm{d}\nu(x)$ whenever this is finite. The space $\mathcal{P}_2$ consists of laws with $m_2 < \infty$. We use the standard quadratic transport cost

$$W_2^2(\nu, \eta) = \inf_{\pi \in \Pi(\nu,\eta)} \int |x - y|^2 \, \mathrm{d}\pi(x, y). \tag{A.1}$$

All vector norms are Euclidean. Matrix operator norms are induced by that norm. $\left\| D^2 T \right\|_{\mathrm{F}}$ means the Euclidean norm of all its second derivatives, or equivalently $\left(\sum_i \|\partial_i DT\|_{\mathrm{F}}^2 \right)^{1/2}$. The notation $A \preceq B$ applies only to symmetric matrices.

The entropy convention is important on the unbounded domain. For $\nu = \rho \, \mathrm{d}x \in \mathcal{P}_2$, the negative part of $\rho \log \rho$ is integrable. One way to see this is to compare with the normalized Gaussian density $g_a(x) = (a/\pi)^{d/2} e^{-a|x|^2}$, $a > 0$. The identity

$$\mathrm{Ent}(\nu) = \mathrm{KL}(\nu \| g_a \, \mathrm{d}x) - a m_2(\nu) + \frac{d}{2} \log(a/\pi) \tag{A.2}$$

holds as an extended-real identity. In particular

$$\mathrm{Ent}(\nu) \geq -a m_2(\nu) + \frac{d}{2} \log(a/\pi). \tag{A.3}$$

The extended energy (2) is therefore defined without an $\infty - \infty$ ambiguity under the lower bounds below. Its effective domain contains only absolutely continuous laws.

Assumption A.1 (Energy class)**.** $V, W \in C^2(\mathbb{R}^d)$, $\nabla^2 V \succeq \lambda I$ for a fixed $\lambda \geq 0$, $W(-x) = W(x)$, and $\nabla^2 W \succeq 0$. The functions and derivatives needed in an assertion are of at most polynomial growth. For assertions involving derivative enclosures, the corresponding functions are also supplied with computable interval extensions and polynomial majorants.

Convexity gives $V(x) \geq V(0) + \nabla V(0) \cdot x + \lambda |x|^2/2$. Even convexity gives $W(x) \geq W(0)$. Adding constants to V or W changes the energy by a constant, and never changes a residual, gap, minimizer, or flow. No strict convexity of W is assumed. In particular the interaction force can vanish on translations.

Assumption A.2 (Source for a single-step certificate)**.** The source is $\mu = \rho_\mu \, \mathrm{d}x$ with $\rho_\mu > 0$ and $\rho_\mu \in C^1$. It has finite absolute entropy, a square-integrable score s_μ, and all polynomial moments required by the candidate and energy. In particular $\mu \in \mathcal{P}_2$. All source integrals below refer to this exact law, not to its empirical measure.

The score requirement is used for integrability of the full residual and the boundary calculation, not for an unproved statistical estimate of that score. A known smooth generator supplies the score explicitly by (5). If the source is known only through data, a separate initial-law or score certificate is needed; Appendix R states exactly how such a certificate enters.

Assumption A.3 (Admissible candidate)**.** $T = \nabla u$, where $u \in C^3(\mathbb{R}^d)$, $mI \preceq D^2 u \preceq MI$ for finite constants $0 < m \leq M$, and $D^3 u$ is bounded. The coefficients of an audited neural candidate specify one exact map. Higher derivatives are assumed only when an audit based on them is invoked.

The C^3 regularity of the potential is the C^2 regularity of the map. It is sufficient for the explicit score and interpolation formulas. A Lipschitz audit of the residual may use one further derivative; a direct interval enclosure need not differentiate the residual again. These are different evaluation strategies.

A.2 Three levels of guarantee

A *mathematical certificate* is an inequality between exact quantities, for example (8). An *evaluation certificate* is finite data proving that a returned number is an upper bound on a mathematical quantity, conditional on sound arithmetic and supplied enclosures. A *termination theorem* proves that an algorithm eventually produces such data. Validity of the first two does not imply the third.

Similarly, three errors have different meanings. The *neural optimization error* compares an output with the best output in a specified parameter family. The *unrestricted JKO error* compares with the minimizer over all laws. The *physical-time error* compares with the continuous flow at the prescribed time. We do not assume that bounds on these errors coincide.

For a source μ, let $\mathcal{B}_\mu$ be the collection of $L^2(\mu)$ equivalence classes of gradients of convex functions. Brenier's theorem identifies each target in $\mathcal{P}_2$ with one such optimal map, up to μ-null sets. This does not identify the distance between two targets with the $L^2(\mu)$ distance between their maps in dimensions greater than one. Only the inequality

$$W_2(S_{\#}\mu, T_{\#}\mu) \leq \|S - T\|_{L^2(\mu)} \tag{A.4}$$

is automatic, because $(S, T)_{\#}\mu$ is an admissible coupling.

A.3 What remains fixed during an audit

The source representation, current step size, energy, candidate coefficients, and arithmetic interpretation are fixed. Changing any of them creates a new proposed step and requires a new certificate. In a stochastic audit, the model is fixed conditionally on all past information before new audit samples are drawn. In a deterministic audit, adaptively refining the spatial partition is harmless because every cell bound remains a pointwise enclosure of the same fixed function.

The main results use exact arithmetic in their mathematical statements. An implementation may instead return rational intervals that enclose all operations. A floating-point value accompanied only by a quadrature library's heuristic error estimate does not meet the contract. The paper does not assert the correctness of an arbitrary interval software package; that is a separate trusted component in any implementation (Rump, 2010).

B TRANSPORT PREPARATION AND THE UNRESTRICTED JKO PROBLEM

B.1 GLOBAL INVERTIBILITY OF A CANDIDATE

Lemma B.1 (Uniform convexity supplies an optimal diffeomorphism)**.** *Under Assumption A.3, T is a global C^2 diffeomorphism, T^{-1} is m^{-1}-Lipschitz, and T is M-Lipschitz. For every absolutely continuous source in $\mathcal{P}_2$, T is its optimal quadratic transport to $T_{\#}\mu$.*

Proof. For $x, y \in \mathbb{R}^d$, the fundamental theorem of calculus gives

$$(T(x) - T(y)) \cdot (x - y) = \int_0^1 (x - y)^T K(y + t(x - y))(x - y)\,\mathrm{d}t \geq m|x - y|^2. \tag{B.1}$$

Cauchy–Schwarz yields $|T(x) - T(y)| \geq m|x - y|$. Thus T is injective and its inverse on its image is m^{-1}-Lipschitz. The upper bound follows from $K \preceq MI$. For any z, the function $u(x) - z \cdot x$ is coercive because it is bounded below by an affine function plus $m|x|^2/2$. It has a minimizer, at which $T(x) = z$. Hence the image is all of $\mathbb{R}^d$. The inverse function theorem applies because K is everywhere invertible and C^1. Finally T is the gradient of a convex function, so Brenier's characterization proves optimality. The linear growth bound $|T(x)| \leq |T(0)| + M|x|$ ensures a finite second moment of the target. □

The last assertion is more than an architectural convenience. It is what makes the transport penalty in (4) equal the true Wasserstein penalty. For a general invertible generator, the same integral is only an upper bound; Appendix O treats that distinct case.

B.2 THE ENTROPY CHANGE OF VARIABLES, INCLUDING ROUGH TRANSPORTS

For an admissible candidate the ordinary change-of-variables theorem gives

$$\rho_{T_{\#}\mu}(T(x)) = \frac{\rho_\mu(x)}{\det DT(x)}, \qquad \mathrm{Ent}(T_{\#}\mu) = \mathrm{Ent}(\mu) - \int \log\det DT\,\mathrm{d}\mu. \tag{B.2}$$

The logarithm is integrable because $d\log m \leq \log\det DT \leq d\log M$. For comparison maps we need the corresponding statement without global smoothness.

Lemma B.2 (Jacobian identity for a finite-energy comparison)**.** *Let $S = \nabla v$ be the Brenier map from the positive-density source μ to an absolutely continuous target $\eta \in \mathcal{P}_2$ with finite entropy. The distributional derivative $DS = D^2 v$ is a locally finite positive semidefinite matrix-valued measure. Write $D_a S$ for its Lebesgue density. Then $D_a S$ is positive definite μ-almost everywhere and*

$$\mathrm{Ent}(\eta) - \mathrm{Ent}(\mu) = -\int \log\det D_a S(x)\,\mathrm{d}\mu(x). \tag{B.3}$$

Proof. A convex Brenier potential can be chosen finite on all of $\mathbb{R}^d$: its convex effective domain has full measure for a source with an everywhere positive density, so its interior is $\mathbb{R}^d$. It is locally Lipschitz. Its gradient is locally of bounded variation and its distributional Hessian is positive semidefinite. Alexandrov differentiability identifies the density of this Hessian with $D_a S$ almost everywhere.

The almost-everywhere Monge–Ampère change-of-variables identity for optimal maps between absolutely continuous laws is

$$\rho_\mu(x) = \rho_\eta(S(x))\det D_a S(x) \quad \text{for } \mu\text{-almost every } x. \tag{B.4}$$

This is the standard Jacobian form of Brenier transport; see Villani (2009) or Santambrogio (2015). The target density is positive and finite at η-almost every point carrying mass. Therefore the determinant is positive μ-almost everywhere. Taking logarithms, integrating, and using $S_{\#}\mu = \eta$ proves (B.3). Both entropies are well defined and finite, so the identity also supplies integrability of the logarithmic determinant. Singular parts of DS are not included in this pointwise determinant. They reappear, with a favorable sign, in the distributional integration by parts below. □

This use of the almost-everywhere Jacobian identity does not assume that the comparison map is globally injective as a classical smooth function, or that its distributional derivative has no singular part. Those would be unnecessary restrictions on the unrestricted optimum.

B.3 Existence by a direct argument

Proposition B.3 (Existence of an unrestricted minimizer)**.** *Under Assumptions A.1–A.2, for every $h > 0$ the functional $\mathcal{J}_{h,\mu}$ attains its minimum at a finite-energy law.*

Proof. The source itself is a finite-energy competitor under the moment assumptions. Consider a minimizing sequence ν_n and optimal couplings $\pi_n \in \Pi(\mu, \nu_n)$. Fix $a = 1/(8h)$ and use (A.2). The elementary inequality

$$|y - x|^2 \geq \tfrac{1}{2}|y|^2 - |x|^2 \tag{B.5}$$

gives

$$\mathcal{J}_{h,\mu}(\nu_n) \geq \mathrm{KL}(\nu_n \| g_a \, \mathrm{d}y) + \frac{1}{8h} m_2(\nu_n) - \frac{1}{2h} m_2(\mu) - c_0 - c_1 m_2(\nu_n)^{1/2}. \tag{B.6}$$

Here the affine lower bound on V and the constant lower bound on W have been used. The quadratic term dominates the linear square-root term, so the target second moments are uniformly bounded. The couplings are tight, and a subsequence converges narrowly to a coupling π with first marginal μ and second marginal $\nu \in \mathcal{P}_2$.

To verify lower semicontinuity without separately subtracting an unbounded moment, keep the Gaussian relative entropy and transport term together. The function

$$H(x, y) = \frac{|y - x|^2}{2h} - a|y|^2 + V(y) \tag{B.7}$$

is continuous and bounded below by $-C(1 + |x|^2)$; completing the square in y proves this bound. Add $C(1 + |x|^2)$ to make it nonnegative. Portmanteau applies to the modified integrand, and its subtracted integral is fixed because the first marginal remains μ. The relative entropy is narrowly lower semicontinuous. The interaction integral is lower semicontinuous because W is continuous and bounded below, and $\nu_n \otimes \nu_n$ converges narrowly to $\nu \otimes \nu$.

It follows that the limit coupling energy is at most the limiting infimum. Replacing its transport integral by $W_2^2(\mu, \nu)$ can only decrease that value. Thus ν is a minimizer. Its energy is finite because the minimizing value is finite, entropy has an integrable negative part, and the other energy terms have the stated lower bounds. □

This argument requires narrow compactness, not compactness of a bounded set in the Wasserstein topology. A uniform second-moment bound alone does not imply the latter, which is why the lower-semicontinuity calculation was performed on couplings.

B.4 Convexity in a common-source chart

Proposition B.4 (Generalized-geodesic strong convexity)**.** *Let S_0, S_1 be the Brenier maps from μ to finite-energy laws ν_0, ν_1. Put $S_t = (1 - t)S_0 + tS_1$ and $\nu_t = (S_t)_{\#}\mu$. Then*

$$\mathcal{J}_{h,\mu}(\nu_t) \leq (1 - t)\mathcal{J}_{h,\mu}(\nu_0) + t\mathcal{J}_{h,\mu}(\nu_1) - \frac{\alpha_h}{2} t(1 - t) \, \|S_1 - S_0\|_{L^2(\mu)}^2 . \tag{B.8}$$

Proof. S_t is again a convex gradient and hence optimal from the common source. Convexity of $-\log\det$ applied to $(1 - t)D_a S_0 + tD_a S_1$, together with the Jacobian identity, gives the entropy inequality. This is also the standard generalized displacement convexity of entropy (McCann, 1997; Ambrosio et al., 2008). Pointwise λ-convexity of V supplies

$$V(S_t(x)) \leq (1 - t)V(S_0(x)) + tV(S_1(x)) - \tfrac{\lambda}{2} t(1 - t)|S_1(x) - S_0(x)|^2. \tag{B.9}$$

For every pair (x, y), the argument of W is the affine interpolation between $S_0(x) - S_0(y)$ and $S_1(x) - S_1(y)$. Its convexity gives the interaction inequality, with no negative curvature term needed. Finally,

$$\|S_t - \mathrm{Id}\|_{L^2(\mu)}^2 = (1 - t) \, \|S_0 - \mathrm{Id}\|_{L^2(\mu)}^2 + t \, \|S_1 - \mathrm{Id}\|_{L^2(\mu)}^2 - t(1 - t) \, \|S_1 - S_0\|_{L^2(\mu)}^2 . \tag{B.10}$$

Adding the four contributions proves the claim. The rough-map version follows directly from the almost-everywhere Jacobians and generalized entropy convexity, not from smoothness of either endpoint. □

Corollary B.5 (Uniqueness and map-space quadratic growth)**.** *The minimizer in Proposition B.3 is unique. If $S^\star$ is its Brenier map, then every finite-energy comparison satisfies*

$$\mathcal{J}_{h,\mu}(S_{\#}\mu) - \mathcal{J}_{h,\mu}(S^\star{}_{\#}\mu) \geq \frac{\alpha_h}{2}\,\|S - S^\star\|^2_{L^2(\mu)} \geq \frac{\alpha_h}{2} W_2^2(S_{\#}\mu, S^\star{}_{\#}\mu). \tag{B.11}$$

Proof. Apply (B.8) to a minimizer and a comparison, use minimality on the left, divide by $t > 0$, and let $t \downarrow 0$. If both endpoints minimize, the resulting positive squared distance must vanish. Their pushforwards are then equal. □

The same source is essential in this proof. It gives no inequality comparing two JKO minimizers generated from different sources. In particular (B.8) is not a proof that the JKO resolvent is $(1+\lambda h)^{-1}$-Lipschitz in W_2 for general d.

C Full-space support and the ambient residual

C.1 The matrix divergence identity

The entropy variation in source coordinates can be written without differentiating an inverse map. We give the index calculation because a transposition error here would change the certificate.

Lemma C.1 (Divergence of an inverse Hessian)**.** *Let $K = D^2u \succ 0$ and $A = K^{-1}$, with $u \in C^3$. Define the divergence of a matrix by $(\operatorname{div} A)_j = \sum_i \partial_i A_{ij}$. Then*

$$\operatorname{div} A = -A\nabla \log\det K, \qquad \rho^{-1}\operatorname{div}(\rho A) = A(\nabla \log\rho - \nabla \log\det K) \tag{C.1}$$

for every positive C^1 density ρ.

Proof. Differentiating $AK = I$ gives $\partial_i A = -A(\partial_i K)A$. Therefore

$$\begin{aligned}(\operatorname{div} A)_j &= -\sum_{i,a,b} A_{ia}(\partial_i K_{ab})A_{bj} = -\sum_b A_{bj}\sum_{i,a} A_{ia}(\partial_b K_{ai}) \\ &= -\sum_b A_{bj}\partial_b \log\det K.\end{aligned} \tag{C.2}$$

The second equality uses symmetry of the third derivatives of u. The last uses Jacobi's formula and symmetry of A. The product rule proves the density-weighted identity. □

The identity need not hold for an arbitrary matrix field K. Its Hessian structure, not merely positive eigenvalues, is part of the argument.

C.2 A distributional integration-by-parts inequality

Lemma C.2 (The singular derivative has the correct sign)**.** *Under Assumptions A.2 and A.3, let S be any Brenier comparison map in $L^2(\mu)$. Then*

$$\int \operatorname{tr}(K^{-1}D_a S)\,\mathrm{d}\mu \leq -\int S\cdot K^{-1}(s_\mu - \nabla \log\det K)\,\mathrm{d}\mu. \tag{C.3}$$

Both sides are finite. Equality holds when DS has no singular part.

Proof. Write the positive semidefinite matrix-valued measure DS as $(D_a S)\,\mathrm{d}x + D_s S$. Set $A = K^{-1}$ and choose nonnegative smooth cutoffs χ_R equal to one on the ball of radius R, supported on the ball of radius $2R$, satisfying $\|\nabla\chi_R\|_\infty \leq c/R$, and increasing pointwise to one as R increases. The test matrix $\chi_R\rho_\mu A$ is positive semidefinite. Hence

$$\begin{aligned}\int \chi_R\rho_\mu \operatorname{tr}(AD_aS)\,\mathrm{d}x &\leq \int \operatorname{tr}(\chi_R\rho_\mu A\,\mathrm{d}DS) \\ &= -\int S\cdot\operatorname{div}(\chi_R\rho_\mu A)\,\mathrm{d}x.\end{aligned} \tag{C.4}$$

Pairing two positive semidefinite matrices has nonnegative trace, which justifies the first inequality even for the singular measure. The integration by parts is the definition of the distributional derivative; all objects are locally integrable because the convex potential is locally Lipschitz.

The derivative of the cutoff creates an error bounded by

$$\left| \int S \cdot A \nabla \chi_R \, \mathrm{d}\mu \right| \leq \frac{c}{mR} \left(\int_{R \leq |x| \leq 2R} |S|^2 \, \mathrm{d}\mu \right)^{1/2} \mu(\{R \leq |x| \leq 2R\})^{1/2}, \tag{C.5}$$

which tends to zero. For the remaining term, Lemma C.1 and bounded D^2T show that $\rho_\mu^{-1} \operatorname{div}(\rho_\mu A) \in L^2(\mu)$. Indeed $|\nabla \log \det K| \leq \sqrt{d}\, m^{-1} \left\| D^2 T \right\|_{\mathrm{F}}$. Its product with S is integrable by Cauchy–Schwarz, so dominated convergence applies.

The integrand on the left of (C.4) is nonnegative. Monotone convergence therefore proves (C.3) and at the same time proves finiteness of that integral. If $D_s S = 0$, the first line of (C.4) is an equality and remains so in the limit. □

The direction of (C.3) is decisive: after multiplication by the minus sign in the entropy support inequality, the singular derivative improves the lower bound. Discarding it with the opposite sign would invalidate comparison with a rough optimum.

C.3 Entropy support on the unrestricted target domain

Proposition C.3 (Entropy support for a smooth candidate)**.** *For a finite-entropy target $\eta = S_{\#}\mu$ and candidate T,*

$$\operatorname{Ent}(\eta) - \operatorname{Ent}(T_{\#}\mu) \geq \int K^{-1}(s_\mu - \nabla \log \det K) \cdot (S - T) \, \mathrm{d}\mu. \tag{C.6}$$

Proof. For positive definite matrices K, B, convexity of $-\log\det$ yields

$$-\log\det B + \log\det K \geq -\operatorname{tr}(K^{-1}(B - K)). \tag{C.7}$$

Apply it almost everywhere with $B = D_a S$. The Jacobian identities for S and T give

$$\operatorname{Ent}(\eta) - \operatorname{Ent}(T_{\#}\mu) \geq d - \int \operatorname{tr}(K^{-1} D_a S) \, \mathrm{d}\mu. \tag{C.8}$$

By Lemma C.2, the right side is at least $d + \int S \cdot K^{-1}(s_\mu - \nabla \log \det K) \, \mathrm{d}\mu$. Apply the equality case of the same integration by parts to T, whose derivative is K, to obtain

$$\int T \cdot K^{-1}(s_\mu - \nabla \log \det K) \, \mathrm{d}\mu = -d. \tag{C.9}$$

Subtracting this identity gives (C.6). The trace terms are integrable by the preceding lemma; no unjustified cancellation of divergent integrals occurs. □

An alternative interpretation follows by differentiating (B.2):

$$\nabla \log \rho_{T_{\#}\mu}(T(x)) = K^{-1}(s_\mu - \nabla \log \det K). \tag{C.10}$$

Thus the entropy part of the residual is the actual spatial score at the proposed law, pulled back to the common source. It is not the differential of a discretized entropy estimator.

C.4 Potential, interaction, and proximal support

For the potential term, ordinary convexity with curvature λ yields

$$\int [V(S) - V(T)] \, \mathrm{d}\mu \geq \int \nabla V(T) \cdot (S - T) \, \mathrm{d}\mu + \frac{\lambda}{2} \|S - T\|^2_{L^2(\mu)}. \tag{C.11}$$

For the interaction term, write $D(x) = S(x) - T(x)$ and use convexity for each pair:

$$\begin{aligned}\frac{1}{2}\iint [W(S(x) - S(y)) - W(T(x) - T(y))]\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\ \geq \frac{1}{2}\iint \nabla W(T(x) - T(y)) \cdot (D(x) - D(y))\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\ = \int I_\mu[T](x) \cdot D(x)\,\mathrm{d}\mu(x). \end{aligned} \tag{C.12}$$

Evenness of W makes ∇W odd, so the two halves of the middle expression are equal. Polynomial growth, linear growth of T, the source moments, and $D \in L^2(\mu)$ justify Fubini and the linear terms. The comparison interaction itself is finite because only finite-energy comparisons are being used.

Finally, since both S and T are optimal from μ,

$$\frac{W_2^2(S_\#\mu, \mu) - W_2^2(T_\#\mu, \mu)}{2h} = \frac{1}{h}\int (T - x) \cdot (S - T)\,\mathrm{d}\mu + \frac{1}{2h}\|S - T\|_{L^2(\mu)}^2. \tag{C.13}$$

Adding (C.6), (C.11), (C.12), and (C.13) proves (7). This completes the full-domain part of Theorem 1.

C.5 Why the residual is in the required space

Let $C_T = |T(0)|$ and suppose $|\nabla V(y)| \leq c_V(1 + |y|^p)$ and $|\nabla W(y)| \leq c_W(1 + |y|^q)$. Then

$$|\nabla V(T(x))| \leq c_V[1 + (C_T + M|x|)^p]. \tag{C.14}$$

By Jensen and the polynomial bound,

$$\int |I_\mu[T](x)|^2\,\mathrm{d}\mu(x) \leq \iint |\nabla W(T(x) - T(y))|^2\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) < \infty. \tag{C.15}$$

The proximal term is square-integrable by linear growth. The score term is bounded in norm by

$$\left\|K^{-1}(s_\mu - \nabla \log\det K)\right\|_{L^2(\mu)} \leq m^{-1}\|s_\mu\|_{L^2(\mu)} + \sqrt{d}\,m^{-2}\left\|D^2T\right\|_{\infty,\mathrm{F}}. \tag{C.16}$$

Thus every pairing used in the support inequality is justified in $L^2(\mu)$.

C.6 The complete certificate proof

Let $d_T = \|T - S^\star\|_{L^2(\mu)}$. The support inequality at $S^\star$ implies

$$P_h(T) - \mathcal{J}^\star \leq Rd_T - \frac{\alpha_h}{2}d_T^2 \leq \frac{R^2}{2\alpha_h}. \tag{C.17}$$

The last inequality is the scalar identity $Rd - \alpha_h d^2/2 = R^2/(2\alpha_h) - \alpha_h(d - R/\alpha_h)^2/2$. On the other hand, (B.11) gives $\alpha_h d_T^2/2 \leq P_h(T) - \mathcal{J}^\star$. Combining the two inequalities gives $\alpha_h d_T^2 \leq Rd_T$. When $d_T = 0$ the conclusion is immediate; otherwise division by $\alpha_h d_T > 0$ yields $d_T \leq R/\alpha_h$. Finally (A.4) gives the Wasserstein bound.

Notice why a one-sided support argument alone is not enough for the stated radius: dropping the positive gap would give only $d_T \leq 2R/\alpha_h$. The second half of the argument, namely common-source quadratic growth at the minimizer, is what supplies the sharper constant.

D Sharp constants, dissipation, and summed error budgets

D.1 Translation sharpness

Proposition D.1 (Neither constant can be uniformly improved)**.** *Fix $h > 0$ and $\lambda > 0$. Take $W = 0$, $V(x) = \lambda|x|^2/2$, $\mu = N(0, \lambda^{-1}I)$, and $T_a(x) = x + a$. Then $S^\star = \mathrm{Id}$ and*

$$r_h[T_a] = \alpha_h a, \quad P_h(T_a) - \mathcal{J}^\star = \frac{\alpha_h}{2}|a|^2, \quad W_2(T_{a\#}\mu, \mu) = |a|. \tag{D.1}$$

Hence both inequalities in (8) *are equalities with $R = \|r_h[T_a]\|_{L^2(\mu)}$.*

Proof. The energy differs from $\mathrm{KL}(\nu\|\mu)$ by a constant, so μ minimizes the energy and also the JKO problem based at itself. Translation does not change entropy. The potential increases by $\lambda|a|^2/2$, since the mean of μ is zero, and the transport penalty increases by $|a|^2/(2h)$. The score at the translated density, evaluated at $x+a$, is $-\lambda x$, while the potential gradient there is $\lambda(x+a)$. This gives the constant residual. The translation coupling attains cost $|a|^2$, and the difference of means provides the matching lower bound for Wasserstein distance. □

This example proves sharpness of the universal constants, not sharpness of every computed audit. A loose norm enclosure can remain very conservative even when the analytic inequality is optimal.

D.2 A VALID ENERGY-DECREASE TEST

Proposition D.2 (Absolute and relative energy guarantees)**.** *Let $L = \|T - \mathrm{Id}\|_{L^2(\mu)}$ and let $g \geq P_h(T) - \mathcal{J}^\star$ be any valid gap certificate. Then*

$$\mathcal{F}(T_{\#}\mu) - \mathcal{F}(\mu) \leq -\frac{L^2}{2h} + g. \tag{D.2}$$

If $g \leq \theta L^2/(2h)$ with $0 \leq \theta < 1$, then

$$\mathcal{F}(T_{\#}\mu) \leq \mathcal{F}(\mu) - \frac{1-\theta}{2h} L^2. \tag{D.3}$$

For the residual certificate it suffices that $R^2 \leq \theta \alpha_h L^2/h$.

Proof. The source μ is an admissible competitor with $\mathcal{J}_{h,\mu}(\mu) = \mathcal{F}(\mu)$. Thus $P_h(T) \leq \mathcal{J}^\star + g \leq \mathcal{F}(\mu) + g$. Substitute the exact identity $P_h(T) = \mathcal{F}(T_{\#}\mu) + L^2/(2h)$ and rearrange. The last condition is obtained by substituting $g = R^2/(2\alpha_h)$. □

For implementation, the relative test requires an *upper* bound on g and a *lower* bound on L^2. Replacing both by approximate point estimates does not preserve the logic. If L is close to zero, an absolute energy allowance can be more appropriate. A relative test demanding zero error at an equilibrium need not terminate under generic interval arithmetic, even though the candidate is mathematically exact. An identity that proves $r_h \equiv 0$ can certify such an equilibrium directly.

D.3 DISCRETE ACTION AND CONTINUITY OF INTERPOLATIONS

Proposition D.3 (An energy-based action estimate)**.** *Suppose $\mu_{k+1} = T_{k\#}\mu_k$, $L_k = W_2(\mu_k, \mu_{k+1})$, and valid gaps g_k are available. If $\mathcal{F}$ is bounded below by $F_{\inf}$ on these iterates, then*

$$\sum_{k=0}^{N-1} \frac{L_k^2}{h_k} \leq 2[\mathcal{F}(\mu_0) - F_{\inf}] + 2\sum_{k=0}^{N-1} g_k. \tag{D.4}$$

For their constant-speed geodesic interpolation $\rho(t)$, with $t_N = \sum_{k<N} h_k$,

$$W_2(\rho(s), \rho(t)) \leq \sqrt{|t-s|}\left(\sum_{k<N} \frac{L_k^2}{h_k}\right)^{1/2}, \quad 0 \leq s \leq t \leq t_N. \tag{D.5}$$

Proof. Sum (D.2); intermediate energies telescope. The metric speed on step k is L_k/h_k. The length bound and Cauchy–Schwarz give

$$W_2(\rho(s), \rho(t)) \leq \int_s^t |\dot\rho|(r)\,\mathrm{d}r \leq \sqrt{t-s}\left(\int_s^t |\dot\rho|^2(r)\,\mathrm{d}r\right)^{1/2}. \tag{D.6}$$

The integral over the entire time interval is exactly the sum in (D.4). □

An action bound is useful for compactness, but it is not by itself a physical-time error bound. In particular it does not determine which limiting evolution is obtained if the optimization defects fail to vanish. The residual stability theorem supplies that additional identification.

D.4 Objective and distance tolerances are different currencies

An objective-gap allowance g gives the radius $\sqrt{2g/\alpha_h}$ from quadratic growth. A full residual allowance R gives R/α_h directly, and the induced objective gap is $R^2/(2\alpha_h)$. These are consistent but should not be mixed without their factors of h.

For example, if $R_h \le ch$ and $\lambda \ge 0$, then

$$g_h \le \frac{c^2h^3}{2(1+\lambda h)}, \qquad \eta_h \le \frac{ch^2}{1+\lambda h}. \tag{D.7}$$

A constant residual tolerance instead gives $g_h = O(h)$ and $\eta_h = O(h)$ as $h \downarrow 0$. Such a local radius can accumulate to an order-one trajectory error over $O(h^{-1})$ steps. The physical-time theorem explains the appropriate accumulation directly, without needing a stability estimate for the discrete resolvent.

Conversely, a desired one-step radius η_h allows a residual $R_h \le (\lambda + h^{-1})\eta_h$. Reducing the step while keeping the same radius can make this residual test easier, even though it increases the number of steps. This is why an outer controller must allocate a total error budget rather than enforce one fixed radius independently at every step.

D.5 A posteriori lower bounds need additional information

It is tempting to interpret a large value of R as proof of a large gap. This is not valid: R is only an upper norm enclosure, and even the exact residual norm can be large in a very steep direction while the gap is small. The theorem supplies a one-sided estimate.

A valid lower gap bound can instead be obtained by exhibiting a feasible improvement. If T_s is an admissible comparison and valid evaluations give $P_h(T) \ge \underline{P}$ and $P_h(T_s) \le \overline{P}_s$, then

$$P_h(T) - \mathcal{J}^\star \ge P_h(T) - P_h(T_s) \ge \underline{P} - \overline{P}_s. \tag{D.8}$$

The directional construction in Appendix F produces such improvements without knowing the unrestricted minimizer. The primal–dual construction later supplies upper rather than lower gaps; keeping these signs explicit prevents an overoptimistic stopping rule.

E Auditable neural transports and known source scores

E.1 A convex ridge architecture

Consider the potential

$$u_\theta(x) = \tfrac{1}{2}x^TBx + b\cdot x + \sum_{j=1}^{m} c_j \,\mathrm{softplus}(w_j\cdot x + \beta_j), \qquad B = B^T \succeq m_0 I, \quad c_j \ge 0, \tag{E.1}$$

where $\mathrm{softplus}(t) = \log(1+e^t)$. With $\sigma(t) = (1+e^{-t})^{-1}$,

$$T_\theta(x) = Bx + b + \sum_j c_j\sigma(w_j\cdot x + \beta_j)w_j, \tag{E.2}$$

$$DT_\theta(x) = B + \sum_j c_j\sigma_j(1-\sigma_j)w_jw_j^T. \tag{E.3}$$

Thus

$$m_0 I \preceq DT_\theta(x) \preceq B + \tfrac{1}{4}\sum_j c_jw_jw_j^T. \tag{E.4}$$

All required global constants are explicit functions of the coefficients. For example,

$$|T_\theta(x)| \le \|B\|_{\mathrm{op}}\,|x| + |b| + \sum_j c_j|w_j|, \qquad \left\|D^2T_\theta(x)\right\|_{\mathrm{F}} \le \tfrac{1}{4}\sum_j c_j|w_j|^3. \tag{E.5}$$

The second estimate uses $|\sigma''| = |\sigma(1-\sigma)(1-2\sigma)| \le 1/4$; it is not the sharpest constant, but it is a valid one. Higher derivatives are bounded by explicit polynomials in $\sigma \in [0,1]$. A stable implementation can enclose the sigmoid directly rather than evaluating e^t when t is large.

This construction is related to input-convex neural architectures (Amos et al., 2017); we do not claim that convex ridge networks were introduced here. Their role is to make global curvature and tail bounds visible. With fixed w_j, β_j and a convex admissible coefficient set, T_θ is affine in the trainable coefficients. The lifted objective is convex on that coefficient set by Proposition B.4. Training the feature locations or directions destroys this finite-dimensional convexity, but not the validity of the certificate for the final admissible map.

E.2 A SIGNED BUMP ARCHITECTURE FOR GRAPH APPROXIMATION

Positive ridge weights should not be assumed universal in a derivative norm. For the approximation theorem we use a different, explicitly defined neural family. Choose an even nonnegative C^∞ bump β supported on $[-1, 1]$, positive inside, and normalized by $\int \beta = 1$. One concrete choice is a normalization of

$$\beta_0(t) = \begin{cases} \exp[-1/(1-t^2)], & |t| < 1, \\ 0, & |t| \geq 1. \end{cases} \tag{E.6}$$

Let $\beta_d(x) = \prod_{i=1}^d \beta(x_i)$ and define

$$u_\theta(x) = \tfrac{1}{2} x^T B x + b \cdot x + c + \sum_{j=1}^m a_j \epsilon_j^{-d} \beta_d((x - z_j)/\epsilon_j). \tag{E.7}$$

The coefficients a_j may have either sign. This is a finite tensor-product radial-basis network for the potential; multiplication is an explicitly specified operation. The candidate map is its gradient. It is not claimed to be a standard positive-weight ICNN.

Outside finitely many compact boxes, $D^2 u_\theta = B$. On those boxes, curvature is checked independently. A sufficient but possibly very conservative test is

$$\lambda_{\min}(B) - \sum_j |a_j| \epsilon_j^{-d-2} \left\| D^2 \beta_d \right\|_\infty > 0. \tag{E.8}$$

The complete certification procedure instead subdivides the boxes and encloses the actual Hessian there. If its true minimum eigenvalue is strictly positive, a convergent interval Hessian extension eventually verifies a smaller positive lower bound. The proof is compactness plus uniform continuity. The coarse sum-of-absolute-values test is not used to assert density.

The activation β_0 is smooth across ± 1: every derivative inside is a polynomial in $(1 - t^2)^{-1}$ and t times the exponential, which tends to zero faster than any power as $t \to \pm 1$. Its derivatives and extrema on compact intervals admit convergent enclosures by domain splitting and elementary exponential bounds. Rational coefficient approximations are allowed with strict curvature margins. None of this implies that such an audit is inexpensive in high dimension.

E.3 A COUPLED POLYNOMIAL ENERGY FAMILY

An explicit nonseparable family in the energy class is

$$V(x) = \tfrac{1}{2} x^T K_V x + b_V \cdot x + \sum_{j=1}^{m_V} a_j (v_j \cdot x)^4, \tag{E.9}$$

$$W(x) = \tfrac{1}{2} x^T K_W x + \sum_{j=1}^{m_W} b_j (w_j \cdot x)^4, \tag{E.10}$$

where $K_V \succeq \lambda I$, $K_W \succeq 0$, and all quartic coefficients are nonnegative. General directions v_j, w_j need not share an orthogonal basis; the functions can genuinely couple all coordinates. The interaction is even and convex. Its force can be expressed using moments of $T(X)$ through order three, so a double integration can be reduced to finitely many single-law moment audits. This family is an example of a classical convex interaction energy, not a new physical model.

For $W(x) = b(w \cdot x)^4$, let $Y = T(X)$, $a = w \cdot T(x)$, and $M_k = \mathbb{E}(w \cdot Y)^k$. Then

$$I_\mu[T](x) = 4b\, w\{a^3 - 3a^2 M_1 + 3a M_2 - M_3\}. \tag{E.11}$$

Each moment can be enclosed with Gaussian-tail integration. Interval propagation through (E.11) is sound even though it may overestimate correlations between repeated appearances of the same moment. Refining moment bounds restores convergence.

E.4 THE SOURCE SCORE UNDER A GENERAL DIFFEOMORPHISM

Proposition E.1 (Score representation). *Let $G : \mathbb{R}^d \to \mathbb{R}^d$ be a C^2 orientation-preserving diffeomorphism and $\mu = G_{\#}\gamma_d$. Then* (5) *holds. If $\left\|DG^{-1}\right\|$ and D^2G are uniformly bounded, the score is square-integrable and has an affine bound in the latent variable.*

Proof. The density identity is $\log \rho_\mu(G(z)) = \log \phi_d(z) - \log\det DG(z)$. Differentiate with respect to z and multiply by $DG(z)^{-T}$. Since $\nabla \log \phi_d(z) = -z$, this gives the formula. Jacobi's formula bounds each component of $\nabla \log\det DG$ by the Frobenius norm of DG^{-1} times the corresponding derivative of DG. Hence

$$|s_\mu(G(z))| \le L_{G^{-1}}\big(|z| + \sqrt{d}\, L_{G^{-1}} M_{2,G}\big), \tag{E.12}$$

where $M_{2,G} = \left\|D^2G\right\|_{\infty,\mathrm{F}}$. Gaussian second moments establish square integrability. □

Here DG^{-1} denotes the inverse matrix at the corresponding point, not differentiation of a separately approximated inverse function. Evaluating (5) at $G(z)$ uses only the forward generator and its derivatives.

E.5 CLOSURE OF THE SOURCE REPRESENTATION UNDER ACCEPTED STEPS

Suppose $G_{k+1} = T_k \circ G_k$ and every T_k satisfies (3). For a finite number of accepted steps, the composition is a global diffeomorphism and

$$\|DG_{k+1}\|_\infty \le M_k \|DG_k\|_\infty, \qquad \left\|DG_{k+1}^{-1}\right\|_\infty \le m_k^{-1} \left\|DG_k^{-1}\right\|_\infty. \tag{E.13}$$

The second derivative chain rule gives

$$M_{2,G_{k+1}} \le M_{2,T_k} \|DG_k\|_\infty^2 + M_k M_{2,G_k}. \tag{E.14}$$

These finite recurrences give computable, though potentially poor, bounds. There is no claim that their products stay uniformly bounded over infinitely many steps.

Linear growth of the composition implies moments of every order under the Gaussian reference. Bounded positive Jacobian determinants above and below give finite absolute entropy through the density change of variables. Equation (E.12) then verifies the remaining source assumptions. This proves closure of the *source representation* for every finite accepted trajectory. It does not assert that the composite G_k is Brenier; that property generally fails and is unnecessary.

F PARAMETER-VISIBLE AND MISSING DIRECTIONS

F.1 AN EXACT HILBERT-SPACE DECOMPOSITION

Let $\mathsf{H} = L^2(\mu;\mathbb{R}^d)$ and consider a differentiable parameterization of admissible maps near θ. Put $A = D_\theta T_\theta : \mathbb{R}^p \to \mathsf{H}$. In differentiable parameter directions the first variation of the lifted objective is

$$D_\theta P_h(T_\theta)[v] = \langle r_h[T_\theta], Av\rangle_{\mathsf{H}}. \tag{F.1}$$

For smooth compactly supported variations this follows directly from integration by parts; for finite neural variations the growth bounds provide dominated convergence. At a boundary of a constrained parameter set, only feasible directions can be used for optimality. The algebra below concerns the linear span of the available derivatives, not a claim that every signed parameter direction is feasible.

Proposition F.1 (Visible-residual identity). *Let $r \in \mathsf{H}$, $H = A^*A$, and $b = A^*r$. Then*

$$\Pi r = AH^\dagger b, \qquad \|\Pi r\|_{\mathsf{H}}^2 = b^T H^\dagger b, \qquad \|r\|_{\mathsf{H}}^2 = b^T H^\dagger b + \|r - \Pi r\|_{\mathsf{H}}^2, \tag{F.2}$$

where Π is the orthogonal projection onto $\operatorname{range} A$. *The first term is unchanged by any invertible reparameterization of the same tangent span.*

Proof. The finite-dimensional range is closed. The kernels of H and A agree because $v^T H v = \|Av\|^2$. Also b is orthogonal to $\ker A$, so $b \in \operatorname{range} H$. Let $c = H^\dagger b$. Then $Hc = b$, and for every v,

$$\langle r - Ac, Av \rangle = v^T(b - Hc) = 0. \tag{F.3}$$

Thus $Ac = \Pi r$. Its squared norm is $c^T H c = b^T H^\dagger b$. Orthogonal Pythagoras gives the decomposition. A nonsingular coordinate change does not change $\operatorname{range} A$ and therefore does not change the norm of the orthogonal projection, regardless of how its matrix formula is written. □

A small Euclidean parameter gradient has two possible defects: poor parameter scaling and missing tangent directions. The Gram-normalized first term removes the former but not the latter. In particular computing a pseudoinverse does not reconstruct an unrepresented direction.

Example F.2 (A scaling obstruction)**.** Take the translation model $T_\theta(x) = x + \epsilon\theta e_1$ in the Gaussian setting of Proposition D.1. To represent the same nonzero shift ae_1, choose $\theta = a/\epsilon$. The full residual norm is $\alpha_h|a|$, while the scalar parameter gradient is $\epsilon\alpha_h a$. It tends to zero as $\epsilon \downarrow 0$ although the represented law and its error are unchanged. Here $H = \epsilon^2$, so the normalized term correctly recovers $\alpha_h^2 a^2$.

F.2 A STRICT REPRESENTATION GAP AT AN AFFINE OPTIMUM

Proposition F.3 (Quartic energy and an invisible Hermite direction)**.** *Let $\mu = N(0,1)$, $h > 0$, $W = 0$, $V(y) = \lambda y^2/2 + \kappa y^4/4$, with $\lambda \geq 0$, $\kappa > 0$. In the family $T_s(z) = sz$, $s > 0$, there is a unique minimizer $s_\star$. At that minimizer the parameter gradient is zero but*

$$r_h[T_{s_\star}](z) = \kappa s_\star^3(z^3 - 3z), \qquad \|r_h[T_{s_\star}]\|^2_{L^2(\gamma_1)} = 6\kappa^2 s_\star^6. \tag{F.4}$$

The unrestricted JKO gap of this map is strictly positive.

Proof. Up to a constant, the restricted objective is

$$p(s) = -\log s + \frac{\lambda}{2}s^2 + \frac{3\kappa}{4}s^4 + \frac{(s-1)^2}{2h}. \tag{F.5}$$

It diverges at both endpoints of $(0, \infty)$ and has second derivative $1/s^2 + \lambda + 9\kappa s^2 + 1/h > 0$. Thus it has a unique minimizer, satisfying

$$\lambda s_\star + 3\kappa s_\star^3 + (s_\star - 1)/h - 1/s_\star = 0. \tag{F.6}$$

The full residual is $[\lambda s + (s-1)/h - 1/s]z + \kappa s^3 z^3$. Substitute (F.6) to get (F.4). Gaussian moments $\mathbb{E}Z^2 = 1$, $\mathbb{E}Z^4 = 3$, and $\mathbb{E}Z^6 = 15$ give $15 - 18 + 9 = 6$ for the squared Hermite norm.

To prove strict suboptimality, choose a nonzero nonnegative C^∞ compactly supported cutoff χ that is positive on a region where $r_h \neq 0$, and set $v = -\chi r_h$. Then

$$\langle r_h, v \rangle = -\int \chi r_h^2 \, \mathrm{d}\gamma_1 < 0. \tag{F.7}$$

For sufficiently small $t > 0$, the map $T_{s_\star} + tv$ has derivative at least $s_\star/2 > 0$ because v' is bounded. It is an admissible increasing diffeomorphism with linear tails. Differentiating its objective at $t = 0$ gives the negative pairing above. A small feasible step therefore has strictly smaller objective. This proves that the unrestricted minimum is strictly below the affine-family minimum. The argument does not infer a lower gap merely from a large upper residual bound. □

The cutoff is necessary. The untruncated direction $-(z^3 - 3z)$ has a derivative unbounded below and would eventually destroy monotonicity for every positive step size. This example illustrates why a formal negative-gradient direction need not be globally admissible in a constrained transport family.

F.3 Quantitative certified feature insertion

Proposition F.4 (A feasible direction supplies a lower gap)**.** *Let T be admissible, $v = \nabla\psi$ be a smooth compactly supported field, and suppose $DT \succeq mI$. Choose $\bar{t} > 0$ with $\bar{t}\,\|Dv\|_\infty \le m/2$. Suppose certified constants $a > 0$ and $L < \infty$ satisfy*

$$\langle r_h[T], v\rangle \le -a, \qquad \sup_{0\le t\le \bar{t}} \frac{\mathrm{d}^2}{\mathrm{d}t^2} P_h(T + tv) \le L. \tag{F.8}$$

For $t_0 = \min\{\bar{t}, a/L\}$ when $L > 0$, one has

$$P_h(T) - \mathcal{J}^\star \ge at_0 - \frac{L}{2}t_0^2 > 0. \tag{F.9}$$

When $L = 0$, choose $t_0 = \bar{t}$ and the lower bound is $a\bar{t}$.

Proof. The Hessian lower bound makes every $T + tv$ in the stated interval admissible. Taylor's theorem with integral remainder gives $P_h(T+tv) \le P_h(T) - at + Lt^2/2$. The unrestricted optimum is at most this feasible value. For $L > 0$, $t_0 \le a/L$ gives $at_0 - Lt_0^2/2 \ge at_0/2 > 0$. The $L = 0$ case is immediate. □

The required curvature is itself explicit. With $K_t = DT + tDv$,

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} P_h(T + tv) = &\int \left[v^T \nabla^2 V(T + tv)v + \mathrm{tr}(K_t^{-1} Dv K_t^{-1} Dv) + \frac{|v|^2}{h} \right] \mathrm{d}\mu \\ &+ \frac{1}{2} \iint (v(x) - v(y))^T \nabla^2 W(T_t(x) - T_t(y))(v(x) - v(y))\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y).\end{aligned} \tag{F.10}$$

Dv is symmetric, so the trace term is nonnegative and bounded using $K_t \succeq mI/2$. For a polynomial interaction the remaining integrals have explicit moment majorants. In a network with signed compact bump features, v is itself a feature insertion or a finite linear combination of features. The proposition therefore gives a mathematically justified enrichment action, not merely the advice to use a larger network.

F.4 What an unsuccessful certificate diagnoses

Suppose a candidate has a small visible term in (F.2), a nontrivial lower bound on a feasible improvement from (F.9), and an upper certificate larger than the requested tolerance. Then a representation or restricted-optimization limitation is evidenced by an actual feasible better law. Without the feasible improvement, the data may also indicate a loose audit or an ill-conditioned full objective. The controller is allowed to refine the audit, restart optimization, or add a feature; it is not allowed to label one of these explanations as proved from the parameter gradient alone.

G Global deterministic evaluation of a residual norm

G.1 The evaluation contract

Let $f : \mathbb{R}^d \to \mathbb{R}^q$ be the latent-coordinate function whose Gaussian norm is needed. For the endpoint certificate, $f = r_h[T] \circ G$ and $q = d$. The inputs to the evaluator are an exact finite model description, sound interval extensions on compact boxes, and a nonnegative polynomial P such that

$$|f(z)| \le P(|z_1|, \ldots, |z_d|) \quad \text{for all } z. \tag{G.1}$$

Sound means that every exact value over a box lies in the returned interval. Convergent means that over any fixed compact set the widths tend uniformly to zero as box diameters and arithmetic roundoff tend to zero. These are assumptions on the supplied function evaluation mechanism, not consequences of observing apparent convergence of ordinary quadrature.

For the neural classes in Appendix E, lower curvature bounds keep inverse Jacobians away from singularity; source scores grow at most linearly in z; polynomial potentials and interactions give polynomial residual growth. Thus a polynomial envelope can be built from the network coefficients and Gaussian moment bounds. In a general implementation, validity of those ingredient bounds must also be checked.

G.2 CORE BOX BOUNDS

Let $Q_L = [-L, L]^d$ and let $\mathcal{P}$ be a finite partition into boxes. Gaussian box probabilities factor:

$$p_B = \gamma_d(B) = \prod_{i=1}^{d} [\Phi(b_i) - \Phi(a_i)], \qquad B = \prod_i [a_i, b_i]. \tag{G.2}$$

They can be enclosed independently with a validated normal distribution function. Suppose an interval evaluation supplies $\underline{q}_B \le |f(z)|^2 \le \overline{q}_B$ for every $z \in B$. Nonnegativity lets us replace $\underline{q}_B$ by $\max(0, \underline{q}_B)$. Then

$$\sum_B p_B \underline{q}_B \le \int_{Q_L} |f|^2 \, \mathrm{d}\gamma_d \le \sum_B p_B \overline{q}_B. \tag{G.3}$$

If probabilities are also intervals, lower products use lower nonnegative endpoints and upper products use upper endpoints. All sums are rounded outward in a numerical implementation.

Alternatively, suppose a verified center evaluation gives $|f(z_B)| \le a_B$ and a derivative bound gives $\|Df\|_{\mathrm{op}} \le L_B$ on the box. Let $\delta_B = \sup_{z \in B} |z - z_B|$. The mean-value theorem gives

$$|f(z)| \le a_B + L_B \delta_B, \tag{G.4}$$

which proves (16). The method can mix direct interval bounds and center-plus-derivative bounds, retaining the smaller valid upper bound. This improves tightness without changing soundness.

G.3 EXACT GAUSSIAN TAIL RECURRENCES

Write

$$M_k(L) = 2 \int_L^\infty z^k \phi(z) \, \mathrm{d}z, \quad L \ge 0, \qquad \phi(z) = (2\pi)^{-1/2} e^{-z^2/2}. \tag{G.5}$$

For integer $k \ge 0$,

$$\begin{aligned} M_0(L) &= \operatorname{erfc}(L/\sqrt{2}), \qquad M_1(L) = 2\phi(L), \\ M_k(L) &= 2L^{k-1}\phi(L) + (k-1)M_{k-2}(L), \qquad k \ge 2. \end{aligned} \tag{G.6}$$

The recurrence follows by integrating $z^{k-1}(-\phi'(z))$ by parts. Every term is nonnegative, making outward interval evaluation straightforward. In particular the recurrence does not subtract two nearly equal tail integrals.

Expand the squared envelope as $P(|z|)^2 = \sum_\alpha c_\alpha \prod_i |z_i|^{\alpha_i}$ with $c_\alpha \ge 0$. The union bound for the complement of Q_L and independence give

$$\int_{Q_L^c} |f|^2 \, \mathrm{d}\gamma_d \le \sum_\alpha c_\alpha \sum_{j=1}^{d} M_{\alpha_j}(L) \prod_{i \ne j} M_{\alpha_i}(0) =: \mathcal{T}_P(L). \tag{G.7}$$

This bound double-counts points lying beyond more than one face of the cube; that only makes it conservative. Every tail moment tends to zero, so $\mathcal{T}_P(L) \to 0$. A radial polynomial bound can first be converted to a polynomial in $|z_i|$ using $|z| \le \sum_i |z_i|$.

G.4 SOUNDNESS AND CONVERGENCE

Theorem G.1 (A convergent upper norm audit)**.** *Under the evaluation contract and* (G.1)*, there are finite computable intervals* $[\ell_n, u_n]$ *with*

$$0 \le \ell_n \le \int |f|^2 \, \mathrm{d}\gamma_d \le u_n, \qquad u_n - \ell_n \longrightarrow 0. \tag{G.8}$$

In particular $R_n = \sqrt{u_n}$ *is a valid upper norm bound converging to* $\|f\|_{L^2(\gamma_d)}$*.*

Proof. At precision level n, choose L_n so that a validated version of $\mathcal{T}_P(L_n)$ is at most 2^{-n}. Such an L_n is found by increasing it until the strictly positive tail target is met. On Q_{L_n}, refine the partition

and arithmetic precision until the sum of the upper core contributions minus the sum of the lower core contributions is at most 2^{-n}. Convergent interval extensions and compactness guarantee that a sufficiently fine finite partition and sufficient arithmetic precision satisfy this condition. The box probabilities are also refined to make their total contribution to the bracket width at most 2^{-n}.

Define ℓ_n as the lower core sum and u_n as the upper core sum plus the tail allowance. All quantities are computed by outward operations. The core inequalities and tail bound prove soundness. Their gap is at most $3 \cdot 2^{-n}$ after increasing the precision constants if needed, proving convergence. There is no need for u_n to be monotone; replacing it by the minimum of its previous values makes it monotone without affecting validity. □

The order of choices matters. Taking $L_n = n$ and a mesh size $1/n$ without controlling the increasingly large compact-domain modulus need not converge. The proof first fixes the tail target and then chooses the mesh and precision required on that particular core.

Corollary G.2 (Strict-slack certification)**.** *If $\|f\|_{L^2(\gamma_d)} < r_{\mathrm{target}}$, a finite audit eventually returns $R_n < r_{\mathrm{target}}$. If the exact norm equals the threshold, finite success is not guaranteed. If it is larger, no sound audit can report success.*

Proof. The first assertion follows from convergence and the positive gap between the exact norm and the threshold. The other two follow respectively from the possibility that every upper enclosure remains strictly above a boundary value and from soundness. □

G.5 Uniform interpolation-parameter audits

For the clock certificate we must bound $C_\star = \sup_{\tau\in[0,1]} \|f_\tau\|_{L^2(\gamma_d)}$, not merely one sampled value of τ. Assume joint continuity and a uniform polynomial envelope. Partition $[0,1]$ into intervals J and evaluate on $J \times B$. Define

$$U_J = \sum_B p_B \overset{\text{enclosed}}{\sup_{(\tau,z)\in J\times B}} |f_\tau(z)|^2 + \mathcal{T}_P(L), \qquad C = \sqrt{\max_J U_J}. \tag{G.9}$$

This is a valid upper bound on $C_\star$. Refining both the parameter partition and spatial partition makes it converge to $C_\star$, because the integrals are uniformly continuous in τ on the truncated compact set and the tail is uniform. More explicitly, the oscillation of $|f_\tau(z)|^2$ on each product cell tends uniformly to zero, so the maximum of the upper cell-integral sums differs from the maximum exact integral by a quantity tending to zero.

It would be safe but potentially irreducibly loose to compute only $\int \sup_\tau |f_\tau(z)|^2 \,\mathrm{d}\gamma_d$. In general the supremum and integral do not commute. Partitioning the parameter interval as in (G.9) is what proves convergence to the desired supremum of norms.

G.6 Interactions and nested evaluations

For polynomial W, the finite moment expansion such as (E.11) reduces the inner force integral to finitely many scalar expectation enclosures. Use a new precision level for those moments each time the outer audit is refined. On a fixed latent core, their propagated errors vanish uniformly; a polynomial tail bound controls the remaining domain.

For a general smooth W with polynomial derivative bounds, one may instead partition a second latent variable z' and evaluate $\nabla W(T(G(z)) - T(G(z')))$ on product boxes. A uniform-in-z tail majorant for the z' integral is available on each fixed outer core. Dominated convergence and compact uniform continuity establish the same strict-slack completeness. This is a $2d$-dimensional integration problem in the direct implementation. We make no tractability claim for it.

G.7 An explicit curse-of-dimensionality bound

Suppose a scalar or vector integrand is L_f-Lipschitz on Q_L and its norm is bounded by M_f there. A cubical grid with side length δ has box radius $\sqrt{d}\,\delta/2$. The upper-minus-lower squared-integrand oscillation is at most a constant times

$$M_f L_f \sqrt{d}\,\delta + L_f^2 d\delta^2. \tag{G.10}$$

Gaussian box probabilities sum to at most one, so the same estimate bounds the total core bracket. Reaching a squared-norm uncertainty τ therefore requires, for this simple grid, on the order of

$$\left(\frac{LM_f L_f \sqrt{d}}{\tau}\right)^d \tag{G.11}$$

boxes in the small-τ regime, up to problem-dependent constants. Adaptive anisotropic partitions can improve individual instances, but this proof gives no dimension-free replacement for the generic integration cost.

H WHY FINITE POINTWISE INFORMATION IS INSUFFICIENT

H.1 THE PRECISE ORACLE MODEL

The following result concerns an evaluator given a map through pointwise oracle access. A query consists of a real point and a finite derivative order, and returns all derivatives through that order at that point. The algorithm can choose later queries adaptively using earlier answers. It is not given a global formula, network coefficients, a uniform curvature ceiling, or a tail envelope. The impossibility result should not be read as a statement about an auditor that does possess those additional data.

Theorem H.1 (Adaptive finite-jet impossibility)**.** *Fix $h > 0$, $\mu = \gamma_1$, and $V(y) = y^2/2$, $W = 0$. Let an algorithm be deterministic and terminate after finitely many queries on every smooth increasing map whose derivative is bounded above and below by positive map-dependent constants and whose tails are affine. It cannot return a finite valid upper bound on the unrestricted JKO gap for every map in that class from the query answers alone.*

Proof. Run the algorithm on $T_0(z) = z$. Suppose it queries finitely many points, and let M exceed all of them. Since it terminates, it returns a finite number B. Choose a C^∞ nondecreasing step s satisfying $s(t) = 0$ for $t \le 0$ and $s(t) = 1$ for $t \ge 1$. For $a > 0$, set

$$T_a(z) = z + as(z - M - 1). \tag{H.1}$$

This map and all its derivatives agree with the identity near every queried point. Its derivative is at least one and at most $1 + a\,\|s'\|_\infty$, and it has affine tails. By induction on the adaptive queries, the algorithm receives exactly the same transcript and returns exactly the same B.

The identity is the unrestricted minimizer, as it is the Gaussian equilibrium. Common-source quadratic growth gives

$$P_h(T_a) - \mathcal{J}^\star \ge \frac{1 + h^{-1}}{2} \int |T_a(z) - z|^2 \,\mathrm{d}\gamma_1(z) \ge \frac{1 + h^{-1}}{2} a^2 \gamma_1([M + 2, \infty)). \tag{H.2}$$

The last Gaussian probability is strictly positive. Choosing a sufficiently large makes the gap larger than B, a contradiction. All maps used in the argument have finite moments and finite entropy; increasing a changes the global upper derivative bound but does not leave the stated class. □

The potential whose gradient is (H.1) is $z^2/2 + a\int_0^z s(t - M - 1)\,\mathrm{d}t$. It is uniformly convex. A smooth step can be obtained by integrating a compactly supported nonnegative bump. Thus the obstruction is compatible with smooth finite-feature transport architectures; it is not an example using a discontinuous or noninvertible map.

H.2 A SOURCE-SAMPLING CONSEQUENCE

Suppose an evaluator makes N independent point queries from γ_1. Choose the transition in (H.1) so far into the tail that the probability p of reaching its support is as small as desired. With probability at least $(1 - p)^N$, all observations agree with the identity. Yet the amplitude can make the true gap arbitrarily large. Consequently sample size alone cannot yield a distribution-free finite upper gap certificate over this unbounded class. A uniform envelope, a known feature description, or another restriction is indispensable.

This observation does not contradict concentration inequalities. Those inequalities require bounds on the queried random variables, a moment condition, or another specified tail class. The stochastic audit in the next appendix supplies such a bound using model information before sampling.

H.3 A FIXED QUADRATURE COUNTEREXAMPLE WITH ONE ANALYTIC FEATURE

The preceding proof uses a compact step. A separate counterexample applies to a fixed quadrature rule for a smooth analytic monotone neuron. Let

$$g(u) = \frac{u}{\sqrt{1+u^2}}, \qquad T_w(z) = z + ag((z-c)/w), \quad a, w > 0. \tag{H.3}$$

Then $T_w' \geq 1$, it is smooth and invertible, and its derivatives are globally bounded for each fixed $w > 0$. Let the source and energy be the same as above and put $\alpha = 1 + h^{-1}$.

Proposition H.2 (A fixed residual quadrature can be an invalid gap upper bound)**.** *Fix finitely many nodes z_i, nonnegative weights ω_i summing to one, and $c \notin \{z_i\}$. Then*

$$\frac{1}{2\alpha}\sum_i \omega_i r_h[T_w](z_i)^2 \longrightarrow \frac{\alpha a^2}{2},$$

$$P_h(T_w) - \mathcal{J}^\star \longrightarrow \frac{\alpha a^2}{2} + 2a\phi(c) \qquad (w \downarrow 0). \tag{H.4}$$

The limiting ratio of the actual gap to the quadrature expression can be arbitrarily large by choosing a small. Therefore no fixed rule of this form is a uniform valid upper certificate over this one-neuron family.

Proof. At any fixed $z \neq c$, $g((z-c)/w) \to \operatorname{sign}(z-c)$, $T_w'(z) \to 1$, and $T_w''(z) \to 0$. The scalar residual is

$$r_h[T_w] = T_w + (T_w - z)/h - \frac{z}{T_w'} - \frac{T_w''}{(T_w')^2}. \tag{H.5}$$

Hence $r_h[T_w](z_i) \to \alpha a \operatorname{sign}(z_i - c)$. Finite summation proves the first limit.

Since the identity minimizes the problem, direct evaluation gives

$$P_h(T_w) - \mathcal{J}^\star = -\mathbb{E}\log T_w'(Z) + a\mathbb{E}[Zg((Z-c)/w)] + \frac{\alpha a^2}{2}\mathbb{E}[g((Z-c)/w)^2]. \tag{H.6}$$

The last expectation tends to one by dominated convergence. The middle expectation tends to $\mathbb{E}[Z\operatorname{sign}(Z-c)] = 2\phi(c)$, since $\int_c^\infty z\phi(z)\,\mathrm{d}z = \phi(c)$.

It remains to control the logarithm rather than simply assert its pointwise convergence. Write $u = (z-c)/w$, so $T_w' = 1 + (a/w)(1+u^2)^{-3/2}$. For any $q \in (1/3, 1)$, the inequality $\log(1+t) \leq t^q/q$ gives

$$0 \leq \mathbb{E}\log T_w'(Z) \leq \frac{a^q w^{1-q}}{q\sqrt{2\pi}} \int_{\mathbb{R}} (1+u^2)^{-3q/2}\,\mathrm{d}u \longrightarrow 0. \tag{H.7}$$

The integral is finite precisely because $3q > 1$. This proves the second limit. The limiting ratio is $1 + 4\phi(c)/(\alpha a)$, which diverges as $a \downarrow 0$. □

H.4 WHAT THE COUNTEREXAMPLE DOES AND DOES NOT SHOW

The false quantity in (H.4) is a purported upper bound on the *objective gap*. Its failure is not merely a disagreement between two unverified numerical solvers. The right side is obtained by an exact asymptotic calculation. Nor does the proposition claim that every Wasserstein distance estimate based on the same nodes fails on exactly this example. The actual distance to the equilibrium tends to a, consistent with the fact that concentrated derivative effects can produce a much larger objective gap than a naive quadratic-in-a estimate.

Adding more predetermined nodes does not give a uniform remedy: the transition center can be chosen away from any finite set, and the width can then be reduced. A coefficient-aware interval audit can discover the narrow transition and refine locally. Its higher cost is not an implementation accident that may be ignored in a certificate claim; it is the price of information that the fixed quadrature lacks.

I Optional probabilistic audits and adaptive validity

I.1 Frozen-model stratified integration

The deterministic theorem is the default certification result. A probabilistic alternative is possible when exact fresh sampling and valid envelopes are available. Let f be fixed after conditioning on all previous information. Partition a core into boxes $B_1, \dots, B_m$ with probabilities p_i, and suppose

$$0 \le |f(z)|^2 \le M_i \quad (z \in B_i), \qquad \int_{\mathrm{core}^c} |f|^2 \, \mathrm{d}\gamma_d \le \mathcal{T}. \tag{I.1}$$

Draw n_i fresh independent points from the conditional Gaussian law on B_i, independently across cells. Set

$$\widehat{S} = \sum_{i=1}^{m} \frac{p_i}{n_i} \sum_{j=1}^{n_i} |f(Z_{ij})|^2. \tag{I.2}$$

The weights p_i, envelopes M_i, sample counts n_i, and frozen model may all depend on the past, but not on these new sample values.

Proposition I.1 (A one-sided finite-sample norm certificate)**.** *For $0 < \delta < 1$, with conditional probability at least $1 - \delta$,*

$$\|f\|^2_{L^2(\gamma_d)} \le \widehat{S} + \sqrt{\frac{1}{2}\log(1/\delta) \sum_i \frac{p_i^2 M_i^2}{n_i}} + \mathcal{T}. \tag{I.3}$$

Proof. Condition on the model and all design choices. Each summand $p_i|f(Z_{ij})|^2/n_i$ is independent and lies in an interval of length $p_i M_i / n_i$. Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr\left(S - \widehat{S} \ge t \mid \text{past}\right) \le \exp\left[-\frac{2t^2}{\sum_i p_i^2 M_i^2 / n_i}\right], \tag{I.4}$$

where S is the exact core integral. Choose the displayed value of t and add the deterministic tail bound. □

Sound interval evaluation of $|f(Z_{ij})|^2$ can replace pointwise exact evaluation by using upper endpoints in the sample average. Interval probability bounds need corresponding conservative weighting; alternatively one can regard the exact conditional probabilities as part of the sampling oracle. A practical claim of rigorous sampling must also specify random generation and numerical distribution-function error, which are outside the abstract proposition.

I.2 Optimal stratification for the displayed bound

For a fixed total real-valued allocation $N = \sum_i n_i$, the variance proxy in (I.3) is minimized by

$$n_i = N \frac{p_i M_i}{\sum_j p_j M_j} \tag{I.5}$$

when all $p_i M_i > 0$. Cauchy–Schwarz proves

$$\left(\sum_i p_i M_i\right)^2 \le \left(\sum_i \frac{p_i^2 M_i^2}{n_i}\right)\left(\sum_i n_i\right), \tag{I.6}$$

with equality exactly at (I.5). Integer counts can be rounded upward, increasing the total budget by at most one per cell. Zero-envelope cells need no samples. This optimizes the stated worst-case concentration bound, not an unknown empirical variance.

The formula also explains a limitation: very conservative envelopes can dominate the sample complexity. Increasing neural training accuracy does not tighten those envelopes automatically. The result is a validity theorem, not an assertion that this stochastic auditor is more efficient than deterministic interval integration.

I.3 Repeated checks and data-dependent stopping

A stopping rule may examine many candidates and many audit resolutions. A confidence guarantee at one fixed check does not automatically survive selecting a successful check from many trials.

Theorem I.2 (Simultaneous validity under adaptive proposals)**.** *Index every stochastic audit, including audits of rejected candidates, by $j \geq 1$. Let $\mathcal{H}_{j-1}$ be all information before the fresh samples for check j. Suppose that conditionally on $\mathcal{H}_{j-1}$ its certificate fails with probability at most δ_j, where $\delta_j \geq 0$ is $\mathcal{H}_{j-1}$-measurable and $\sum_j \delta_j \leq \delta$ almost surely. Then all certificates are valid simultaneously with probability at least $1 - \delta$. In particular any accepted trajectory constructed from them satisfies its deterministic error account on this event.*

Proof. Let F_j be failure at check j. By the tower property, $\Pr(F_j) = \mathbb{E}[\Pr(F_j \mid \mathcal{H}_{j-1})] \leq \mathbb{E}\delta_j$. The countable union bound and monotone convergence give

$$\Pr\Big(\bigcup_{j\geq 1} F_j\Big) \leq \sum_j \mathbb{E}\delta_j = \mathbb{E}\sum_j \delta_j \leq \delta. \tag{I.7}$$

On the complement every numerical bound used in acceptance is correct. The deterministic induction proving the trajectory theorem then applies pathwise. The number and locations of accepted steps may be data-dependent; no additional union bound over their possible values is needed. □

One convenient schedule is $\delta_j = 6\delta/(\pi^2 j^2)$. It is not necessary to know the number of future audits. If a single audit also checks several norm quantities, its allocated failure budget must be split between them or a joint inequality must be used.

I.4 Why a repeatedly searched holdout is different

Consider a finite set of points already revealed to a training procedure. A smooth compact perturbation can be placed in an interval containing none of them. The resulting model may have zero perturbation at every holdout point while differing elsewhere. This is the same information mechanism as Theorem H.1, now applied to adaptive model selection. The sample is no longer independent of the chosen integrand, so the conditional Hoeffding proof above does not apply.

There are other ways to justify reuse, such as a uniform complexity bound for a predetermined function class or a suitable time-uniform concentration argument. Those are additional assumptions and are not asserted here. The simple valid protocol in this paper freezes the candidate and draws fresh samples at each check. Merely calling a dataset "validation data" does not establish that protocol.

I.5 Statistical failure and deterministic failure are separate

A stochastic audit can fail because its confidence bound is too large, because the model's true residual is too large, or because a probability-δ event occurred. A deterministic audit can also be inconclusive because its enclosure is too wide, but it has no probabilistic escape from an invalid upper bound. The algorithm must report which type of certificate it uses. An overall statement such as "error at most ε" should therefore either be deterministic under the arithmetic assumptions or explicitly state the simultaneous probability $1 - \delta$.

J The transport interpolation and its physical defect

J.1 A curve of laws with an explicit velocity

Fix a source μ and an admissible candidate T. Throughout this appendix h is fixed and positive. Write

$$\Delta = T - \mathrm{Id}, \qquad Q_\tau = \mathrm{Id} + \tau\Delta, \qquad B_\tau = DQ_\tau = I + \tau D\Delta, \qquad 0 \leq \tau \leq 1. \tag{J.1}$$

The eigenvalues of B_τ lie between $b_* := \min\{1, m\} > 0$ and $b^* := \max\{1, M\}$. Therefore Q_τ is a global diffeomorphism and a Brenier map from μ. Let $\rho_\tau \, \mathrm{d}y = (Q_\tau)_\# \mu$. In physical time $s = h\tau$, define

$$v_s(Q_{s/h}(x)) = \Delta(x)/h. \tag{J.2}$$

For every smooth compactly supported test function ζ,

$$\frac{\mathrm{d}}{\mathrm{d}s}\int \zeta(y)\rho_{s/h}(y)\,\mathrm{d}y = \frac{1}{h}\int \nabla\zeta(Q_{s/h}(x))\cdot\Delta(x)\,\mathrm{d}\mu(x)$$
$$= \int \nabla\zeta(y)\cdot v_s(y)\rho_{s/h}(y)\,\mathrm{d}y. \tag{J.3}$$

Differentiation is justified by the compact support of $\nabla\zeta$, linear growth of the maps, and the finite second moment. This proves the continuity equation $\partial_s\rho_{s/h} + \mathrm{div}(\rho_{s/h}v_s) = 0$. Moreover

$$\int |v_s|^2\rho_{s/h}\,\mathrm{d}y = \|\Delta\|^2_{L^2(\mu)}/h^2. \tag{J.4}$$

Thus this is an absolutely continuous Wasserstein curve. It is in fact the displacement interpolation from μ to $T_{\#}\mu$, but the explicit continuity equation is all that is needed for the defect estimate.

The velocity is a gradient field. Its derivative in Eulerian coordinates is $h^{-1}D\Delta\, B_\tau^{-1}$, which is symmetric because B_τ is a polynomial in the symmetric matrix $D\Delta$ at the same spatial point. Since $\mathbb{R}^d$ is simply connected, a continuously differentiable field with symmetric derivative is a gradient. This observation also puts the velocity in the usual Wasserstein tangent space, after the standard L^2 closure.

J.2 The spatial gradient of the energy along the curve

Define $c_\tau = \nabla \log\det B_\tau$. The score change of variables gives

$$\nabla\log\rho_\tau(Q_\tau(x)) = B_\tau^{-1}\big(s_\mu(x) - c_\tau(x)\big). \tag{J.5}$$

For the interaction, $(Q_\tau)_{\#}\mu = \rho_\tau\,\mathrm{d}y$ implies

$$(\nabla W * \rho_\tau)(Q_\tau(x)) = \int \nabla W(Q_\tau(x) - Q_\tau(y))\,\mathrm{d}\mu(y). \tag{J.6}$$

Consequently the pullback of the energy subgradient is

$$A_\tau(x) = \nabla V(Q_\tau(x)) + \int \nabla W(Q_\tau(x) - Q_\tau(y))\,\mathrm{d}\mu(y)$$
$$+ B_\tau^{-1}\big(s_\mu(x) - c_\tau(x)\big). \tag{J.7}$$

Under the moment and polynomial-growth assumptions, $A_\tau \in L^2(\mu)$, uniformly in τ. Its Eulerian counterpart is the usual minimal subgradient $\xi_\tau = \nabla\log\rho_\tau + \nabla V + \nabla W * \rho_\tau$. For completeness, this subgradient statement can also be obtained directly from the full support proof: use $\rho_\tau\,\mathrm{d}y$ as source, the identity as candidate, and omit the proximal penalty. The resulting inequality is the displacement subgradient inequality with curvature λ. Thus the physical-time argument uses an energy subgradient on laws, not a parameter-space derivative.

J.3 A matrix identity that avoids a commutativity mistake

Set $D = D\Delta$. Since $B_\tau = I + \tau D$,

$$\partial_\tau B_\tau^{-1} = -B_\tau^{-1}DB_\tau^{-1}, \qquad (c_\tau)_i = \tau\,\mathrm{tr}(B_\tau^{-1}\partial_i D). \tag{J.8}$$

Differentiating the second expression gives

$$(\dot c_\tau)_i = \mathrm{tr}(B_\tau^{-1}\partial_i D) - \tau\,\mathrm{tr}(B_\tau^{-1}DB_\tau^{-1}\partial_i D)$$
$$= \mathrm{tr}\left[(B_\tau^{-1} - \tau B_\tau^{-1}DB_\tau^{-1})\partial_i D\right] = \mathrm{tr}(B_\tau^{-2}\partial_i D). \tag{J.9}$$

Indeed $B_\tau^{-1} - \tau B_\tau^{-1}DB_\tau^{-1} = B_\tau^{-1}(B_\tau - \tau D)B_\tau^{-1} = B_\tau^{-2}$. This identity does not assume that $\partial_i D$ commutes with D or B_τ. Keeping it inside the trace is essential. It requires second derivatives of the map, not fourth derivatives of the map.

Proposition J.1 (Exact interpolation derivative)**.** *Under Assumptions A.1–A.3, with the moments needed for the displayed expressions, $\tau \mapsto A_\tau$ is continuously differentiable into $L^2(\mu)$ and*

$$\dot A_\tau(x) = \nabla^2 V(Q_\tau(x))\Delta(x)$$
$$+ \int \nabla^2 W(Q_\tau(x) - Q_\tau(y))(\Delta(x) - \Delta(y))\,\mathrm{d}\mu(y)$$
$$- B_\tau^{-1}DB_\tau^{-1}(s_\mu - c_\tau) - B_\tau^{-1}\dot c_\tau. \tag{J.10}$$

Proof. Apply the ordinary chain rule to the first term of (J.7). For the interaction, differentiate under the integral. A polynomial in $|x| + |y|$ bounds the differentiated integrand uniformly in τ, because Q_τ and Δ have linear growth. Jensen's inequality and the assumed source moments supply an $L^2(\mu)$ dominating function. The entropy term follows from the inverse derivative and (J.9). Uniform invertibility, bounded $D\Delta$ and $D^2\Delta$, and $s_\mu \in L^2(\mu)$ supply domination there. The same bounds and continuity of the coefficients give continuity of the derivative in L^2 by dominated convergence. □

J.4 The endpoint residual is only part of the defect

Let $e_s = v_s + \xi_{s/h}$, so the continuity equation can be written

$$\partial_s \rho_{s/h} = \operatorname{div}(\rho_{s/h}\xi_{s/h}) - \operatorname{div}(\rho_{s/h} e_s). \tag{J.11}$$

An exact gradient flow has $e_s = 0$. In source coordinates,

$$e_s(Q_{s/h}(x)) = \Delta(x)/h + A_{s/h}(x) = r_h[T](x) + A_{s/h}(x) - A_1(x). \tag{J.12}$$

Therefore valid numbers $R \geq \|r_h[T]\|_{L^2(\mu)}$ and $C \geq \sup_{0\leq\tau\leq 1} \left\|\dot{A}_\tau\right\|_{L^2(\mu)}$ imply

$$\|e_s\|_{L^2(\rho_{s/h})} = \left\|r_h[T] + A_{s/h} - A_1\right\|_{L^2(\mu)} \leq R + (1 - s/h)C. \tag{J.13}$$

The equality of the two norms is exact change of variables. The inequality uses the Bochner fundamental theorem of calculus $A_\tau - A_1 = -\int_\tau^1 \dot{A}_\sigma \, \mathrm{d}\sigma$ and the triangle inequality.

A perfect inner solve makes $R = 0$ but generally not $C = 0$. Hence an exact JKO step is not automatically an exact physical-time step. Conversely, C measures variation of the energy vector field along a specific interpolation; it is not an estimate of the distance from the candidate to the JKO optimum. These are logically independent roles.

J.5 Useful a priori envelopes

If $\left\|\nabla^2 V\right\|_{\mathrm{op}} \leq L_V$ and $\left\|\nabla^2 W\right\|_{\mathrm{op}} \leq L_W$ globally, the first two lines of (J.10) have norm at most $(L_V + 2L_W)\,\|\Delta\|_{L^2(\mu)}$. The interaction estimate follows from $\|\Delta(x) - \Delta(y)\|_{L^2(\mu\otimes\mu)} \leq 2\,\|\Delta\|_{L^2(\mu)}$. Writing $d_0 = \|D\Delta\|_{\infty,\mathrm{op}}$ and $d_1 = \left\|D^2\Delta\right\|_{\infty,\mathrm{F}}$ gives the entropy bound

$$C \leq (L_V + 2L_W)\,\|\Delta\|_{L^2(\mu)} + b_*^{-2} d_0 \left(\|s_\mu\|_{L^2(\mu)} + \sqrt{d}\, b_*^{-1} d_1\right) + \sqrt{d}\, b_*^{-3} d_1. \tag{J.14}$$

Here $|c_\tau| \leq \sqrt{d}\, b_*^{-1} d_1$ and $|\dot{c}_\tau| \leq \sqrt{d}\, b_*^{-2} d_1$. For polynomial potentials use moment-weighted bounds instead of a nonexistent global L_V. Formula (J.14) is conservative, but proves that C is finite and explains why $C = O(h)$ is plausible when the step displacement and its first two derivatives are $O(h)$. Such regularity is an additional hypothesis, not a consequence of a small parameter gradient.

K Wasserstein stability of the perturbed evolution

K.1 The precise flow hypothesis

Let σ_t be the λ-EVI gradient flow of $\mathcal{F}$ in $\mathcal{P}_2$, with the finite-energy initial law used for comparison. The EVI property means that for every finite-energy comparison law η,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\sigma_t, \eta) + \frac{\lambda}{2} W_2^2(\sigma_t, \eta) \leq \mathcal{F}(\eta) - \mathcal{F}(\sigma_t) \tag{K.1}$$

holds for almost every positive time in its standard integrated sense. For entropy, a λ-convex potential, and an even convex interaction on Euclidean space, this is the classical displacement-convex gradient-flow setting (Ambrosio et al., 2008). We use the standard EVI existence and Wasserstein distance-calculus results as foundational inputs; we do not deduce EVI existence from neural training.

The approximate curve ρ_t is locally absolutely continuous, solves the continuity equation with velocity v_t, and has a square-integrable energy subgradient ξ_t satisfying the displacement support inequality.

For our interpolations these requirements were established in Appendix J. Write $e_t = v_t + \xi_t$. Assume its norm is locally integrable in time. The curve is allowed to change its velocity at accepted step boundaries; a finite number of such boundaries on a compact interval has no effect on an almost-everywhere differential inequality.

Lemma K.1 (Residual stability)**.** *Under these hypotheses, for* $0 \le a \le t$*,*

$$W_2(\rho_t, \sigma_t) \le e^{-\lambda(t-a)} W_2(\rho_a, \sigma_a) + \int_a^t e^{-\lambda(t-s)} \, \|e_s\|_{L^2(\rho_s)} \, \mathrm{d}s. \tag{K.2}$$

Proof. We first give the calculation at times at which the exact flow has its square-integrable velocity $w_t = -\xi_t^\star$ and the Wasserstein distance derivative exists. Let π_t be an optimal coupling of ρ_t and σ_t. The distance-calculus inequality is

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\rho_t, \sigma_t) \le \int (x - y) \cdot (v_t(x) - w_t(y)) \, \mathrm{d}\pi_t(x, y). \tag{K.3}$$

It follows by transporting an optimal coupling for an infinitesimal time using the two continuity equations and expanding the squared cost; the rigorous absolutely continuous version is part of the Wasserstein differential calculus in Ambrosio et al. (2008). The subgradient inequalities at the two laws read

$$\mathcal{F}(\sigma_t) - \mathcal{F}(\rho_t) \ge \int \xi_t(x) \cdot (y - x) \, \mathrm{d}\pi_t + \tfrac{\lambda}{2} W_2^2, \tag{K.4}$$

$$\mathcal{F}(\rho_t) - \mathcal{F}(\sigma_t) \ge \int \xi_t^\star(y) \cdot (x - y) \, \mathrm{d}\pi_t + \tfrac{\lambda}{2} W_2^2. \tag{K.5}$$

Adding them proves

$$\int (x - y) \cdot (\xi_t(x) - \xi_t^\star(y)) \, \mathrm{d}\pi_t \ge \lambda W_2^2. \tag{K.6}$$

Substitute $v_t = -\xi_t + e_t$ and $w_t = -\xi_t^\star$ in (K.3), and apply Cauchy–Schwarz to the remaining error term. With $d(t) = W_2(\rho_t, \sigma_t)$ and $f(t) = \|e_t\|_{L^2(\rho_t)}$ this gives

$$\tfrac{1}{2}(d^2)' \le -\lambda d^2 + df. \tag{K.7}$$

The same inequality for an EVI flow at nonsmooth times follows by its integrated EVI and standard time regularization. Equivalently, perform the argument from a positive time where the flow has its metric velocity and then use continuity at the initial time.

To avoid dividing by d at a collision, let $d_\epsilon = (d^2 + \epsilon^2)^{1/2}$. The chain rule and $d/d_\epsilon \le 1$ imply

$$d_\epsilon' \le -\lambda d_\epsilon + f + \lambda\epsilon \tag{K.8}$$

for $\lambda \ge 0$, since $\epsilon^2/d_\epsilon \le \epsilon$. Multiply by $e^{\lambda t}$, integrate, and let $\epsilon \downarrow 0$. This proves (K.2). For $\lambda = 0$ the regularization term is absent. □

This lemma is an established residual-stability mechanism, specialized here to an auditable neural-JKO defect. Related a posteriori Wasserstein residual bounds appear in Gonnella et al. (2026). The new object supplied by the present construction is the pair of computable endpoint/interpolation bounds, not the scalar Gronwall inequality itself.

K.2 Exact integration of the local envelope

For one step with initial comparison error at most E_0, substitute (J.13) in (K.2). At the endpoint,

$$E_1 = e^{-\lambda h} E_0 + a_h R + b_h C, \tag{K.9}$$

where

$$a_h = \int_0^h e^{-\lambda(h-s)} \, \mathrm{d}s, \qquad b_h = \int_0^h e^{-\lambda(h-s)} (1 - s/h) \, \mathrm{d}s. \tag{K.10}$$

For $\lambda > 0$, the substitution $u = h - s$ gives

$$a_h = \frac{1 - e^{-\lambda h}}{\lambda}, \qquad b_h = \frac{1 - (1 + \lambda h)e^{-\lambda h}}{h\lambda^2}. \tag{K.11}$$

At $\lambda = 0$, use the integrals directly to obtain $a_h = h$ and $b_h = h/2$. Both coefficients are nonnegative. They satisfy $a_h \le h$ and $b_h \le h/2$ for $\lambda \ge 0$.

The alternative expression $h^{-1} \int_0^h u e^{-\lambda u} \, \mathrm{d}u$ for b_h is useful for checking its sign and its small-h limit. Expanding the numerator alone in machine arithmetic can suffer cancellation when λh is small. An implementation should use a validated series, an enclosing exponential implementation, or the integral representation rather than trust a difference of nearly equal floating-point values.

K.3 The error bound holds between grid points as well

Let $0 \le s \le h$. Define a_s as in (K.10), and

$$c_{s,h} = \int_0^s e^{-\lambda(s-u)}(1 - u/h) \, \mathrm{d}u. \tag{K.12}$$

Then

$$W_2(\rho_{s/h}, \sigma_s) \le e^{-\lambda s} E_0 + a_s R + c_{s,h} C. \tag{K.13}$$

For $\lambda > 0$, a numerically useful identity is

$$c_{s,h} = (1 - s/h)a_s + \frac{1 - (1 + \lambda s)e^{-\lambda s}}{h\lambda^2}; \tag{K.14}$$

for $\lambda = 0$ it is $s - s^2/(2h)$. Every term in its defining integral is nonnegative. At $s = h$, $c_{h,h} = b_h$.

This is a certificate for the specified interpolation, not a claim that every path joining the two endpoint laws has the same error. The path and its physical duration are part of the algorithm's output. Merely relabeling the endpoints by a different time grid changes the velocity, residual, and error accounting.

K.4 Composition of local ledgers

Let $t_{k+1} = t_k + h_k$, and let valid R_k, C_k refer to the actual accepted source and map at step k. Define

$$E_N = e^{-\lambda t_N} E_0 + \sum_{k=0}^{N-1} e^{-\lambda(t_N - t_{k+1})}(a_{h_k} R_k + b_{h_k} C_k). \tag{K.15}$$

Induction using the local bound proves $W_2(\rho_{t_N}, \sigma_{t_N}) \le E_N$. Equivalently the ledger can be updated recursively, requiring constant storage apart from the audit records. Rejected trial steps do not advance time and do not enter (K.15). They still count toward computational work and, for probabilistic audits, toward the failure-probability allocation. A changed source law is never silently compared with the exact JKO step from a previous source. This is why the continuous-time ledger remains valid without a general Wasserstein resolvent-contraction theorem.

L Stopping rules, step-size control, and quantitative rates

L.1 A finite-horizon acceptance theorem

Choose a horizon $T_{\mathrm{fin}} > 0$, a total permissible error $\varepsilon > 0$, a certified initial error $E_0 < \varepsilon$, and a split $\theta \in (0, 1)$. Put

$$\zeta = (\varepsilon - E_0)/T_{\mathrm{fin}}. \tag{L.1}$$

At a trial step of length h, accept only if independently valid bounds satisfy

$$a_h R \le \theta \zeta h, \qquad b_h C \le (1 - \theta)\zeta h. \tag{L.2}$$

The two tests have different remediation paths. Failure of the first permits more training, feature enrichment, or a tighter audit. Failure of the second calls for a smaller time step unless the clock audit itself is unnecessarily loose. Neither a failed solver exit nor an exhausted audit budget is acceptance.

Theorem L.1 (Sound finite-horizon endpoint control)**.** *Suppose the accepted steps reach T_{fin} and each satisfies* (L.2)*. Then*

$$W_2(\rho_{t_k}, \sigma_{t_k}) \le E_k \le E_0 + \zeta t_k \le \varepsilon. \tag{L.3}$$

Every interior point has the independently valid bound (K.13)*. To enforce the same ε-tube at all interior points, it suffices to impose the stronger, contraction-free tests*

$$R \le \theta\zeta, \qquad C/2 \le (1-\theta)\zeta. \tag{L.4}$$

Proof. Since $e^{-\lambda h} \le 1$, the endpoint recursion gives $E_{k+1} \le E_k + \zeta h_k$. Induction proves the first conclusion. Under (L.4), at a partial step $s \le h$, the nonnegative integral of the defect envelope is at most its integral over the whole step:

$$a_s R + c_{s,h} C \le sR + (s - s^2/(2h))C \le hR + hC/2 \le \zeta h. \tag{L.5}$$

Thus the error at every point of that step is at most $E_0 + \zeta(t_k + h_k) \le \varepsilon$. This proves the uniform-tube conclusion. The stronger tests imply the weighted endpoint tests because $a_h \le h$ and $b_h \le h/2$. □

A weighted endpoint budget alone need not bound an interior maximum by the same number: a decreasing forcing term can make the scalar error envelope peak before the endpoint. The distinction is retained in the theorem. One may use the sharper weighted tests and explicitly audit the maximum of (K.13), or use (L.4) and avoid that extra calculation. The mathematical path certificate is available in both cases.

L.2 A CONTRACTION-AWARE INVARIANT TUBE

When $\lambda > 0$ and the horizon is large, a budget proportional to $1 - e^{-\lambda h}$ avoids accumulating a fixed amount of permitted error forever.

Proposition L.2 (Endpoint and continuous invariant tubes)**.** *Suppose $E_0 \le \varepsilon$ and every accepted step satisfies*

$$a_h R + b_h C \le \varepsilon(1 - e^{-\lambda h}). \tag{L.6}$$

Then all endpoint errors are at most ε. If in addition $R + C \le \lambda\varepsilon$ on each step, the whole interpolation remains in that tube.

Proof. The endpoint recursion gives $E_{k+1} \le e^{-\lambda h}\varepsilon + \varepsilon(1 - e^{-\lambda h}) = \varepsilon$. For the continuous assertion, the defect norm is at most $R + C \le \lambda\varepsilon$. Lemma K.1 then gives $E(s) \le e^{-\lambda s}\varepsilon + \lambda\varepsilon a_s = \varepsilon$. □

A convenient split of (L.6) is $R \le \theta\lambda\varepsilon$ and $b_h C \le (1-\theta)\varepsilon(1 - e^{-\lambda h})$. The first equivalence follows from $a_h = (1 - e^{-\lambda h})/\lambda$. These inequalities distinguish the terminal-error objective from minimizing every local objective to the same arbitrary tolerance.

L.3 FIRST-ORDER PHYSICAL-TIME CONVERGENCE WITH VERIFIED INNER WORK

Theorem L.3 (A rate from residual and clock scalings)**.** *Let the steps reach T_{fin} and let $h_{\max} = \max_k h_k$. If*

$$R_k \le c_R h_k, \qquad C_k \le c_C h_k \tag{L.7}$$

for constants independent of the mesh, then

$$W_2(\rho_{T_{\mathrm{fin}}}, \sigma_{T_{\mathrm{fin}}}) \le e^{-\lambda T_{\mathrm{fin}}} E_0 + (c_R + c_C/2) T_{\mathrm{fin}} h_{\max}. \tag{L.8}$$

For every interior point, the same right side with E_0 in place of $e^{-\lambda T_{\mathrm{fin}}} E_0$ is a uniform upper bound over the horizon.

Proof. Each local contribution is at most $a_{h_k} R_k + b_{h_k} C_k \le (c_R + c_C/2) h_k^2$. In (K.15), omit exponential factors only from the nonnegative forcing contributions and use $\sum h_k^2 \le h_{\max} \sum h_k = T_{\mathrm{fin}} h_{\max}$. For interior times, bound the partial-step integral by the corresponding whole-step integral as in the preceding theorem; the same sum is an upper bound. □

For the objective gap, the local residual certificate gives

$$\mathcal{J}_{h_k,\mu_k}(\mu_{k+1}) - \mathcal{J}^\star_{h_k,\mu_k} \le \frac{c_R^2 h_k^3}{2(1+\lambda h_k)}. \tag{L.9}$$

Thus an $O(h^3)$ verified gap is compatible with first-order physical-time accuracy, but the converse implication does not follow from (L.9): a small gap need not give a small full residual, and neither statement controls the clock term without a separate bound. The theorem assumes (L.7), not merely a training-loss tolerance labeled "cubic."

L.4 Why tightening only the inner tolerance is insufficient

Consider a one-dimensional Gaussian translation in the Ornstein–Uhlenbeck flow with $V(x) = \lambda x^2/2$, $\lambda > 0$, and stationary variance $1/\lambda$. Its mean solves $m' = -\lambda m$. An exact JKO step has mean $m_1 = m_0/(1+\lambda h)$, whereas the exact flow has mean $m_0 e^{-\lambda h}$. Consequently

$$W_2(\mu_1, \sigma_h) = |m_0| \left|(1+\lambda h)^{-1} - e^{-\lambda h}\right| > 0 \tag{L.10}$$

when $m_0 \neq 0$ and $h > 0$. The inequality is strict since $\log(1+x) < x$ for $x > 0$. The inner residual and unrestricted objective gap are both exactly zero. No additional inner training changes (L.10).

For this example the displacement is constant and the clock derivative is $\lambda\Delta$, so $C = \lambda|\Delta|$. It is precisely the omitted clock quantity. This analytic example isolates a physical-time error without representation error, numerical quadrature error, or imperfect optimization.

L.5 Interpreting a rejection

A failed upper-bound test has three different possible causes. The candidate may truly be inaccurate; the architecture may not approximate a satisfactory candidate; or the enclosure may be too loose. The upper bound alone does not identify which cause occurred. Certified descent directions from Appendix F and lower enclosures of residual integrals from Appendix G can sometimes distinguish them. A lower residual enclosure proves only that the residual is large, not that the objective gap is correspondingly large.

The algorithm therefore returns one of two mathematically meaningful outcomes: an accepted law with its certificate, or a failure to certify within the supplied resources. It does not convert the latter into a successful output by dropping a difficult audit term. The next appendix treats existence of good neural approximations, and the following one states the extra scheduling assumptions required to obtain finite termination.

M Graph approximation and residual consistency

M.1 The topology needed for a certificate

Density of neural outputs in L^2 does not imply density of their entropy residuals. The score term contains $(DT)^{-1}$ and D^2T, so a sequence that converges only in value can develop narrow derivative spikes. The obstruction in Appendix H is one concrete manifestation. We therefore specify a stronger approximation topology.

Definition M.1 (A curvature-controlled local graph sequence)**.** Fix a C^2 map T_* with $m_* I \preceq DT_* \preceq M_* I$. A sequence T_n is a local graph approximation if there is a compact set K_0 such that $T_n = T_*$ outside K_0, all $DT_n \succeq mI$ for one $m > 0$, and

$$\|T_n - T_*\|_\infty + orm DT_n - DT_*{}_{\infty,\mathrm{op}} + \left\|D^2 T_n - D^2 T_*\right\|_{\infty,\mathrm{F}} \longrightarrow 0. \tag{M.1}$$

A weighted graph approximation may also perturb a common affine tail, provided the three differences are dominated by fixed polynomial envelopes and converge pointwise with their derivatives, and the same uniform lower curvature bound holds.

The weighted version is useful for rationalizing the coefficients of a quadratic potential. In either version the assumptions ensure a common integrable bound on the residuals under the source moment and score conditions. No approximation of the inverse map is needed.

Proposition M.2 (Residual consistency)**.** *For a fixed source and step, a sequence as in Definition M.1, with common polynomial envelopes, satisfies*

$$\|r_h[T_n] - r_h[T_*]\|_{L^2(\mu)} \longrightarrow 0. \tag{M.2}$$

The interpolation clock functions also converge uniformly in τ in $L^2(\mu)$:

$$\sup_{\tau\in[0,1]} \left\| \dot{A}_\tau[T_n] - \dot{A}_\tau[T_*] \right\|_{L^2(\mu)} \longrightarrow 0. \tag{M.3}$$

Here continuity of $\nabla^2 V$ and $\nabla^2 W$, polynomial derivative bounds, and the moments appearing in (J.10) *are assumed.*

Proof. Uniform positive curvature makes matrix inversion uniformly continuous on the relevant compact matrix sets. The formula $\partial_i \log\det DT = \mathrm{tr}((DT)^{-1}\partial_i DT)$ therefore converges pointwise with a common bound. The part multiplying s_μ is dominated in L^2 by a constant times $|s_\mu|$. All remaining terms in the residual have a polynomial envelope in $|x|$, or, for the interaction, in $|x| + |y|$. Dominated convergence and Jensen's inequality prove (M.2).

For the clock, $B_\tau = I + \tau(DT - I)$ is uniformly bounded below by $\min\{1, m\}I$. The inverse and all matrix products in (J.10) converge uniformly in τ at each spatial point. On each compact spatial set the convergence of the potential and interaction Hessians is uniform by uniform continuity. Taking the supremum in τ of the pointwise difference before applying dominated convergence gives an L^2 bound that tends to zero. For the interaction the same argument is applied on the product source space and then Jensen's inequality is used. This proves (M.3). □

In particular, if $r_h[T_*] = 0$, the true residual norms of the approximating neural maps tend to zero. The conclusion is conditional on the existence of this regular map T_*. The unrestricted minimizer from Appendix B was not assumed to have this regularity for the validity theorem.

M.2 A QUANTITATIVE RESIDUAL PERTURBATION BOUND

For $W = 0$, suppose T, S agree outside a compact set, have eigenvalues of their derivatives at least m, and have second derivatives bounded by M_2. Let L_V bound $\left\|\nabla^2 V\right\|_{\mathrm{op}}$ on the compact region swept out between $T(x)$ and $S(x)$ where they differ. Put $\delta_i = \left\|D^i(T - S)\right\|_\infty$, using operator norm for $i = 1$ and Frobenius norm for $i = 2$. Then

$$\begin{aligned}\|r_h[T] - r_h[S]\|_{L^2(\mu)} \le{}& (L_V + h^{-1})\delta_0 \\ &+ \left(m^{-2}\|s_\mu\|_{L^2(\mu)} + 2\sqrt{d}\, m^{-3} M_2\right)\delta_1 + \sqrt{d}\, m^{-2}\delta_2.\end{aligned} \tag{M.4}$$

To check the score term, use $K^{-1} - L^{-1} = K^{-1}(L - K)L^{-1}$, so its operator norm is at most $m^{-2}\delta_1$. For $c_T = \nabla \log\det DT$,

$$|c_T - c_S| \le \sqrt{d}\left(m^{-1}\delta_2 + m^{-2}M_2\delta_1\right), \qquad |c_S| \le \sqrt{d}\, m^{-1} M_2. \tag{M.5}$$

Split $K^{-1}c_T - L^{-1}c_S$ into two terms and obtain (M.4). If $\nabla^2 W$ is globally bounded by L_W, add $2L_W\delta_0$. For quartic interactions a moment-weighted bound is required instead; a global Hessian bound is not available.

This estimate makes the graph requirement concrete. Reducing the value error δ_0 while leaving the derivative errors fixed need not reduce the certified residual. Conversely, a controlled C^2 approximation does reduce it, even if the neural parameter gradient is not used anywhere in the argument.

M.3 A CONSTRUCTIVE DERIVATIVE APPROXIMATION THEOREM

Theorem M.3 (Quadratic-tail potentials admit neural graph approximation)**.** *Let*

$$u_*(x) = q(x) + k(x), \qquad q(x) = \tfrac{1}{2}x^T Bx + b\cdot x + c, \tag{M.6}$$

where $k \in C_c^4(\mathbb{R}^d)$ and $D^2 u_* \succeq 2mI$ globally for some $m > 0$. There are potentials of the finite bump form (E.7) *such that*

$$\|u_n - u_*\|_{C^3(\mathbb{R}^d)} \longrightarrow 0, \qquad D^2 u_n \succeq mI \tag{M.7}$$

when the same quadratic q is retained. Their nonquadratic parts have supports in one common compact set. Thus $T_n = \nabla u_n$ is a curvature-controlled graph approximation to $T_* = \nabla u_*$. Rational approximation of q gives the weighted version of the same conclusion.

Proof. Let $\beta_{d,\epsilon}(x) = \epsilon^{-d}\beta_d(x/\epsilon)$ with the normalized bump from Appendix E. Convolution with this approximate identity gives $k_\epsilon = k * \beta_{d,\epsilon}$. For every multi-index $|\alpha| \le 3$,

$$D^\alpha k_\epsilon(x) - D^\alpha k(x) = \int \beta_d(z)\big[D^\alpha k(x - \epsilon z) - D^\alpha k(x)\big]\,\mathrm{d}z. \tag{M.8}$$

The derivatives of k are uniformly continuous and compactly supported, so the right side tends uniformly to zero as $\epsilon \downarrow 0$. All supports lie in a fixed compact enlargement of $\operatorname{supp} k$ for $\epsilon \le 1$.

Fix $\epsilon > 0$. Express the convolution as

$$k_\epsilon(x) = \int_K k(y)\epsilon^{-d}\beta_d((x-y)/\epsilon)\,\mathrm{d}y, \tag{M.9}$$

where K is a fixed box containing $\operatorname{supp} k$. Approximate this integral by Riemann sums over a partition of K with cell volumes v_j and representative points y_j:

$$k_{\epsilon,N}(x) = \sum_j v_j k(y_j)\epsilon^{-d}\beta_d((x-y_j)/\epsilon). \tag{M.10}$$

For derivatives up to order three in x, the integrands are jointly uniformly continuous on a common compact set in (x, y). Outside that set they vanish. Therefore the Riemann sums converge uniformly in C^3 as the mesh tends to zero. A diagonal sequence in ϵ and mesh size approximates k in C^3.

Choose the errors so that $\big\|D^2(k_n - k)\big\|_{\mathrm{op}} \le m$. Weyl's elementary eigenvalue inequality then gives $D^2(q + k_n) \succeq mI$. For each finite sum, perturbing its coefficients, widths, and centers slightly perturbs all derivatives through order three uniformly on a common compact set. The parameters may therefore be chosen rational while preserving any smaller strict error and curvature margins. If the coefficients of q are also rationalized, the position error has at most affine growth, the first derivative error is constant, and higher derivative errors are unchanged. This is the weighted graph variant. □

The theorem is an approximation result for potentials with a specified regularity and tail form. It does not assert that every exact JKO minimizer has a quadratic tail, that positive ridge networks are C^3-dense in all convex potentials, or that gradient descent discovers these Riemann-sum coefficients.

M.4 From approximation to auditable candidates

Each finite sum is exactly specified by finitely many rational coefficients and the fixed computable activation. Its tail is affine, all derivative envelopes are finite, and strict positive curvature can be verified by finite subdivision on the compact support. Outside that support the Hessian is the constant B. Consequently the candidates belong to the auditable class used by the integration theorem, rather than merely to an abstract closure of smooth functions.

If the source is an auditable generator and the energy has the required computable interval extensions, the true residual norm of each candidate can be enclosed to arbitrary accuracy. Combining this fact with Proposition M.2 proves existence of candidates passing any positive residual threshold whenever the exact regular minimizer is in the stated graph closure. How to arrange the search so that this existential approximation statement becomes finite acceptance is a separate scheduling question, addressed next.

N FINITE CERTIFICATION, FAIR SEARCH, AND TIME-STEP NONACCUMULATION

N.1 THREE DIFFERENT FINITE-TERMINATION CLAIMS

There are three algorithms that could fail to terminate: training a neural proposal, integrating a fixed proposal's residual tightly enough, and advancing enough accepted time steps to reach a horizon. We separate them. A global validity theorem gives no automatic bound on any of these runtimes.

For a fixed admissible map, the integration theorem gives a decreasing sequence of upper enclosures after taking running minima. If the true norm is strictly below the requested threshold, some finite refinement accepts it. If the norm is exactly equal to that threshold, finite acceptance need not occur. If it is larger, an upper-bound-only routine might refine forever unless a resource rule or lower enclosure rejects it. Fair interleaving prevents such a trial from blocking all future candidates.

Definition N.1 (Fair certificate search)**.** Enumerate a countable set of finitely specified candidate potentials and audit tasks. At stage n, execute at least one additional finite computation step for each of the first n tasks that has not finished. A task consists of checking curvature, integrability, and successively refining the required norm enclosures for one candidate. Invalid or definitively failed tasks may be removed. A certificate is returned only after all its finite checks have passed.

This is a dovetailing schedule. It is a proof device, not an efficient recommended optimizer. Neural training may be used to prioritize useful proposals, provided a fair fallback is retained when finite-termination guarantees are desired.

Theorem N.2 (Finite certification of a fixed JKO step)**.** *Fix a source, an energy, $h > 0$, and $\eta > 0$. Suppose the auditable candidate family contains a sequence T_n with common positive curvature margins and*

$$\|r_h[T_n]\|_{L^2(\mu)} \longrightarrow 0. \tag{N.1}$$

Suppose the source and energy have the computable envelopes required by the audit theorem. A fair search eventually returns an admissible candidate with a verified one-step radius at most η.

Proof. The requested residual threshold is $b = \alpha_h \eta > 0$. Choose one candidate T_{n_*} with true residual norm below $b/2$. Its curvature is strictly positive and its required integrals are finite. By the convergent audit theorem, its upper residual enclosure eventually becomes smaller than b. All these checks require only finitely many operations for that particular candidate and accuracy level. Fairness ensures that those operations eventually occur, even if some earlier candidate is invalid, has a residual above the threshold, or lies exactly on an undecidable equality boundary. At that finite stage, the global certificate theorem gives the desired radius. □

The theorem does not require a computable rate for (N.1). Such a rate would be needed for a work bound. A graph approximation to a zero-residual map supplies the hypothesis by Proposition M.2. Nothing in this proof is a convergence theorem for an arbitrary nonconvex training algorithm.

N.2 SIMULTANEOUS OPTIMIZATION AND CLOCK ACCEPTANCE

For a fixed h, both tests can be certified by the same fair-search idea if a candidate exists with strict slack in both. In particular, let an exact regular map T_h^* satisfy

$$r_h[T_h^*] = 0, \qquad C_h^* := \sup_\tau \left\| \dot{A}_\tau[T_h^*] \right\|_{L^2(\mu)} < c_h. \tag{N.2}$$

If it is approximable in the graph topology and the clock threshold is $c_h > 0$, then (M.3) implies that sufficiently accurate neural approximants have clock norm strictly below c_h. Their residual norms are also below every positive endpoint threshold. The two convergent audits then finish after finite work. One must not infer this simultaneous property from L^2 approximation of T_h^* alone.

N.3 A SUFFICIENT HYPOTHESIS EXCLUDING TIME-STEP ACCUMULATION

Let $\mathcal{U}$ be a set of possible accepted source laws. The following assumption is deliberately uniform; merely knowing that each source separately admits some small step is not enough.

Assumption N.3 (Uniform small-step graph regularity)**.** There exist $h_0 > 0$ and $K \geq 0$ such that, for every $\mu \in \mathcal{U}$ and $0 < h \leq h_0$, the exact regular JKO map $T^*_{h,\mu}$ admits auditable graph approximation, has zero full residual, and satisfies

$$C^*_{h,\mu} \leq Kh. \tag{N.3}$$

Every accepted source stays in $\mathcal{U}$. The source and energy representations remain auditable there. All proposed steps used for the theorem have strictly positive optimization budgets.

Membership in $\mathcal{U}$ and its invariance are mathematical assumptions to be proved for a particular application. Finite-step composition alone does not establish their uniform version.

Theorem N.4 (Finite advancement under uniform regularity)**.** *Assume N.3, choose $\zeta > 0$ and $\theta \in (0, 1)$, and use either the weighted endpoint tests or the stronger tests* (L.4)*. Choose $h_* > 0$ such that*

$$h_* \leq h_0, \qquad Kh_* < (1 - \theta)\zeta. \tag{N.4}$$

The second condition is automatic for any finite h_* when $K = 0$. Use trial lengths in $[h_*/2, h_*]$, except for a final remaining interval shorter than $h_*/2$, and use fair candidate/audit search at each such length. Every trial length eventually yields an accepted map. The horizon is reached in at most

$$\left\lceil \frac{2T_{\mathrm{fin}}}{h_*} \right\rceil + 1 \tag{N.5}$$

accepted steps, each with finite certification work.

Proof. For an exact map at any chosen $h \leq h_*$, the optimization residual is zero, strictly below the positive threshold. Its clock satisfies $C^*_h/2 \leq Kh/2 < (1 - \theta)\zeta/2$. Thus there is strict slack even for the stronger clock test $C/2 \leq (1 - \theta)\zeta$. Graph approximation and clock consistency preserve this slack for a sufficiently accurate candidate. Fair search and convergent audits then accept after finite work. The argument applies anew at every accepted source by the uniform assumption. Every nonfinal step advances by at least $h_*/2$, proving the count. The final shorter step has the same strict-slack argument. A finite sum of finite runtimes is finite, although no uniform numerical bound on that runtime has been supplied. □

The theorem is an existence guarantee for a safeguarded scheduling rule. It is not a claim that an arbitrary adaptive controller will avoid endlessly halving its step. It also does not identify an efficient K for general nonlinear equations. That regularity problem is separate from checking a proposed finite certificate.

N.4 Why nonaccumulation needs an explicit condition

Positive step sizes alone do not prevent $\sum_k h_k < \infty$. For example $h_k = 2^{-k}$ is positive but never reaches a horizon larger than its sum. Even when every trial can be certified, a scheduler could deliberately choose such a sequence. A theorem promising finite advancement must constrain either the scheduler, the step-size lower bound, or both.

Uniform $O(h)$ control of the clock is also not automatic from a zero inner residual. It is a regularity statement about how the exact resolvent approaches the identity in a topology strong enough to control the entropy score. In the Gaussian model below it can be verified explicitly on compact sets of positive definite covariances and bounded means. For rough initial data or collapsing curvature bounds it may fail to be uniform. The validity of an individual certificate survives such failure; the unconditional runtime guarantee does not.

O An independent primal–dual certificate for non-Brenier outputs

O.1 Scope and weak duality

This appendix assumes $W = 0$. It gives a second route to an unrestricted objective-gap bound. It can audit an arbitrary smooth diffeomorphism, not just a Brenier map, but requires a valid dual lower bound and a primal upper evaluation. It is not used to justify the full residual for non-Brenier maps.

Set $c_h(x,y) = |x-y|^2/(2h)$ and, for a function f on the target space, define

$$f^c(x) = \inf_y \{c_h(x,y) - f(y)\}. \tag{O.1}$$

Assume all displayed expectations are well defined and $0 < Z_f := \int e^{-V(y)-f(y)}\,\mathrm{d}y < \infty$. By definition, $f(y) + f^c(x) \le c_h(x,y)$. Integrating over any coupling of μ and ν and taking the optimal cost gives

$$\frac{W_2^2(\mu,\nu)}{2h} \ge \int f\,\mathrm{d}\nu + \int f^c\,\mathrm{d}\mu. \tag{O.2}$$

For any absolutely continuous probability law, the Gibbs variational inequality is

$$\mathrm{Ent}(\nu) + \int (V+f)\,\mathrm{d}\nu = \mathrm{KL}\left(\nu \middle\| Z_f^{-1} e^{-V-f}\,\mathrm{d}y\right) - \log Z_f \ge -\log Z_f. \tag{O.3}$$

Therefore

$$D_h(f) := \int f^c\,\mathrm{d}\mu - \log Z_f \le \inf_\nu \mathcal{J}_{h,\mu}(\nu). \tag{O.4}$$

Only weak duality was used. Attainment of a dual maximizer, a minimax interchange, or convergence of dual training is unnecessary for validity.

For a global C^1 diffeomorphism H with positive determinant and suitable integrability, the change-of-variables identity and the coupling $(\mathrm{Id}, H)_{\#}\mu$ give

$$\mathcal{J}_{h,\mu}(H_{\#}\mu) \le P_h^{\mathrm{coup}}(H) := \mathrm{Ent}(\mu) + \mathbb{E}_\mu\left[V(H) - \log\det DH + \frac{|H-x|^2}{2h}\right]. \tag{O.5}$$

The transport inequality may be strict. This is harmless for an upper certificate, but it must not be labeled equality for a nonoptimal map.

Proposition O.1 (Audited weak-duality gap)**.** *If $\overline{P} \ge P_h^{\mathrm{coup}}(H)$ and $\underline{D} \le D_h(f)$ are valid evaluations, then*

$$0 \le \mathcal{J}_{h,\mu}(H_{\#}\mu) - \mathcal{J}^\star_{h,\mu} \le \overline{P} - \underline{D}. \tag{O.6}$$

Under the energy assumptions, the unrestricted minimizer satisfies

$$W_2(H_{\#}\mu, \nu_h^*) \le \sqrt{\frac{2(\overline{P} - \underline{D})}{\alpha_h}}. \tag{O.7}$$

Proof. Combine (O.4) and (O.5). For the distance bound use map-space quadratic growth from Appendix B, now applied to the Brenier map from μ to the output law $H_{\#}\mu$, not to the possibly non-Brenier map H itself. This gives the same measure-space lower growth bound and proves the claim. □

O.2 A BOUNDED-CURVATURE NEURAL WITNESS

Choose a strongly convex C^2 function ψ satisfying

$$\sigma I \preceq \nabla^2\psi \preceq B, \qquad \sigma > 0, \tag{O.8}$$

and define $f(y) = |y|^2/(2h) - \psi(y)$. Expanding the square in (O.1) yields

$$f^c(x) = \frac{|x|^2}{2h} - \psi^*(x/h), \qquad \psi^*(p) = \sup_y \{p\cdot y - \psi(y)\}. \tag{O.9}$$

A positive-weight softplus potential of the form (E.1) supplies explicit lower and upper curvature bounds. The witness architecture is a sufficient class, not a claim that all dual optimizers have that form.

Lemma O.2 (A conjugate upper bound from an approximate solve)**.** *For every trial point z and $p \in \mathbb{R}^d$,*

$$\psi^*(p) \le p\cdot z - \psi(z) + \frac{|\nabla\psi(z) - p|^2}{2\sigma}. \tag{O.10}$$

Proof. Strong convexity gives $\psi(y) \geq \psi(z) + \nabla\psi(z) \cdot (y - z) + \sigma|y - z|^2/2$. Subtract from $p \cdot y$ and maximize the resulting quadratic in $y - z$. The maximum is the right side of (O.10). □

The direction matters: the conjugate needs an *upper* bound so that f^c receives a lower bound. Simply inserting a nonoptimal point in the supremum gives a lower conjugate value and would reverse the required guarantee. An approximate Newton or neural conjugate solver may propose $z(x)$, but the squared residual correction must be retained.

O.3 A LOG-PARTITION LOWER BOUND WITHOUT AN EXACT MODE

Let $U = V + f$ and suppose

$$H_0 := \lambda I + h^{-1} I - B \succ 0. \tag{O.11}$$

Then $\nabla^2 U \succeq H_0$. For any point y_0, set $g_0 = \nabla U(y_0)$. Completing the square in the strong-convexity support of U gives

$$U(y) \geq U(y_0) + g_0 \cdot (y - y_0) + \tfrac{1}{2}(y - y_0)^T H_0 (y - y_0), \tag{O.12}$$

$$Z_f \leq e^{-U(y_0) + g_0^T H_0^{-1} g_0/2} \, (2\pi)^{d/2} (\det H_0)^{-1/2}. \tag{O.13}$$

Consequently

$$-\log Z_f \geq U(y_0) - \tfrac{1}{2} g_0^T H_0^{-1} g_0 - \tfrac{d}{2}\log(2\pi) + \tfrac{1}{2}\log\det H_0. \tag{O.14}$$

The gradient correction is zero at the exact minimizer of U, but it is necessary at an approximate one. Replacing y_0 by an approximate numerical mode and dropping that correction is not a valid lower bound in general.

Combining (O.10) and (O.14) gives a completely explicit lower expression for (O.4). Remaining source expectations can be evaluated by the global audit. If (O.11) fails, this Gaussian comparison is unavailable; it does not follow that the dual expression is invalid or that Z_f is infinite. A different verified partition bound may be used.

O.4 HYBRID CERTIFICATES AND A VALIDITY–SHARPNESS DISTINCTION

For a Brenier output T, both methods can be applied. If $g_R = R^2/(2\alpha_h)$ and $g_D = \overline{P} - \underline{D}$ are valid, then $\min\{g_R, g_D\}$ is a valid gap upper bound. The distance can be bounded by the smaller of R/α_h and $\sqrt{2g_D/\alpha_h}$. A negative difference between correctly rounded primal and dual bounds cannot occur; a negative computed difference is a consistency failure to investigate, not a quantity to clip silently to zero.

A poor dual witness can leave g_D positive even at an exact primal minimizer. A conservative global curvature lower bound for U can likewise make (O.14) loose. Thus this appendix proves validity, but does not claim that its particular witness class and Gaussian partition bound always tighten to zero. The residual audit has a different completeness statement: for a fixed regular map it converges to its actual residual norm. These two statements should not be conflated.

The primal–dual principle is classical and related to established Wasserstein primal–dual solvers (Carrillo et al., 2022). Its purpose here is to provide an alternative audit contract for outputs outside the relative-Brenier architecture. Extending this specific Gibbs formula to a general interaction energy would require an additional valid variational representation; no such extension is asserted by simply reusing (O.4).

P AN EXACT NONCOMMUTING GAUSSIAN INTERACTION MODEL

P.1 THE MODEL AND UNRESTRICTED OPTIMALITY

Let $K = K^T \succ 0$, $L = L^T \succeq 0$, and

$$V(y) = \tfrac{1}{2} y^T K y, \qquad W(y) = \tfrac{1}{2} y^T L y, \qquad \mu = N(m_0, \Sigma_0), \quad \Sigma_0 \succ 0. \tag{P.1}$$

No common eigenbasis is assumed. Set $M = K + L$, $a = h^{-1}$, $B = M + aI$, and $C_0 = \Sigma_0^{-1}$. Define the positive definite matrix

$$X = \frac{aI + (a^2 I + 4B^{1/2} C_0 B^{1/2})^{1/2}}{2}, \qquad A = B^{-1/2} X B^{-1/2}, \tag{P.2}$$

and $m = (I + hK)^{-1} m_0$. All square roots are the symmetric positive definite square roots. The candidate is

$$T(x) = m + A(x - m_0). \tag{P.3}$$

It is a globally uniformly convex gradient map. Its output law is $N(m, A\Sigma_0 A)$.

Theorem P.1 (Exact Gaussian JKO map without commutativity)**.** *The map* (P.3) *has identically zero full residual and is therefore the unique unrestricted JKO minimizer for* (P.1)*.*

Proof. Write $z = x - m_0$. The source score is $s_\mu(x) = -C_0 z$. Since $DT = A$ is constant, its log-determinant gradient vanishes. The target mean is m, so the interaction force is $I_\mu[T](x) = L(T(x) - m) = LAz$. The full residual is consequently

$$r_h[T](x) = Km + a(m - m_0) + [BA - aI - A^{-1} C_0] z. \tag{P.4}$$

The constant term is zero by the definition of m. For the matrix term, X is a function of the symmetric matrix $B^{1/2} C_0 B^{1/2}$ and satisfies

$$X^2 - aX = B^{1/2} C_0 B^{1/2}. \tag{P.5}$$

Conjugating by $B^{-1/2}$ gives

$$ABA - aA = C_0. \tag{P.6}$$

Multiplying on the left by A^{-1} yields $BA - aI - A^{-1} C_0 = 0$. Thus (P.4) vanishes. The global full-space certificate gives gap zero against all finite-energy laws, proving unrestricted optimality. No assumption of Gaussian optimality was used in that argument. □

This also proves that solving the finite matrix equation did not merely optimize within Gaussian laws. The full-space residual is the bridge from the explicit matrix calculation to the unrestricted problem.

P.2 Why the matrix order cannot be changed

In general $A\Sigma_0 \neq \Sigma_0 A$, and BA need not be symmetric. The coefficient in (P.4) is $BA - aI - A^{-1} C_0$, not $AB - aI - C_0 A^{-1}$ with factors freely reordered. The Riccati equation restores symmetry by multiplying on the left with A. Formula (P.2) is obtained by a congruence transformation, not simultaneous diagonalization.

Uniqueness of the positive definite solution of (P.6) can also be seen directly. Any such solution gives a positive definite X satisfying $X^2 - aX = D$ with $D = B^{1/2} C_0 B^{1/2} \succ 0$. Hence X commutes with D, because D is a polynomial in X. In an eigenbasis of X, each positive eigenvalue solves $x^2 - ax = d > 0$, whose only positive root is $(a + \sqrt{a^2 + 4d})/2$. Functional calculus gives (P.2).

P.3 Closed residual norms for nonoptimal Gaussian proposals

For any affine Brenier map $T(x) = m + A(x - m_0)$ with $A \succ 0$, define

$$b = Km + a(m - m_0), \qquad H = BA - aI - A^{-1} C_0. \tag{P.7}$$

Gaussian centering gives the exact identity

$$\|r_h[T]\|^2_{L^2(\mu)} = |b|^2 + \operatorname{tr}(H \Sigma_0 H^T). \tag{P.8}$$

This permits analytic calibration of both the norm and the gap bound. The true objective can also be evaluated explicitly:

$$\begin{aligned} P_h(T) = & -\tfrac{1}{2} \log\left((2\pi e)^d \det(A\Sigma_0 A)\right) + \tfrac{1}{2} m^T K m + \tfrac{1}{2} \operatorname{tr}(MA\Sigma_0 A) \\ & + \frac{1}{2h}\left[|m - m_0|^2 + \operatorname{tr}((A - I)\Sigma_0 (A - I))\right]. \end{aligned} \tag{P.9}$$

The interaction contributes $\mathrm{tr}(LA\Sigma_0 A)/2$ because the covariance of two independent centered target draws is $2A\Sigma_0 A$ and the energy already has its factor $1/2$.

For isotropic one-dimensional $K = \lambda > 0$, $L = \ell \geq 0$, and source standard deviation s_0, the target standard deviation is

$$s = \frac{s_0 + \sqrt{s_0^2 + 4h[1 + h(\lambda + \ell)]}}{2[1 + h(\lambda + \ell)]}. \tag{P.10}$$

The mean contraction still uses λ, not $\lambda + \ell$. The interaction affects fluctuations about the mean, and cannot damp a common translation by itself.

P.4 Exact continuous physical time

The gradient-flow equation is

$$\partial_t \rho = \Delta\rho + \mathrm{div}\left(\rho[Kx + L(x - m(t))]\right). \tag{P.11}$$

For a Gaussian initial law the mean and covariance satisfy

$$m'(t) = -Km(t), \qquad \Sigma'(t) = 2I - M\Sigma(t) - \Sigma(t)M. \tag{P.12}$$

These identities follow either by integrating the equation against x and xx^T or by the corresponding linear stochastic differential equation. Their explicit solutions are

$$\begin{aligned} m(t) &= e^{-Kt} m_0, \\ \Sigma(t) &= e^{-Mt}\Sigma_0 e^{-Mt} + 2\int_0^t e^{-2Ms}\,\mathrm{d}s \\ &= M^{-1} + e^{-Mt}(\Sigma_0 - M^{-1})e^{-Mt}. \end{aligned} \tag{P.13}$$

No commutativity of Σ_0 and M was used. The last expression uses only that M commutes with its own exponential and inverse.

P.5 The interpolation clock is explicit as well

Let $d_m = m - m_0$, $D = A - I$, and $B_\tau = I + \tau D$. The interpolated mean is $m_\tau = m_0 + \tau d_m$, while its centered transport is $B_\tau z$. In this affine case $c_\tau = \dot{c}_\tau = 0$, and (J.10) reduces to

$$\dot{A}_\tau(x) = Kd_m + [MD + B_\tau^{-1} D B_\tau^{-1} C_0]z. \tag{P.14}$$

Accordingly its squared norm is $|Kd_m|^2$ plus the covariance trace of the bracketed matrix. A supremum over the compact interval $\tau \in [0, 1]$ can be enclosed by one-dimensional interval subdivision. Matrices need not commute with C_0; their order in (P.14) follows directly from the general derivative formula.

P.6 Uniform small-step regularity on compact parameter sets

Suppose K, L range over a bounded set with a common positive lower bound for K, the eigenvalues of Σ_0 stay in $[s_-, s_+]$ with $s_- > 0$, and $|m_0|$ stays bounded. Multiplying (P.6) by h gives

$$hAMA + A^2 - A - hC_0 = 0. \tag{P.15}$$

At $h = 0, A = I$, the derivative with respect to the symmetric matrix A is the identity map: the derivative of $A^2 - A$ at I is $H \mapsto H$. The finite-dimensional implicit function theorem, uniformly on the stated compact parameter set, gives

$$A = I + h(C_0 - M) + O(h^2), \qquad m = m_0 - hKm_0 + O(h^2). \tag{P.16}$$

The constants are uniform on that set. Substitution into (P.14), with B_τ^{-1} uniformly bounded, gives $C_h^* \leq Kh$ for a finite common constant. Thus the nonaccumulation hypothesis is nonempty and explicitly verifiable in this coupled noncommuting model. This is an analytic model class, not an empirical benchmark claim.

Q EXACT NONLINEAR COUPLED MODELS WITH NEURAL MINIMIZERS

Q.1 A NONAFFINE MODEL CONSTRUCTED FROM AN ADMISSIBLE TRANSPORT

Gaussian closure alone would not show that the nonlinear theory has examples. We construct a strongly convex potential whose unrestricted JKO minimizer is an exactly specified nonaffine neural transport. This is a manufactured analytical family, not a claim that an arbitrary prescribed potential can be solved by this construction.

Fix $h > 0$ and a nonzero $\psi \in C_c^\infty(\mathbb{R}^d)$. Choose ψ to be a finite tensor-product bump potential from (E.7), and let

$$u_\epsilon(x) = \tfrac{1}{2}|x|^2 + \epsilon\psi(x), \qquad T_\epsilon(x) = x + \epsilon\nabla\psi(x). \tag{Q.1}$$

For $|\epsilon| \left\|D^2\psi\right\|_{\infty,\mathrm{op}} < 1/2$, the Hessian of u_ϵ has eigenvalues in $[1/2, 3/2]$. Thus T_ϵ is a smooth global Brenier diffeomorphism, equal to the identity outside a compact set. Let $S_\epsilon = T_\epsilon^{-1}$ and $\nu_\epsilon = (T_\epsilon)_\# \gamma_d$, with density ρ_ϵ. Define

$$V_\epsilon(y) = \frac{u_\epsilon^*(y) - |y|^2/2}{h} - \log \rho_\epsilon(y). \tag{Q.2}$$

Every term is an exactly defined smooth function. The inverse and conjugate in this definition are not assumed to have elementary closed forms.

Proposition Q.1 (Zero residual by construction)**.** *For source γ_d, potential* (Q.2)*, and $W = 0$, the map T_ϵ has identically zero residual.*

Proof. The conjugate gradient is $\nabla u_\epsilon^*(y) = S_\epsilon(y)$. Hence

$$\nabla V_\epsilon(y) = \frac{S_\epsilon(y) - y}{h} - \nabla \log \rho_\epsilon(y). \tag{Q.3}$$

At $y = T_\epsilon(x)$, the source point is $S_\epsilon(y) = x$. Adding the proximal displacement $(T_\epsilon(x) - x)/h$ and the pulled-back target score cancels both terms. This is exactly the full residual formula. □

Q.2 STRONG CONVEXITY OF THE MANUFACTURED ENERGY

At $\epsilon = 0$, $u_0^* = |y|^2/2$, $\nu_0 = \gamma_d$, and $V_0(y) = |y|^2/2 + (d/2)\log(2\pi)$. The next statement shows that the construction remains in the convex energy class for sufficiently small nonzero ϵ.

Lemma Q.2 (Uniform C^2 control of the manufactured potential)**.** *For fixed h and fixed $\psi \in C_c^\infty$, there are $\epsilon_0 > 0$ and $C_h < \infty$ such that, for $|\epsilon| \le \epsilon_0$,*

$$\left\|\nabla^2 V_\epsilon - I\right\|_{\infty,\mathrm{op}} \le C_h|\epsilon|. \tag{Q.4}$$

In particular, after reducing ϵ_0 if necessary, $\nabla^2 V_\epsilon \succeq I/2$ globally. The potential equals a quadratic plus a constant outside a fixed compact set.

Proof. Choose a ball containing the support of ψ and its derivatives. The image of that ball under all T_ϵ with sufficiently small $|\epsilon|$ lies in a fixed larger ball. Outside the latter ball, $S_\epsilon(y) = y$, $u_\epsilon^*(y) = |y|^2/2$, and $\rho_\epsilon(y) = \phi_d(y)$. Thus the potential has the asserted exact quadratic tail.

On the larger ball, the inverse equation $S_\epsilon(y) + \epsilon\nabla\psi(S_\epsilon(y)) = y$ and the uniform curvature bound imply

$$S_\epsilon - \mathrm{Id} = O(\epsilon), \quad DS_\epsilon - I = O(\epsilon), \quad D^2S_\epsilon = O(\epsilon), \quad D^3S_\epsilon = O(\epsilon) \tag{Q.5}$$

uniformly. To verify these estimates, differentiate the inverse equation successively. At each stage the highest derivative of S_ϵ is multiplied by $I + \epsilon D^2\psi(S_\epsilon)$, whose inverse norm is at most two. The other terms contain an explicit factor ϵ and bounded derivatives of ψ and lower derivatives of S_ϵ. Induction proves the displayed estimates.

The density identity is

$$-\log \rho_\epsilon(y) = \tfrac{1}{2}|S_\epsilon(y)|^2 + \tfrac{d}{2}\log(2\pi) + \log\det\big(I + \epsilon D^2\psi(S_\epsilon(y))\big). \tag{Q.6}$$

Differentiating twice on the fixed ball shows that the Hessian of the first term differs from I by $O(\epsilon)$ and that the Hessian of the log-determinant term is $O(\epsilon)$. Bounded derivatives of ψ through order four suffice for these estimates. Finally $\nabla^2 u_\epsilon^* = DS_\epsilon = I + O(\epsilon)$. Insert these bounds into (Q.2), retaining the fixed factor h^{-1}, to obtain (Q.4). Outside the ball the difference vanishes, so the estimate is global. □

Combining the lemma with the zero-residual proposition and the global certificate proves that ν_ϵ is the unique unrestricted JKO minimizer. This is an exact theoretical conclusion, not numerical evidence based on small residual values.

Q.3 Non-Gaussian and coupled behavior

Choose $d \geq 2$ and, for example, a product bump $\psi(x) = \prod_{i=1}^{d} \beta(x_i)$ with nonzero coefficient. Its mixed second derivatives are not identically zero. The map T_ϵ therefore couples coordinates on part of its support and is not affine. It is an exactly represented member of the finite bump architecture.

The output is not Gaussian for $\epsilon \neq 0$. Indeed the Brenier map between two nondegenerate Gaussian laws is affine. Uniqueness of the Brenier map would force T_ϵ to be affine if its output were Gaussian, contradicting the nonzero compactly supported perturbation. Since the map equals the identity outside a compact set, even an affine alternative could only be the identity, which would force $\nabla\psi = 0$ everywhere. A nonzero compact bump does not have this property.

The example establishes nonaffine, coordinate-coupled, non-Gaussian behavior. We do not need to classify every possible orthogonal change of coordinates or claim that every coupled target admits a compact bump representation. Its role is to demonstrate that the full-space theorem is not merely a repackaging of a Gaussian or scalar calculation.

Q.4 Adding a nontrivial convex interaction

The same target can be retained while introducing a convex interaction. Let $L \succeq 0$ be a nonzero matrix with $\|L\|_{\mathrm{op}} \leq 1/4$ and set $W_L(y) = y^T L y/2$. Define

$$\widetilde{V}_\epsilon(y) = V_\epsilon(y) - (W_L * \nu_\epsilon)(y). \tag{Q.7}$$

The convolution is a quadratic polynomial:

$$(W_L * \nu_\epsilon)(y) = \tfrac{1}{2} y^T L y - y^T L m_\epsilon + \tfrac{1}{2}\mathbb{E}_{Y\sim\nu_\epsilon} Y^T L Y. \tag{Q.8}$$

Consequently $\nabla^2 \widetilde{V}_\epsilon = \nabla^2 V_\epsilon - L \succeq I/4$ for the small ϵ chosen above. At the target law, the added interaction force exactly cancels the subtracted convolution gradient. Thus the full residual remains zero for the energy with potential $\widetilde{V}_\epsilon$ and interaction W_L.

The constant involving the second moment has no effect on the residual or minimizer. If ψ is even, symmetry gives $m_\epsilon = 0$. This yields a fully specified nonlinear model with a genuinely nonzero interaction and a strongly convex external potential.

Q.5 Computability of the manufactured coefficients and evaluations

For computable h, ϵ, and bump coefficients, the inverse value $S_\epsilon(y)$ is the unique minimizer of $u_\epsilon(x) - y \cdot x$. Its gradient is strongly monotone with constant at least $1/2$. An approximate point x_0 has the verified distance estimate

$$|S_\epsilon(y) - x_0| \leq 2|T_\epsilon(x_0) - y|. \tag{Q.9}$$

Together with a global localization bound and interval evaluation, this gives convergent inverse enclosures. Differentiating the inverse equation gives enclosures for the derivatives used in (Q.6). Hence the potential admits the required compact-domain evaluations and exact quadratic-tail envelopes.

This construction is a proof of existence and computability for a manufactured family. It is not used as a fast numerical benchmark, and no unrun experiments are reported. Its exact residual cancellation also provides a useful symbolic test case for independently checking signs in the entropy and proximal terms.

R UNCERTAIN SCORES, APPROXIMATE FORCES, AND EXPORTED MODELS

R.1 A CERTIFIED SOURCE-SCORE ERROR

Suppose the source law itself is fixed and known, but the score used in evaluation is $\widehat{s}_\mu$ rather than s_μ. Assume an independently valid bound

$$\|\widehat{s}_\mu - s_\mu\|_{L^2(\mu)} \le \sigma. \tag{R.1}$$

Let $\widehat{r}$ be the residual obtained by replacing the source score and changing no other term. Since the score enters linearly,

$$r_h[T] - \widehat{r} = (DT)^{-1}(s_\mu - \widehat{s}_\mu), \qquad \|r_h[T]\|_{L^2(\mu)} \le \|\widehat{r}\|_{L^2(\mu)} + \sigma/m. \tag{R.2}$$

Thus an audited $\widehat{R} \ge \|\widehat{r}\|_{L^2(\mu)}$ yields the valid replacement

$$R = \widehat{R} + \sigma/m. \tag{R.3}$$

The bound σ must concern the population $L^2(\mu)$ norm. A small score-training loss evaluated on reused particles does not establish it.

The clock correction is different. Let $D = D\Delta$ and $b_* = \min\{1, m\}$. In (J.10) the source score appears only in $-B_\tau^{-1} D B_\tau^{-1} s_\mu$. Therefore

$$C = \widehat{C} + b_*^{-2} \|D\Delta\|_{\infty,\mathrm{op}}\, \sigma \tag{R.4}$$

is valid when $\widehat{C}$ bounds the clock derivative computed with $\widehat{s}_\mu$. Omitting this correction while enlarging only the endpoint residual can leave the physical-time ledger unjustified.

R.2 APPROXIMATE INTERACTION FORCES

For a fixed candidate, suppose a force approximation $\widehat{I}[T]$ has a certified error $\left\|\widehat{I}[T] - I[T]\right\|_{L^2(\mu)} \le \eta_I$. Add η_I to the endpoint residual bound. To use the derivative-based clock estimate, one needs a corresponding uniform bound for the interpolation-force derivative:

$$\sup_{\tau\in[0,1]} \left\|\partial_\tau \widehat{I}[Q_\tau] - \partial_\tau I[Q_\tau]\right\|_{L^2(\mu)} \le \eta_{I,1}. \tag{R.5}$$

Then add $\eta_{I,1}$ to the clock bound. A bound on the force itself at $\tau = 1$ does not imply this derivative estimate.

There is an alternative when only a uniform force error along the whole path is known. If $\sup_\tau \left\|\widehat{I}[Q_\tau] - I[Q_\tau]\right\|_{L^2(\mu)} \le \eta_{I,0}$, a direct audit of the approximate physical defect can be enlarged by $\eta_{I,0}$ at every time. Apply the continuous stability lemma to that enlarged envelope. This does not require differentiating the force error, but it may lose the favorable vanishing factor $(1-\tau)$ in the clock decomposition.

For polynomial interactions, moment intervals can be propagated directly through formulas such as (E.11). The interval bounds then enclose the true force without introducing a separate unverified moment estimator. Repeated moment appearances can increase interval width, but cannot reverse the guarantee.

R.3 INITIAL-LAW UNCERTAINTY IS NOT SCORE UNCERTAINTY

Suppose σ_t is the exact flow from a desired initial law σ_0, while the numerical source starts at a represented law μ_0. If $W_2(\mu_0, \sigma_0) \le E_0$ is known, the ledger simply starts from E_0. No score of σ_0 is needed for this comparison, because the numerical residual is computed relative to the actual represented source μ_0 and the exact flow is compared through measure-level stability.

This does not imply that a score learned from σ_0 is an accurate score for μ_0. Wasserstein proximity alone does not control Fisher-information-type quantities. Conversely, a source-score error bound without a law-distance bound does not certify the desired initial condition. The two inputs enter different parts of the theorem and should be recorded separately.

R.4 Model export and parameter rounding

The certificate concerns one exact exported map. Floating-point optimization may produce a vector of machine numbers, after which each number is interpreted as an exact rational binary value or an exact decimal value. The auditor must use that interpretation consistently. It then audits the exported model from scratch; correctness of the optimizer's floating-point trajectory is irrelevant to validity.

If instead one wishes to transfer a certificate from a mathematical map T to a perturbed exported map $\widetilde{T}$, first verify that the curvature lower bound survives. On a common compact perturbation region, if $\left\|D\widetilde{T} - DT\right\|_{\infty,\mathrm{op}} < m/2$, then $D\widetilde{T} \succeq mI/2$. The graph perturbation estimates from Appendix M then provide an explicit residual correction. The clock needs its own derivative perturbation estimate or a fresh audit. Merely observing that parameter rounding was small does not control either quantity without a parameter-to-derivative bound.

For the ridge architecture, such bounds can be derived from the displayed sigmoid derivatives and coefficient intervals. For the bump architecture they depend on inverse powers of the feature widths. Narrow features can amplify apparently small coefficient or center errors. This is one reason to freeze the exact export before auditing, rather than interpret a certificate as robust under arbitrary later edits.

R.5 Transferring a law-distance certificate to a different output

Suppose a Brenier map T has a certified radius η_T to the JKO minimizer, but the delivered generator is another measurable map H from the same source. If an independent coupling bound satisfies $\|H - T\|_{L^2(\mu)} \le \delta$, then

$$W_2(H_{\#}\mu, \nu_h^*) \le W_2(H_{\#}\mu, T_{\#}\mu) + W_2(T_{\#}\mu, \nu_h^*) \le \delta + \eta_T. \tag{R.6}$$

No invertibility of H is needed for this distance statement. It does not transfer an objective-gap certificate: entropy is not uniformly continuous in W_2, so a nearby output law can have very different entropy or even be singular. For an objective certificate of a non-Brenier diffeomorphism, use the separate primal–dual upper evaluation in Appendix O.

R.6 Perturbed energy fields

Let an approximate curve be generated using a different energy field $\widehat{\xi}$ and suppose a valid along-curve bound is known:

$$\left\|\widehat{\xi}_t - \xi_t\right\|_{L^2(\rho_t)} \le \delta_t. \tag{R.7}$$

If the approximate physical defect relative to $\widehat{\xi}$ is bounded by $\widehat{f}_t$, its defect relative to the desired energy is bounded by $\widehat{f}_t + \delta_t$. The stability lemma then gives

$$W_2(\rho_t, \sigma_t) \le e^{-\lambda t} E_0 + \int_0^t e^{-\lambda(t-s)} (\widehat{f}_s + \delta_s)\, \mathrm{d}s. \tag{R.8}$$

For example, a uniform bound on $\nabla\widehat{V} - \nabla V$ and a uniform bound on the difference of interaction forces give such a δ_t. The curvature constant is that of the desired energy in the comparison theorem. This is a perturbation result for a flow field, not an unstated Lipschitz theorem for the multidimensional JKO resolvent under arbitrary source changes.

R.7 Combining deterministic and probabilistic inputs

If the score error, moment enclosures, or initial-law distance are probabilistic, let Ω be the event that all those inputs and all subsequent audit bounds hold simultaneously. On Ω, the deterministic inequalities in this paper apply without modification. The overall coverage is at least $1 - \delta$ when the probability of the complement of Ω is bounded by δ using the conditional allocation in Appendix I. Numerical roundoff, statistical estimation, and representation error are distinct terms; none is removed merely by labeling the final number a confidence bound.

S TOLERANCE ALLOCATION UNDER AN EXPLICIT WORK MODEL

S.1 THE TERMINAL CONTRIBUTION OF A LOCAL ERROR

For a fixed accepted time grid, the physical-time ledger has the linear form

$$E_N = e^{-\lambda T} E_0 + \sum_{j=1}^{J} w_j \eta_j, \qquad w_j > 0. \tag{S.1}$$

An index j may represent an endpoint residual contribution or a clock contribution. For example the weights multiplying R_k and C_k are respectively $e^{-\lambda(T-t_{k+1})} a_{h_k}$ and $e^{-\lambda(T-t_{k+1})} b_{h_k}$. Once the grid is fixed, allocating tolerances to these terms is a finite-dimensional optimization problem. It is different from proving that the underlying candidate search can realize each tolerance.

We analyze a stylized work model

$$\mathsf{Work}_j(\eta_j) = c_j \eta_j^{-p}, \qquad c_j > 0, \quad p > 0, \tag{S.2}$$

where smaller requested error costs more work. This model is an explicit assumption, not an empirical fit or a claimed complexity law for all neural solvers. The exponent could describe a certified quadrature refinement regime or a prescribed deterministic optimization routine. Actual algorithms may not obey it.

Proposition S.1 (Optimal weighted tolerance allocation)**.** *For $\varepsilon_b > 0$, the unique minimizer of*

$$\min_{\eta_j > 0} \sum_j c_j \eta_j^{-p} \quad \textit{subject to} \quad \sum_j w_j \eta_j \le \varepsilon_b \tag{S.3}$$

is

$$S = \sum_j c_j^{1/(p+1)} w_j^{p/(p+1)}, \qquad \eta_j^* = \frac{\varepsilon_b}{S} \left(\frac{c_j}{w_j} \right)^{1/(p+1)}. \tag{S.4}$$

The optimal work is $S^{p+1} \varepsilon_b^{-p}$.

Proof. The objective is strictly convex on the positive orthant and strictly decreasing in each coordinate, so the constraint is active at an optimum. The Lagrange equation for multiplier $\tau > 0$ is $-p c_j \eta_j^{-p-1} + \tau w_j = 0$. Thus $\eta_j = (p c_j / (\tau w_j))^{1/(p+1)}$. Substitution in the active constraint determines the common factor as ε_b / S, proving (S.4). Substituting back in the objective gives

$$\sum_j c_j \left[\frac{\varepsilon_b}{S} \left(\frac{c_j}{w_j} \right)^{1/(p+1)} \right]^{-p} = S^p \varepsilon_b^{-p} \sum_j c_j^{1/(p+1)} w_j^{p/(p+1)}. \tag{S.5}$$

This is the stated value. Strict convexity proves uniqueness. Existence follows as well from the positive explicit solution and the convex optimality conditions. □

This is a classical constrained resource-allocation calculation, not a new generic optimization theorem. Its relevance is that the weights have an exact distributional meaning supplied by the physical-time certificate.

S.2 WHAT THE ALLOCATION SAYS AND DOES NOT SAY

Earlier errors receive smaller terminal weights when $\lambda > 0$, because the exact flow contracts them before the terminal time. At equal c_j, formula (S.4) allows larger local errors where w_j is smaller. Conversely, an especially expensive audit receives a larger tolerance at fixed influence weight. These statements follow directly from the powers in the formula.

The allocation is terminal-time optimal under (S.2); it need not minimize the maximum error along the whole trajectory. Requiring a uniform tube adds intermediate-time constraints and can prevent large early local tolerances. The stronger tests in Appendix L are a simple feasible solution to that different problem.

If lower or upper admissible tolerances are present, the optimum is obtained from the same multiplier relation clipped to the permitted interval. The clipping threshold is determined by monotone one-dimensional search for the multiplier. This follows from the KKT conditions for the box-constrained convex problem. When some weights vanish, their errors do not affect the specified terminal bound; the positivity assumption in the proposition deliberately excludes such degenerate terms.

S.3 Residual budgets versus objective-gap budgets

For a fixed step, a residual threshold $R \le \alpha_h \eta$ certifies Wasserstein radius η. A gap threshold $g \le \alpha_h \eta^2/2$ gives the same radius through quadratic growth. Under the residual construction $g = R^2/(2\alpha_h)$, these two parameterizations coincide. But the cost of reducing R or g need not have the same power law: g is quadratic in R at fixed h.

Across varying steps, $\alpha_h = \lambda + h^{-1}$ changes. Thus imposing the same numerical gap tolerance on every step does not impose the same local law-distance tolerance. Likewise, imposing the same residual norm does not imply the same terminal contribution, because the factor a_h and the remaining physical time vary. The ledger exposes these conversions explicitly rather than burying them in optimizer hyperparameters.

S.4 A feasibility budget for the clock

Unlike an inner residual, the interpolation clock generally cannot be reduced to zero while holding an arbitrary nonzero time step and its exact minimizer fixed. The time-grid design must therefore account for a lower attainable clock level. In a regime with $C_h^* \le Kh$ and accurate inner solves, the local clock contribution is at most $Kh^2/2$. Summing over the grid gives a first-order $O(h_{\max})$ bound. Reducing the optimization tolerance far below that level cannot improve this particular guaranteed order.

A practical theoretical policy is to reserve a prescribed fraction of the total budget for time discretization and allocate the remaining fraction among residual audits using (S.4). This is valid only after the time-grid clock contribution has itself been bounded. Guessing the constant K from a few runs without a valid envelope is not a certificate.

S.5 A distinction between cost optimality and validity

The main certificate theorem is independent of the work model. Even a highly inefficient schedule is mathematically sound when all accepted bounds are valid. Conversely, an apparently efficient empirical stopping rule is not sound merely because it minimizes observed runtime. The resource-allocation result concerns the efficiency of a family of already valid tolerances under an additional cost assumption; it does not turn an invalid diagnostic into a certificate.

T An anisotropic certificate separating mean and fluctuations

T.1 Matrix curvature contains more information than one scalar

The scalar constant $\alpha_h = \lambda + h^{-1}$ is convenient but can discard known curvature. Suppose stronger matrix lower bounds are available:

$$\nabla^2 V(y) \succeq K_0 \succeq 0, \qquad \nabla^2 W(y) \succeq L_0 \succeq 0 \quad \text{for every } y. \tag{T.1}$$

The two matrices are symmetric and need not commute. Define

$$A_0 = K_0 + h^{-1} I, \qquad A_1 = A_0 + L_0. \tag{T.2}$$

For a field $D \in L^2(\mu; \mathbb{R}^d)$, write $\bar{D} = \mathbb{E}_\mu D$ and $D^\circ = D - \bar{D}$. Introduce the bounded, self-adjoint, strictly positive Hilbert-space operator

$$\mathsf{M} D = A_0 \bar{D} + A_1 D^\circ = A_0 D + L_0 (D - \bar{D}). \tag{T.3}$$

Constants and mean-zero fields are orthogonal subspaces, and both are invariant under this operator. Its quadratic form and inverse are

$$\begin{aligned} \|D\|_{\mathsf{M}}^2 &= \bar{D}^T A_0 \bar{D} + \mathbb{E}_\mu[(D^\circ)^T A_1 D^\circ], \\ \mathsf{M}^{-1} r &= A_0^{-1} \bar{r} + A_1^{-1} (r - \bar{r}). \end{aligned} \tag{T.4}$$

This is a norm on source-coordinate vector fields for the fixed step. It is not a new global metric imposed on all of Wasserstein space.

T.2 The interaction curvature acts only on fluctuations

Let S, T be comparison maps and $D = S - T$. Strong convexity of W with matrix L_0 gives, for each pair (x, y), a quadratic remainder $\frac{1}{2}(D(x) - D(y))^T L_0 (D(x) - D(y))$. After multiplication by the factor $1/2$ in the interaction energy, the remainder becomes

$$\begin{aligned}\frac{1}{4}\mathbb{E}_{\mu\otimes\mu}[(D(X) - D(Y))^T L_0 (D(X) - D(Y))] &= \frac{1}{2}\left(\mathbb{E}_\mu D^T L_0 D - \bar{D}^T L_0 \bar{D}\right) \\ &= \frac{1}{2}\mathbb{E}_\mu[(D^\circ)^T L_0 D^\circ]. \end{aligned} \tag{T.5}$$

The identity follows by expanding the square and using independence for the cross term. A common translation has $D^\circ = 0$, so interaction curvature supplies no translation curvature. This prevents the incorrect replacement of λ by $\lambda + \lambda_{\min}(L_0)$ in all directions.

The potential contributes $\mathbb{E}D^T K_0 D/2$, and the proximal term contributes $\mathbb{E}|D|^2/(2h)$. Adding (T.5) yields $\|D\|^2_{\mathsf{M}}/2$. The entropy support proof is unchanged.

Theorem T.1 (Matrix-weighted full-space certificate)**.** *Under the source and candidate hypotheses and* (T.1)*, every finite-energy comparison law $\eta = S_{\#}\mu$ satisfies*

$$\mathcal{J}_{h,\mu}(\eta) \geq P_h(T) + \langle r_h[T], S - T\rangle_{L^2(\mu)} + \tfrac{1}{2}\|S - T\|^2_{\mathsf{M}}. \tag{T.6}$$

If S^* is the unrestricted minimizing map and $r = r_h[T]$, then

$$\begin{aligned} P_h(T) - \mathcal{J}^* &\leq \tfrac{1}{2}\|r\|^2_{\mathsf{M}^{-1}}, \\ \|T - S^*\|_{\mathsf{M}} &\leq \|r\|_{\mathsf{M}^{-1}}, \\ W_2(T_{\#}\mu, \nu_h^*) &\leq \frac{\|r\|_{\mathsf{M}^{-1}}}{\sqrt{\lambda_{\min}(A_0)}}. \end{aligned} \tag{T.7}$$

The right sides can be replaced by any independently valid upper enclosures.

Proof. The preceding curvature calculation proves (T.6). Repeating generalized-geodesic convexity with the same matrix remainders gives quadratic growth at the optimum: $P_h(T) - \mathcal{J}^* \geq \|T - S^*\|^2_{\mathsf{M}}/2$. At S^*, the support inequality gives the upper bound

$$P_h(T) - \mathcal{J}^* \leq \langle r, T - S^*\rangle - \tfrac{1}{2}\|T - S^*\|^2_{\mathsf{M}}. \tag{T.8}$$

Complete the square in the M inner product, or apply the dual-norm inequality, to bound this expression by $\|r\|^2_{\mathsf{M}^{-1}}/2$. Combining upper and lower growth as in the scalar proof gives $\|T - S^*\|_{\mathsf{M}} \leq \|r\|_{\mathsf{M}^{-1}}$. Finally $\mathsf{M} \succeq \lambda_{\min}(A_0) I$ on $L^2(\mu)$, and the common-source coupling bounds W_2 by $\|T - S^*\|_{L^2(\mu)}$. This proves the last inequality. □

The source mean and fluctuations are separated only inside this fixed-source quadratic form. The theorem does not claim that the unknown displacement's mean and fluctuating errors can each be bounded by just the corresponding residual component without accounting for their possible nonlinear coupling in the energy.

T.3 A subtraction-free expression for auditing

The squared dual norm can be written

$$\|r\|^2_{\mathsf{M}^{-1}} = \mathbb{E}_\mu[r^T A_1^{-1} r] + \bar{r}^T (A_0^{-1} - A_1^{-1})\bar{r}. \tag{T.9}$$

Both terms are nonnegative. To verify this, first expand the centered expression from (T.4), and then note the inverse-order inequality $A_1 \succeq A_0 \succ 0 \Rightarrow A_0^{-1} - A_1^{-1} \succeq 0$. One proof writes $A_1 = A_0^{1/2}(I + A_0^{-1/2} L_0 A_0^{-1/2}) A_0^{1/2}$ and uses that the inverse of the middle matrix is at most the identity. No commutativity assumption is needed.

An auditor can enclose $\bar{r}$ componentwise and the expectation of $r^T A_1^{-1} r$ using the same Gaussian core-and-tail construction. The second term is then bounded over the mean enclosure by interval matrix arithmetic or a separate small convex quadratic maximization over its box. Evaluating a centered second moment as the difference of two uncertain positive quantities without the proper interval directions would be unsafe. Formula (T.9) avoids that subtraction.

When $K_0 \succeq \lambda I$, the operator satisfies $\mathsf{M} \succeq \alpha_h I$. Thus

$$\tfrac{1}{2}\|r\|^2_{\mathsf{M}^{-1}} \le \frac{\|r\|^2_{L^2(\mu)}}{2\alpha_h}. \tag{T.10}$$

At the level of exact residual evaluation, the anisotropic gap is never worse than the scalar one. Numerical enclosure overhead can still make a particular computed anisotropic bound less useful; taking the minimum of two valid bounds remains safe.

T.4 An explicit improvement for a centered residual

Suppose $K_0 = \lambda I$ and $L_0 = \ell I$, with $\ell > 0$. If $\bar{r} = 0$, then

$$\tfrac{1}{2}\|r\|^2_{\mathsf{M}^{-1}} = \frac{\|r\|^2_{L^2(\mu)}}{2(\lambda + \ell + h^{-1})}. \tag{T.11}$$

The gap improvement factor over the scalar certificate is $(\lambda + \ell + h^{-1})/(\lambda + h^{-1})$. If instead r is constant, the dual norm is $|r|^2/(\lambda + h^{-1})$, with no improvement from ℓ. These two exact cases explain why mean/fluctuation splitting is more informative than simply adding interaction curvature to a scalar constant.

For a quantitative model with a small but nonzero mean residual, (T.4) interpolates continuously between these cases. The main difficulty is not defining the weighted norm but evaluating it globally and retaining the mean uncertainty with the correct sign.

T.5 Relation to the physical-time controller

The anisotropic certificate can replace the scalar gap or single-step radius when deciding whether a local JKO approximation is acceptable. The physical-time theorem still uses an L^2 bound on the actual forcing defect and the EVI curvature of the energy. It does not automatically inherit the step-dependent operator M, because that operator includes the artificial proximal curvature $h^{-1}I$ and is tied to one source chart.

A sound hybrid controller may therefore use the sharper anisotropic certificate for local optimality, while retaining the original endpoint and interpolation norm audits for physical time. Replacing the flow contraction exponent by an eigenvalue of M would be incorrect: the h^{-1} contribution belongs to the JKO objective, not to the continuous energy gradient flow.

U Proof dependencies and the finite certificate contract

U.1 What a single-step record must specify

A mathematical certificate must identify the problem to which it applies. For an accepted step the record consists of the exact source representation, energy, step size, candidate coefficients, and verified curvature margins. It also specifies the quantities being bounded: the full residual norm, the interpolation derivative norm when physical time is claimed, and all auxiliary errors introduced by approximations. Without the source law, a residual number does not determine a JKO problem.

For deterministic Gaussian-reference integration, the finite record contains a core domain, its finite partition, pointwise enclosures on every cell, enclosures of the Gaussian cell masses, and the coefficients and evaluation of the tail majorant. A clock audit additionally records the time-parameter partition and uses the maximum of integrals over spatial cells at each time-parameter cell, as in Appendix G. An implementation that instead reports an integral of a pointwise supremum may still be sound but has a different, potentially much looser, convergence target.

The record also contains the accepted error ledger and arithmetic enclosures for its coefficients. A downstream verifier need not trust the optimizer or know how the candidate was trained. It needs to check that the supplied finite inequalities cover the actual exported model and the entire integration domain. This separation between an untrusted proposer and an independently checked auditor is the algorithmic interface studied in the paper.

U.2 Dependencies of the principal claims

Claim	Required inputs and proof location
Unrestricted objective gap	Source density and moments, convex energy, a relative Brenier map, and a valid full residual norm; Appendices B–C.
One-step law radius	The same inputs plus common-source quadratic growth, proved without regularity of the exact optimum; Appendix B.
Finite norm enclosure	Frozen computable coefficients, sound convergent compact enclosures, and an integrable computable tail; Appendix G.
Probability coverage	Fresh conditional samples, valid bounded envelopes, and a summable allocation over all audits; Appendix I.
Physical-time error	A valid interpolation defect and a displacement-convex EVI flow; Appendices J–K.
Finite candidate acceptance	Existence of a strict-slack auditable candidate and fair search; Appendix N.
Finite horizon advancement	Uniform small-step graph regularity and a safeguarded time scheduler; Appendix N.
Non-Brenier output gap	A valid coupling-based primal upper bound and Gibbs/transport dual lower bound, with $W = 0$; Appendix O.

No row follows solely from convergence of a parameter gradient. The dependency structure prevents an assumption needed for termination from being mistaken for an assumption of the already valid local inequality, and prevents a local inequality from being mistaken for a continuous-time guarantee.

U.3 External mathematical inputs

The transport foundations are Brenier's optimal-map theorem, the almost-everywhere Jacobian identity for absolutely continuous source and target laws, and generalized displacement convexity of entropy. The flow foundations are the EVI theory and distance/velocity calculus in the classical convex setting. These are cited standard results, not original claims (Brenier, 1991; McCann, 1997; Ambrosio et al., 2008; Villani, 2009; Santambrogio, 2015). The proof involving distributional Hessians makes explicit how the smooth candidate is compared with nonsmooth optimal maps; it does not replace the foundational Monge–Ampère theorem.

The concentration inequality in the optional stochastic audit is Hoeffding's bounded-sum inequality (Hoeffding, 1963). The matrix square-root and implicit-function arguments in the Gaussian example use ordinary finite-dimensional spectral calculus. The interval contract assumes sound outward enclosures, whose software verification is a separate numerical-analysis issue (Rump, 2010). None of these external results is claimed as a new contribution.

U.4 The scope of novelty

Computable variational diagnostics, primal–dual Wasserstein optimization, inexact JKO convergence, and residual-based Wasserstein stability already exist in the literature (Carrillo et al., 2022; Di Marino et al., 2025; Chazal et al., 2026; Gonnella et al., 2026). The present theoretical construction links a full-space neural transport residual to a global audit and a separately differentiated physical-time defect in a multidimensional convex interaction model. It further makes representation failure and finite-information obstructions explicit. A priority claim stronger than this scoped comparison would require a broader independent literature assessment.

The architecture is chosen for transparent derivatives, curvature, and tails, not because its training complexity is proved superior to existing neural solvers. The manufactured nonlinear model is an existence example for the assumptions, not a newly discovered physical law. The Gaussian family

is an exact analytical calibration case, not evidence of high-dimensional nonlinear computational scalability.

U.5 Auxiliary code and what it verifies

The accompanying small code package evaluates the closed Gaussian formulas, the error-ledger coefficients, and the resource-allocation identities, and runs numerical or symbolic consistency tests. It is not a full implementation of the multidimensional global auditor. Its floating-point checks are not proofs of the theorems, and the manuscript's results do not depend on unreported empirical runs. No trained-model benchmark or experimental performance comparison is claimed in this pure-theory version.

A proof-assistant formalization and a verified numerical implementation would have distinct roles. The former could check the mathematical deductions relative to formalized transport foundations; the latter could verify individual finite bounds for exported models. This manuscript supplies mathematical proofs and a precise numerical contract, but does not assert that either entire formal-verification project has been completed.

U.6 Limitations that remain after the proofs

The source must be a genuine density with the required score and moments, or have separately certified approximations. Atomic empirical measures cannot be inserted into the displayed entropy-score formula. The potential and interaction are convex in the stated senses. Nonconvex energies require a different error-bound and stability analysis; adding a neural architecture does not create those properties.

The relative maps have strictly positive curvature. General generators can be handled by the separate dual route only under its own conditions. The full multidimensional common-source map distance is an upper coupling bound, not an isometry of Wasserstein space. The trajectory comparison deliberately avoids a general resolvent-contraction claim.

Finally, the global integration can have exponential dependence on dimension and severe dependence on narrow feature widths and inverse curvature. The finite-search theorem may be computationally impractical. The nonaccumulation theorem is conditional on uniform regularity and invariance of a source class. These are substantive limitations of the conclusions, not merely implementation details omitted from a universal guarantee.

U.7 The answer to the training question

For a proposed relative Brenier map, training is sufficient for a requested one-step Wasserstein radius η once an independent full-space audit returns $R \leq (\lambda + h^{-1})\eta$. For a requested physical-time accuracy, that endpoint condition is not enough: the interpolation defect and the initial-law error must also fit a valid accumulated ledger. The certificate may require more training, a richer model, a better audit, or a smaller time step. A successful return identifies which mathematical inequalities have passed; an unsuccessful return is a failure to certify, not an assertion that a training loss has become small enough by analogy.