IN-CONTEXT MAPS AS A COLLISION OPERATOR: STOCHASTIC JUMP TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a *stochastic jump transformer*, an extension of attention in which learned, context-dependent interaction rates govern the occurrence of pairwise jumps, while learned jump maps determine their outcomes. Composing small-time blocks yields McKean–Vlasov jump dynamics, extending the Vlasov continuous-depth limit of deterministic transformers. On compact convex state spaces, rate attention extends transformer universality from continuous deterministic in-context maps to continuous conditional laws, with deterministic maps recovered as Dirac-valued kernels. Because individual interactions are represented explicitly, physical structure can be imposed directly on the jump mechanism. Elastic pair maps, for example, preserve momentum and energy exactly; reversible jumps with symmetric forward–reverse rates yield entropy dissipation. For spatially nonhomogeneous Maxwell–Boltzmann dynamics with a fixed spatial mollifier, we show that their conditional transition laws fit into the stochastic jump transformer framework. More broadly, this suggests a kinetic route to physics foundation models: learn reusable, structure-preserving local interaction laws and compose them across particles and time.

1 INTRODUCTION

We introduce an architecture, driven by stochastic jump interactions, that extends Vlasov transformer dynamics to McKean–Vlasov jump dynamics. The motivation comes from kinetic theory, where macroscopic motion emerges from microscopic interactions. For a gas, one evolves the distribution of molecular states, with suitable scaling limits connecting this evolution to fluid equations (Bardos et al., 1991; Golse and Saint-Raymond, 2004). This suggests a route to foundation models for physics: learn interaction laws, then compose them to predict collective dynamics. Learned particle simulators and neural operators already demonstrate the value of such reuse (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Li et al., 2021). Our question is how to make *random interactions* the primitive of a neural model shared across initial states and resolutions.

A transformer already composes interactions. An attention layer updates each token $x \in \mathbb{R}^d$ through an in-context map of the whole context $\mu^N = N^{-1}\sum_{i=1}^N \delta_{x_i}$: a multi-head attention layer can be written as $\mathrm{MAtt}_\theta(x_1,\dots,x_N) = \big(G_\theta(\mu^N, x_i)\big)_{i=1}^N$, where

$$G_\theta(\mu^N, x) = x + \sum_{h=1}^H W_h \int \frac{\exp\Big(\langle Q_h x, K_h y\rangle/\sqrt{k}\Big)}{\displaystyle\int \exp\Big(\langle Q_h x, K_h y'\rangle/\sqrt{k}\Big)\ \mu^N(dy')} V_h y\, \mu^N(dy).$$

Here Q_h, K_h, V_h, W_h are the query, key, value, and output matrices of head h, H is the number of heads, and k is the query–key dimension (Vaswani et al., 2017). The layer moves the context to $G_\theta(\mu^N,\cdot)_\#\mu^N = N^{-1}\sum_i \delta_{G_\theta(\mu^N,x_i)}$, and deep transformers with small residual updates follow a Vlasov-type transport of this distribution (Furuya et al., 2025a; Castin et al., 2025). The update is deterministic, however, and a deterministic map cannot split an atom: a context supported on two states is sent to a context with at most two states, whereas the outgoing states of a collision can be spread over a continuum. For a scattering target whose output has more than two support points at such an input, deterministic in-context maps therefore incur a positive output-law approximation error; Proposition B.3 gives an explicit example. To represent scattering at a fixed input, a layer must send each token to a distribution rather than a single state.

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.

McKean–Vlasov jump processes provide the mathematical template: the current state distribution determines both the event rates and the state changes (Sznitman, 1991; Andreis et al., 2018; Erny, 2022; Chaintron & Diez, 2022). Boltzmann particle methods fit into this template by sampling collision partners and encounter geometries (Kac, 1956; Bird, 1963; Nanbu, 1980; Graham & Méléard, 1997). We build on the separation, common to both, between how often an interaction occurs and what it does. In the mechanism we call *rate attention*, attention reads the evolving context to assign a rate to a proposed pair interaction, a jump map specifies the pair's new states, and a random gate accepts the jump with probability about the step size times the rate. The *stochastic jump transformer* is the network built from such layers (Section 1.1).

Related work. Stochastic deep networks represent continuous maps between probability distributions (De Bie et al., 2019), and probabilistic transformers approximate probability-valued maps in Wasserstein distance (Kratsios, 2023). For deterministic in-context maps, transformers are universal on compact token domains (Furuya et al., 2025b). The support-preserving characterization of Furuya et al. (2025a) identifies conditions under which a measure transformation admits a continuous in-context pushforward representation. Continuous-depth analyses of transformers lead to mean-field or Vlasov-type transport equations (Castin et al., 2025; Furuya et al., 2025a), while neural jump models parameterize continuous motion together with event intensities and jump responses (Jia & Benson, 2019). A complementary kinetic-learning direction is NeurDE, which lifts nonlinear conservation laws to a kinetic representation and learns the local equilibrium within a BGK/lattice-Boltzmann evolution (Lara Benitez et al., 2025). The present architecture instead learns finite-rate pair interactions: small deterministic updates generate transport, while accepted pairwise events generate a McKean–Vlasov jump term. Randomness has also been introduced into attention for different purposes. YOSO uses Bernoulli sampling to estimate self-attention (Zeng et al., 2021), while SANTA samples the post-softmax value aggregation and also proposes Bernoulli sampling of query–key scores (Lee et al., 2026). Fedorov et al. (2026) obtain an interacting-particle diffusion driven by common Brownian noise from random value-matrix initialization under diffusion scaling and token-wise RMS normalization. Sinkformers use "doubly stochastic" in a different sense, imposing unit row and column sums on the attention matrix (Sander et al., 2022).

Results. On compact convex state spaces, Theorem A identifies the uniform W_1 closure of the architecture's conditional laws as the jointly continuous nonlinear Markov kernels: conditional transition laws that depend continuously on the context and input state. It also characterizes this class through the joint input-output law. Theorem B shows that composing small-step blocks yields a McKean–Vlasov drift–jump process whose kernels lie in this class, and Theorem B' gives conditions under which a family of kernels comes from one such mechanism. Because each interaction is explicit, physical structure can be imposed on the jump itself (Section 4): rotating the relative velocity of a pair conserves its momentum and energy for every learned parameter, and symmetrizing the rate yields an H-theorem. Section 4 also shows that the conditional kernels of spatially nonhomogeneous Maxwell–Boltzmann dynamics fit the representation and continuous-depth framework developed above.

1.1 Rate attention: in-context maps determine jump updates

We introduce *rate attention*, in which an in-context map assigns a nonnegative jump rate, or intensity, to a proposed pair interaction, and a separate jump map determines its update. Let $E \subset \mathbb{R}^d$ be the particle state space and $\mathcal{P}(E)$ its Borel probability measures. For $N \geq 2$ states $z_1, \dots, z_N \in E$, write $\mu^N = N^{-1} \sum_{i=1}^N \delta_{z_i} \in \mathcal{P}(E)$. As in Figure 1, the layer begins with a classical in-context map $G_\theta : \mathcal{P}(E) \times E \to E$, producing $\bar{z}_i = G_\theta(\mu^N, z_i)$ and $\bar{\mu}^N = G_\theta(\mu^N, \cdot)_\# \mu^N$ as in the introduction.

Propose a pair and choose its update. Select two distinct indices uniformly from $\{1, \dots, N\}$, with streamed states $\bar{z}, \bar{z}_*$. Independently draw a *mark* $\xi \sim \nu$ from a fixed law $\nu \in \mathcal{P}(\Xi)$, $\Xi \subseteq \mathbb{R}^{d_\xi}$. In a collision model the mark can specify an impact direction; in general it supplies randomness that distinguishes possible outcomes for the same pair. The proposed interaction is the pair query $q = (\bar{z}, \bar{z}_*, \xi) \in E^2 \times \Xi$, and two learned modules return its *intensity*, or rate, $\lambda(\bar{\mu}^N, q) \geq 0$ and its *pair output* $\Phi(\bar{\mu}^N, q) = (\Phi^1, \Phi^2)(\bar{\mu}^N, q) \in E^2$. For a step size $h > 0$, the acceptance probability of a pair interaction is $p_h(\lambda) = h\lambda$ (for rates with $0 \leq h\lambda \leq 1$) or $p_h(\lambda) = 1 - e^{-h\lambda}$. Draw a gate and

update the pair:

$$A \mid (\bar{\mu}^N, q) \sim \mathrm{Bernoulli}\big(p_h\big(\lambda_\vartheta(\bar{\mu}^N, q)\big)\big), \qquad (z^+, z_*^+) = \begin{cases} \Phi_\eta(\bar{\mu}^N, q), & A = 1, \\ (\bar{z}, \bar{z}_*), & A = 0. \end{cases} \tag{1}$$

The gate selects the jump or the identity (skip) branch. Rejection ($A = 0$) retains the streamed pair; acceptance ($A = 1$) replaces it, giving

$$\mu^{N,+} = \bar{\mu}^N + \frac{1}{N}\left(\delta_{z^+} + \delta_{z_*^+} - \delta_{\bar{z}} - \delta_{\bar{z}_*}\right). \tag{2}$$

The pair, mark, and acceptance draws are the sources of randomness. Both gate conventions satisfy $p_h(\lambda) = h\lambda + o(h)$: the probability of a jump vanishes with the step, while the jump itself may remain of order one. This is the small-time scaling of a finite-rate jump process, and it produces McKean–Vlasov jump dynamics in the continuous-depth limit (Section 3). For $\lambda > 0$, a Poisson clock of rate λ has independent exponential waiting times of mean $1/\lambda$; $1 - e^{-h\lambda}$ is the probability of at least one event during time h. Pair updates can be parallelized by drawing a random matching: for even N, every particle receives one proposal per round, giving a per-particle time step h. In Figure 1, only two particles receive a proposal, so a label is selected with probability $2/N$ and the corresponding physical step is $\Delta = 2h/N$; Appendix D.1 derives these physical-time normalizations and the required transport scaling. Algorithm 1 in Appendix D implements the jump stage in parallel by drawing a random *matching round* and updating all matched pairs simultaneously. We call the pair rule *exchange symmetric* if exchanging the two proposed states, together with a fixed measure-preserving relabeling of the mark, leaves the rate unchanged and exchanges the two outputs; the order of a proposed pair then does not matter in law, and averaging over the two orders imposes this exactly (Appendix A).

Compute the intensity with pair-query attention. The proposed pair reads the entire streamed context with its own key, query, and value matrices $Q_\vartheta \in \mathbb{R}^{k\times(2d+d_\xi)}$, $K_\vartheta \in \mathbb{R}^{k\times d}$, and $V_\vartheta \in \mathbb{R}^{d_{\mathrm{val}}\times d}$.

The score $s_\vartheta(q, w) = \exp\langle Q_\vartheta q, K_\vartheta w\rangle$ compares the proposed interaction with a context state $w \in E$. Its unnormalized read is

$$C_\vartheta(\bar{\mu}^N, q) = \int_E s_\vartheta(q, w) V_\vartheta w \, \bar{\mu}^N(dw) = \frac{1}{N}\sum_{i=1}^N s_\vartheta(q, \bar{z}_i) V_\vartheta \bar{z}_i \in \mathbb{R}^{d_{\mathrm{val}}}. \tag{3}$$

We omit the normalization by the total score so that the rate can use that score, for instance through a constant coordinate in the value (Appendix A); the read can then record how much of the context resembles the proposed pair, as a collision rate depends on the local density of partners. It does not directly update the state: it determines the event rate and the jump outcome, while the jump map Φ updates the selected pair when the jump is accepted. The two coefficients are

$$\lambda_\vartheta(\bar{\mu}^N, q) = \mathrm{MLP}_\vartheta\big(q, C_\vartheta(\bar{\mu}^N, q)\big), \qquad \Phi_\eta(\bar{\mu}^N, q) = \Pi_{E^2}\,\mathrm{MLP}_\eta\big(q, C_\vartheta(\bar{\mu}^N, q)\big), \tag{4}$$

where MLP_ϑ ends with a nonnegative activation such as ReLU (with its output bounded when required by the theory), and Π_{E^2} is the metric projection on the closed convex domains of Theorem A; physical jump maps instead enforce their output geometry directly. In this shared-read variant, ϑ contains the reader and rate-head parameters, η indexes the jump head, and Φ_η abbreviates $\Phi_{\vartheta,\eta}$. The weights are shared across pairs, while the rates and outcomes vary with the pair query and the context. Additional attention and MLP layers can be used to evaluate each coefficient without adding pair-update steps. We also study the restricted variant $\Phi_\eta(q)$, in which the jump map receives only the pair and mark, so all learned context dependence enters through λ_ϑ. Both variants recover the same conditional-kernel approximation class, even without the deterministic part (i.e. $G_\theta = \mathrm{Id}$); see Theorem A.

Composition of layers We extend the usual $\diamond$-composition of in-context maps (Furuya et al., 2025b) to pair queries. For either pair coefficient $\Psi \in \{\lambda_\vartheta, \Phi_\eta\}$, define

$$(\Psi \diamond G_\theta)(\mu^N, (z, z_*, \xi)) := \Psi\big(\bar{\mu}^N, (\bar{z}, \bar{z}_*, \xi)\big), \tag{5}$$

where $\bar{z} = G_\theta(\mu^N, z)$ and $\bar{z}_* = G_\theta(\mu^N, z_*)$. Thus composition transforms both particles and their context, while retaining the mark. Applying the Bernoulli branch to these composed coefficients

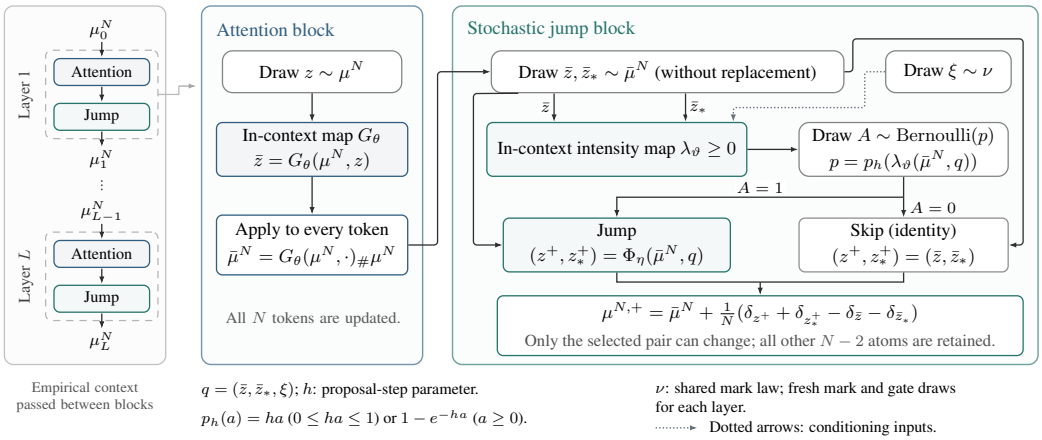

Figure 1: The *stochastic jump transformer*. (Left) Each block passes its realized empirical measure to the next. (Right) The attention map updates every token; the stochastic jump block selects two distinct particle indices uniformly and draws a fresh mark and Bernoulli gate. Acceptance replaces the selected pair with the jump-map outputs, while rejection preserves it. Many parallel jumps per block are implemented in Algorithm 1.

gives the complete layer in Figure 1. Write $\Theta = (\theta, \vartheta, \eta)$ for the full parameter tuple, with θ reserved for the deterministic stream, ϑ for the reader and rate head, and η for the jump head. We call G_θ the deterministic in-context update; when it approximates continuous motion, it is the transport or streaming step. A network of depth L repeats this stochastic block with fresh random draws and layer-dependent parameters.

1.2 Stochastic jump block: the conditional law

Given an input state $z \in E$ and context $\mu \in \mathcal{P}(E)$, the stochastic jump block produces a random output $Z^+ \in E$. Its conditional kernel is a nonlinear Markov kernel

$$K_{h,\Theta} : \mathcal{P}(E) \times E \longrightarrow \mathcal{P}(E), \qquad K_{h,\Theta}(\mu, z;\, \mathrm{d}y) = \mathrm{Law}(Z^+ \mid \mu, z)(\,\mathrm{d}y). \tag{6}$$

The in-context integral (3) defines the same coefficient maps for empirical and general contexts. As in Section 1.1, $\bar{\mu}$ is the context produced by the preceding attention block and $\bar{z}$ is the updated input state. In the *law-level block*, the input particle receives one proposal with its partner drawn from $\bar{\mu}$ and its mark from ν. Averaging over the partner, the mark, and the gate, gives the block *in law*: for every Borel set $D \subseteq E$,

$$\begin{aligned} K_{h,\Theta}(\mu, z; D) = \int_{E\times\Xi} \Big[& (1 - p_h(\lambda_\vartheta(\bar{\mu}, q)))\mathbf{1}_D(\bar{z}) \\ & + p_h(\lambda_\vartheta(\bar{\mu}, q))\mathbf{1}_D\big(\Phi_\eta^1(\bar{\mu}, q)\big)\Big] \bar{\mu}(\,\mathrm{d}\bar{z}_*)\nu(\,\mathrm{d}\xi). \end{aligned} \tag{7}$$

Here $K_{h,\Theta}(\mu, z; D)$ is the probability that Z^+ lies in D. The two terms are the conditional expectation over rejection and acceptance. Only the first component of the jump map, Φ_η^1, appears because the observed input z is taken to be the first member of the proposed pair.

One-particle output law. If the input state is drawn from the context μ, taking expectation also over that draw gives a measure on possible output states $y \in E$,

$$F_{h,\Theta}(\mu)(\,\mathrm{d}y) = \int_E K_{h,\Theta}(\mu, z;\, \mathrm{d}y)\, \mu(\,\mathrm{d}z) \in \mathcal{P}(E). \tag{8}$$

This is the output distribution of a representative particle after one law-level block. For a deterministic kernel $K(\mu, z;\, \mathrm{d}y) = \delta_{G(\mu,z)}(\,\mathrm{d}y)$, the same formula gives the pushforward $G(\mu, \cdot)_{\#}\mu$, including the empirical context update. The atom replacement in (2) produces a random empirical context.

Select an output index uniformly and independently of the block's random draws, and denote the corresponding state by $Z^{+,N}$. Then

$$\mathbb{E}[\mu^{N,+} \mid \mu^N] = \mathrm{Law}(Z^{+,N} \mid \mu^N). \tag{9}$$

The expected empirical context is therefore a one-particle marginal of the finite block. For a matching round with fixed continuous coefficients on compact E, if the input contexts converge to μ, the random output context concentrates around $F_{h,\Theta}(\mu)$ as $N \to \infty$, and its one-particle marginal converges to that same law (Appendix D.1, Proposition D.1). A single-pair layer gives a particle a proposal only with probability $2/N$, so its jump effect vanishes over one layer as N grows. Repeating these layers with physical increments $\Delta = 2h/N$, and transporting over the same increments, recovers the same continuous-depth evolution as refined matching rounds. Appendix D.1 proves the fixed-round limit and the bounded-coefficient single-pair limit as $h \downarrow 0$ and $N \to \infty$.

Joint law. For one or several composed blocks, write K_Θ for the conditional kernel. Draw an initial particle $Z_0 \sim \mu$ and its output from $K_\Theta(\mu, Z_0; \cdot)$. Their joint distribution and its output marginal are

$$\widehat{F}_\Theta(\mu)(\,\mathrm{d}z,\,\mathrm{d}y) = \mu(\,\mathrm{d}z)K_\Theta(\mu, z;\,\mathrm{d}y), \qquad F_\Theta(\mu) = (\pi_2)_\# \widehat{F}_\Theta(\mu), \tag{10}$$

where $\pi_1(z, y) = z$ and $\pi_2(z, y) = y$ are the coordinate projections. For a single block, $K_\Theta = K_{h,\Theta}$ and $F_\Theta = F_{h,\Theta}$. Retaining the input z as a *tag* lets us ask whether the architecture reproduces the response to each specified input. For example, keeping two equally weighted states in place and exchanging them produce the same output context. Their joint laws distinguish the two rules by recording which output belongs to each input. This is the information needed to characterize conditional approximation.

2 Which conditional laws can rate attention represent?

We study the uniform closure: we ask which conditional kernels the stochastic jump block(s) can approximate uniformly over contexts and input states. On compact convex state spaces, Theorem A identifies this class as the jointly continuous kernels $K : \mathcal{P}(E) \times E \to \mathcal{P}(E)$. Their joint laws $\widehat{F}(\mu)(\,\mathrm{d}z,\,\mathrm{d}y) = \mu(\,\mathrm{d}z)K(\mu, z;\,\mathrm{d}y)$ also give an intrinsic way to recognize the class: which observable properties of $\widehat{F}$ ensure that such a continuous conditional kernel exists? For this section assume E is compact, with $d_E(z, w) = |z - w|$. Write W_1 for Wasserstein distance with this cost, and use the sum metric on E^2. Continuity of $\widehat{F} : \mathcal{P}(E) \to \mathcal{P}(E^2)$ means continuity for these Wasserstein metrics. With the coordinate projections defined after (10), we call $\widehat{F}$ *ancestry preserving* if $(\pi_1)_\# \widehat{F}(\mu) = \mu$: the first coordinate records each output token's input ancestor. To read the response of one state directly from $\widehat{F}$, consider a finite weighted context $\rho = \sum_{i=1}^k a_i \delta_{z_i}$, where the z_i are distinct, $a_i > 0$, and $\sum_i a_i = 1$. Such contexts include empirical measures μ^N but also allow arbitrary positive weights. At z_i, ancestry preservation makes $\widehat{F}(\rho)(\{z_i\} \times E) = a_i$. Hence

$$K^{\mathrm{fin}}_{\widehat{F}}(\rho, z_i; D) = \frac{\widehat{F}(\rho)(\{z_i\} \times D)}{a_i}, \qquad a_i = \rho(\{z_i\}) > 0, \quad D \subset E \text{ Borel}, \tag{11}$$

is the conditional output probability for that input atom. The denominator is the scalar mass a_i. No density of ρ is required. We require these conditional laws to vary continuously with a common modulus: there is a nondecreasing $\omega : [0, \infty) \to [0, \infty)$ with $\omega(0) = 0$ and $\lim_{a \downarrow 0} \omega(a) = 0$ such that

$$W_1(K^{\mathrm{fin}}_{\widehat{F}}(\rho, z), K^{\mathrm{fin}}_{\widehat{F}}(\widetilde{\rho}, w)) \le \omega\big(W_1(\rho, \widetilde{\rho}) + d_E(z, w)\big) \tag{12}$$

for every two finite weighted contexts and their positive-mass atoms. For example, a Lipschitz conditional rule has $\omega(a) = Ca$; the theorem also allows slower moduli of continuity. The same modulus applies independently of the atom weights. The map $\widehat{F}$ itself is defined on all probability measures, including diffuse ones, and at those inputs the conditional kernel will be recovered by continuous extension. We compare two ancestry-preserving maps through their finite conditional responses,

$$d_{\mathrm{anc}}(\widehat{F}, \widehat{H}) = \sup_{\substack{\rho \text{ finitely supported} \\ z \in \mathrm{supp}\,\rho}} W_1(K^{\mathrm{fin}}_{\widehat{F}}(\rho, z), K^{\mathrm{fin}}_{\widehat{H}}(\rho, z)). \tag{13}$$

For the universality constructions below, set the deterministic update $G_\theta = \mathrm{Id}$.

Theorem A (Intrinsic characterization and two universal realizations). *Let $E \subset \mathbb{R}^d$ be nonempty, compact, and convex, and let $\widehat{F} : \mathcal{P}(E) \to \mathcal{P}(E^2)$ be continuous. Fix $h > 0$ and the linear gate $p_h(\lambda) = h\lambda$ with intensities in $[0, h^{-1}]$. In* (iii) *the outcome mark is uniform on $[0,1]$; in* (iv) *it is uniform on $\{-1,+1\}$. Marks and the auxiliary uniforms used for the gates are independent, with fresh draws in each block. The following are equivalent:*

(i) $\widehat{F}$ *preserves ancestry and satisfies* (12)*.*

(ii) There is a unique jointly continuous kernel $K : \mathcal{P}(E) \times E \to \mathcal{P}(E)$ with $\widehat{F}(\mu)(\,\mathrm{d}z,\,\mathrm{d}y) = \mu(\,\mathrm{d}z)K(\mu, z;\,\mathrm{d}y)$.

(iii) $\widehat{F}$ preserves ancestry and can be approximated arbitrarily closely in d_{anc} by one law-level block with a contextual jump map.

(iv) $\widehat{F}$ preserves ancestry and can be approximated arbitrarily closely in d_{anc} by a finite composition of law-level blocks with context-independent jump maps $\Phi_\ell(z, z_*, \xi)$, whose only context-dependent coefficients are $\lambda_\ell(\mu, z, z_*, \xi)$.

All states remain in E. The token embedding dimension, query–key dimension, and number of heads per attention layer can be bounded independently of the approximation accuracy and context size. The number of internal attention layers and the hidden widths of the tokenwise MLPs may grow; in (iv)*, the number of stochastic blocks may also grow.*

Interpretation. Theorem A is uniform over the context and input state, so one approximating network applies to every empirical context μ^N. A matching round realizes the one-block output law as $N \to \infty$ (Proposition D.1); a single-pair layer has the additional $2/N$ participation factor and the separate time scaling described in Appendix D.1. The constructions are exchange symmetric and impose no conservation constraint. For continuous representing kernels, $d_{\mathrm{anc}}(\widehat{F}, \widetilde{F}) = \sup_{\mu,z} W_1(K(\mu, z), \widetilde{K}(\mu, z))$. The class therefore contains, for instance, deterministic in-context maps and continuous mixtures of such maps.

Why conditional stability is needed. The joint law $\widehat{F}(\mu)(dz, dy) = \mu(dz)K(\mu, z; dy)$ can hide unstable behavior at atoms of vanishing mass. Condition (12) instead controls the conditional response uniformly in the query state, independently of its mass, and therefore yields a unique continuous kernel also at zero-mass query states and diffuse contexts. The regular kinetic kernels of Theorem B satisfy this condition by the stability estimate established there. Appendix B gives a simple small-atom example showing why joint-law continuity alone is insufficient.

How the approximation is realized. For (iii), first approximate $K(\mu, z)$ uniformly by a finite mixture $\sum_{a=1}^{m} p_a(\mu, z)\delta_{y_a}$. A uniform seed $u \in [0,1]$ is then used to select an output: the interval is partitioned into pieces of lengths $p_a(\mu, z)$, and the jump map sends u to the corresponding y_a. Smoothing the selector near the interval boundaries makes this map continuous while changing the law only on a set of small seed probability. Thus one stochastic block suffices, with the required accuracy carried by its attention–MLP coefficient network. For (iv), only the rates read the context. The proposal rule is fixed, but its law still depends on the context through partner sampling. Proposition B.4 shows that one such block with identity stream is not universal. Several blocks instead select successive candidates through contextual acceptance probabilities. Intermediate states retain the initial state and the selected outcome, so later rates can recover the initial context and a final context-independent jump decodes the output. For $0 < \varepsilon < \mathrm{diam}(E)$ and an $\varepsilon/4$-net of E with m points, the construction uses at most

$$2 + (m-1)\left\lceil \log_2 \frac{4m\,\mathrm{diam}(E)}{\varepsilon} \right\rceil$$

stochastic blocks (Appendix B.3).

3 PHYSICAL DEPTH: THE MCKEAN–VLASOV JUMP LIMIT

Now we study the continuous-depth class. Deep transformers whose residual updates are small produce context-dependent transport and, in the limit, Vlasov dynamics (Furuya et al., 2025a; Castin et al., 2025). We use the same depth scaling and add finite jumps whose probability is proportional to the step h. Fix a horizon $T > 0$, let L be the number of blocks, and set $h = T/L$ and $t_j = jh$. The continuous layer coordinate t is included as an input, so the coefficients may vary with depth.

The limiting coefficients are a drift $b_t(\mu, z) \in \mathbb{R}^d$, a rate $\lambda_t(\mu, z, z_*, \xi) \geq 0$, and a pair map $\Phi_t = (\Phi_t^1, \Phi_t^2) \in E^2$. For a smooth test function $\varphi : E \to \mathbb{R}$, they define the generator

$$(\mathcal{L}_{t,\mu}\varphi)(z) = b_t(\mu, z) \cdot \nabla\varphi(z) + \int_{E\times\Xi} \lambda_t(\mu, z, z_*, \xi)\big[\varphi(\Phi_t^1(\mu, z, z_*, \xi)) - \varphi(z)\big]\, \mu(\,\mathrm{d}z_*)\, \nu(\,\mathrm{d}\xi). \tag{14}$$

The corresponding McKean–Vlasov process moves with drift $b_t(\mu_t, \cdot)$ between jumps and uses λ_t and Φ_t for its jump rates and outcomes, with μ_t its own current law, so that $\frac{\mathrm{d}}{\mathrm{d}t}\langle\varphi, \mu_t\rangle = \langle\mathcal{L}_{t,\mu_t}\varphi, \mu_t\rangle$. For the population started at $\mu_s = \mu$, let $K_{s,t}(\mu, z; \cdot)$ be the law at time t of a separate tagged particle started at z at time s, with generator $\mathcal{L}_{t',\mu_{t'}}$ at each time t'. The background $\mu_{t'}$ evolves from μ and does not change with the tagged starting state z. The flow of b with the context frozen is the solution of $\dot{z} = b_t(\mu, z)$ for fixed μ, and we require these frozen flows to stay in E so that the dynamics remain E-valued before jumps.

On the network side, with time as an input, linear gates, rates in $[0, \Lambda]$, and $h\Lambda \leq 1$, the repeated blocks give

$$\mu_{j+1}^h = F_{h,\Theta_j}(t_j, \mu_j^h), \quad \mu_0^h = \mu, \qquad Y_{j+1} \mid Y_j \sim K_{h,\Theta_j}(t_j, \mu_j^h, Y_j; \cdot), \quad Y_0 = z. \tag{15}$$

The contexts μ_j^h are deterministic because F_{h,Θ_j} averages over the block randomness, while the tagged state Y_j is random; write $K_{0,t_j}^{h,\Theta}(\mu, z) = \mathrm{Law}(Y_j)$.

Theorem B (The McKean–Vlasov continuous-depth limit)**.** *Let $E \subset \mathbb{R}^d$ be nonempty, compact, and convex, and let $\Xi \subset \mathbb{R}^{d_\xi}$ be compact with fixed mark law $\nu \in \mathcal{P}(\Xi)$. Suppose b is bounded and jointly Lipschitz in (t, μ, z), with frozen flows preserving E. Suppose λ and Φ take values in $[0, \Lambda]$ and E^2, are continuous in ξ, and are uniformly Lipschitz in (t, μ, z, z_*) for $\xi \in \Xi$.*

(i) *The generator* (14) *determines a McKean–Vlasov jump process, unique in law, whose tagged kernels satisfy*

$$W_1\big(K_{s,t}(\mu, z), K_{s,t}(\widetilde{\mu}, w)\big) \leq C_T\big(W_1(\mu, \widetilde{\mu}) + |z - w|\big), \qquad 0 \leq s \leq t \leq T, \tag{16}$$

where C_T depends only on the coefficient bounds, their Lipschitz constants, E, and T.

(ii) *Let $\Theta(L) = (\Theta_j)_{j=0}^{L-1}$ be depth-L networks whose blocks have rates in $[0, \Lambda]$, deterministic updates valued in E, and pair maps valued in E^2, and whose coefficients converge to (b, λ, Φ):*

$$G_{h,\theta_j}(t_j, \mu, z) = z + h\, b_{t_j}(\mu, z) + o(h), \qquad \lambda_{\vartheta_j}(t_j, \cdot) \longrightarrow \lambda_{t_j}, \qquad \Phi_{\eta_j}(t_j, \cdot) \longrightarrow \Phi_{t_j},$$

uniformly over layers and all coefficient inputs as $L \to \infty$. Then

$$\lim_{L\to\infty} \sup_{\mu\in\mathcal{P}(E),\, z\in E} \max_{0\leq j\leq L} W_1\big(K_{0,t_j}^{h,\Theta(L)}(\mu, z), K_{0,t_j}(\mu, z)\big) = 0, \qquad h = T/L. \tag{17}$$

Part (i) is a classical well-posedness and stability result for bounded Lipschitz McKean–Vlasov jump processes (Graham & Méléard, 1997; Erny, 2022; Chaintron & Diez, 2022); we include it because (16) places the associated kernels $K_{s,t}$ in the universality class of Theorem A. Appendix C constructs the process from (14) with a bounded Poisson clock and a self-consistency argument for its law, then compares each block with this process over one step and propagates the error by stability. Appendices A and C also construct finite attention and MLP coefficient networks that satisfy the convergence and range conditions in Theorem B for every admissible triple (b, λ, Φ), so the depth limit applies to actual networks. The theorem concerns law-level blocks. For bounded Lipschitz target coefficients and exchange-symmetric approximants, if the initial empirical laws converge, Appendix D proves that the finite-particle matching algorithm converges to the same population law μ_t, uniformly on $[0, T]$ in probability and in mean, as $N \to \infty$, $h \to 0$, and the coefficient errors vanish. Each matching round represents duration h; a single-pair layer represents $\Delta = 2h/N$, with the deterministic update over the same Δ (Appendix D.1).

The converse. Theorem B starts from a local rule and obtains its finite-time kernels. Conversely, when does a given family $\{K_{s,t}\}_{s\le t}$ come from one such rule? To continue from an intermediate time, define the evolved context $S_{s,t}\mu := \int_E K_{s,t}(\mu, z; \cdot)\,\mu(\mathrm{d}z)$. Direct evolution must agree with evolution through any intermediate time a:

$$K_{s,t}(\mu, z; \cdot) = \int_E K_{a,t}(S_{s,a}\mu, y; \cdot)\, K_{s,a}(\mu, z; \mathrm{d}y), \qquad s \le a \le t, \qquad K_{t,t}(\mu, z) = \delta_z. \quad (18)$$

We also require the short-time change of the kernels to separate into continuous motion and finite-rate jumps.

Theorem B' (From conditional maps to a local jump mechanism)**.** *Let E be as in Theorem B, and let $K_{s,t}(\mu, z)$ be jointly continuous in (s, t, μ, z) and satisfy* (18)*. Suppose there is a bounded, jointly Lipschitz field $b_t(\mu, z)$ with E-preserving frozen flows such that, for every test function φ that is the restriction to E of a C^1_b function on $\mathbb{R}^d$,*

$$B_{t,\mu}\varphi(z) = \lim_{h\downarrow 0} \frac{\langle \varphi, K_{t,t+h}(\mu, z)\rangle - \varphi(z)}{h} - b_t(\mu, z)\cdot\nabla\varphi(z) \quad (19)$$

exists uniformly over (t, μ, z) and extends linearly to $C(E)$. Assume that, for some $\Lambda > 0$,

$$B_{t,\mu}1(z) = 0, \qquad B_{t,\mu}\varphi(z) \ge -\Lambda\varphi(z) \qquad (\varphi \in C(E),\ \varphi \ge 0), \quad (20)$$

and that, for some $C < \infty$ and every Lipschitz $\varphi : E \to \mathbb{R}$,

$$\big|B_{t,\mu}\varphi(z) - B_{s,\widetilde{\mu}}\varphi(w) + \Lambda[\varphi(z) - \varphi(w)]\big| \le C\,\mathrm{Lip}(\varphi)\big(|t-s| + W_1(\mu, \widetilde{\mu}) + |z-w|\big). \quad (21)$$

Then the given kernels are exactly those of the unique finite-rate McKean–Vlasov process with generator $\mathcal{L}_{t,\mu}\varphi(z) = b_t(\mu, z)\cdot\nabla\varphi(z) + B_{t,\mu}\varphi(z)$. Conversely, the limiting kernels of Theorem B satisfy these conditions.

Here b describes the motion between jumps, while B accounts for the jumps' contribution to short-time change; the derivative of the mean displacement can include both terms. Condition (20) makes $\varphi \mapsto B_{t,\mu}\varphi(z) + \Lambda\varphi(z)$ a positive functional of mass Λ, so the jumps occur at rate Λ with a jump law that may leave the state unchanged. Condition (21) controls its variation. The theorem recovers B, not a unique rate and jump map: a zero rate and a positive rate whose jump map is the identity generate the same process. For example, a reset at rate Λ to a fixed $\nu_0 \in \mathcal{P}(E)$ gives $K_{s,t}(\mu, z) = e^{-\Lambda(t-s)}\delta_z + (1 - e^{-\Lambda(t-s)})\nu_0$ and satisfies these conditions.

4 Physical structure and the Maxwell–Boltzmann realization

Rate attention selects a pair outcome rather than averaging possible outcomes. Physical constraints can therefore be imposed directly on the jump.

Conservation, reversibility, and entropy. On $E = \mathbb{T}^{d_v} \times \mathbb{R}^{d_v}$, with unit spatial volume, write $z = (x, v)$, $z_* = (y, w)$, $c = (v+w)/2$, $g = (v-w)/2$, and $I = (c, |g|)$. A learned skew-symmetric matrix $\Omega_\eta(I, \xi) \in \mathbb{R}^{d_v \times d_v}$ defines $O_\eta(I, \xi) = \exp \Omega_\eta(I, \xi)$ and the elastic pair map

$$\Phi_\eta(z, z_*, \xi) = \big((x, c + O_\eta(I, \xi)g), (y, c - O_\eta(I, \xi)g)\big). \quad (22)$$

Every jump conserves momentum and energy. Because the rotation depends only on I and the mark, it fixes c and $|g|$ and preserves pair volume.

Assume the mark law ν is invariant under $\xi \mapsto -\xi$ and $\Omega_\eta(I, -\xi) = -\Omega_\eta(I, \xi)$, so the negative mark reverses the jump. Suppose first that a nonnegative learned rate $\widetilde{\lambda}$ has been symmetrized under exchange of the two particles. For a probability density f, set $\mu_f(\mathrm{d}z) = f(z)\,\mathrm{d}z$ and $(z', z'_*) = \Phi_\eta(z, z_*, \xi)$, and define

$$\lambda[f](z, z_*, \xi) = \tfrac{1}{2}\big[\widetilde{\lambda}(\mu_f, z, z_*, \xi) + \widetilde{\lambda}(\mu_f, z', z'_*, -\xi)\big]. \quad (23)$$

Forward and reverse collisions then have the same rate. With free flight $b(x, v) = (v, 0)$, these symmetries give the Boltzmann H-theorem $\mathrm{d}H(f_t \mid M)/\mathrm{d}t \le 0$, where $H(f \mid M) = \int_E f \log(f/M)\,\mathrm{d}z$ and $M(v)$ is a unit-mass Maxwellian. This requires the regularity and integrability assumptions of Proposition H.2. Appendix H.3 gives the entropy identity; Appendix H gives the separate finite-particle balance conditions.

A spatially nonhomogeneous Maxwell–Kac model. The Maxwell model gives a concrete realization of the architecture. Kac-type dynamics sample a collision partner and scattering parameter and replace the selected pair by its post-collisional state (Kac, 1956; Graham & Méléard, 1997). Adding free transport gives the three components of our block: transport between collisions, a collision rate, and a pair jump. Take $E = \mathbb{T}^3 \times \mathbb{R}^3$, with squared product cost $d_E((x,v),(y,w))^2 = d_{\mathbb{T}^3}(x,y)^2 + |v-w|^2$ for W_2. Free transport over a step h sends (x,v) to $(x+hv,v)$. For isotropic cutoff Maxwell molecules, the collision rate is independent of relative speed (Tanaka, 1978). Let ν be the uniform probability law on $SO(3)$. A mark $R \sim \nu$ gives the elastic collision

$$\Phi_R((x,v),(y,w)) = \big((x, c+Rg),(y, c-Rg)\big), \qquad c = \frac{v+w}{2}, \quad g = \frac{v-w}{2}. \tag{24}$$

Fix an even spatial interaction kernel satisfying

$$\beta \in C^1(\mathbb{T}^3), \qquad 0 < \beta_{\min} \le \beta \le \beta_{\max}, \qquad \|\nabla\beta\|_\infty \le L_\beta. \tag{25}$$

The corresponding choice of coefficients in (14) is

$$b(x,v) = (v,0), \qquad \lambda^{r,\beta}((x,v),(y,w)) = r\beta(x-y), \qquad \Phi = \Phi_R,$$

where $r \ge 0$ sets the collision frequency. Hence, for the current population law $f \in \mathcal{P}_2(E)$ and $\varphi \in C_c^1(E)$,

$$\begin{aligned}(\mathcal{L}_f^{r,\beta}\varphi)(x,v) &= v \cdot \nabla_x \varphi(x,v) \\ &\quad + r \int_E \beta(x-y) \int_{SO(3)} \big[\varphi(x, c+Rg) - \varphi(x,v)\big]\, \nu(\mathrm{d}R)\, f(\mathrm{d}y, \mathrm{d}w).\end{aligned} \tag{26}$$

Thus the Maxwell generator is exactly of the transport–jump form: b gives free transport, $r\beta(x-y)$ gives the collision rate, and Φ_R gives the elastic collision outcome. The evolution is nonlinear because each collision partner is sampled from the evolving population law. The same coefficients give a direct realization by the stochastic jump block. Over one step h, first transport each state by $(x,v) \mapsto (x+hv,v)$. Then draw a partner from the transported population and a fresh collision mark R, and accept the update Φ_R with probability $1 - e^{-hr\beta(\bar{x}-\bar{y})}$, where $\bar{x}, \bar{y}$ are the transported positions; otherwise retain the transported pair. Thus the Maxwell–Kac scheme is precisely the law-level stochastic jump block with this choice of transport, rate, and jump map.

Write $K_t^{r,\beta}(f,z)$ for the tagged law and $S_t^{r,\beta} f = \int_E K_t^{r,\beta}(f,z)\, f(\mathrm{d}z)$ for the population law. Although the physical phase space is noncompact, Appendix I shows that these kernels have the same joint continuity in context and input state that characterizes the class in Theorem A. It also shows that the Maxwell blocks converge to the tagged and population laws as $h \downarrow 0$, uniformly for initial laws with fixed velocity mean and energy and a common fourth-moment bound, with r and the tagged velocity bounded, for each fixed $t \ge 0$. The same conclusion holds when the collision rate is approximated uniformly by an admissible neural rate. Thus the Maxwell model provides a noncompact physical realization of both the conditional-kernel class and the continuous-depth limit developed above.

5 Perspective: reusable kinetic building blocks

The architecture is not tied to one kinetic equation. Its basic ingredients (transport, stochastic interaction rates, and jump maps) can be changed to represent different kinetic laws, while structural constraints can enforce conservation, reversibility, or equilibrium. This makes the transport–jump block a natural reusable component for physics models: the same architecture can represent different microscopic interaction mechanisms rather than being rebuilt for each equation.

The same viewpoint extends beyond kinetic scale. A macroscopic field can be lifted to a kinetic distribution, evolved by transport and interactions, and recovered through its moments. Boltzmann, BGK, and lattice kinetic models recover fluid equations under suitable scaling and regularity assumptions (Bardos et al., 1991; Golse and Saint-Raymond, 2004; Saint-Raymond, 2003; Junk and Yong, 2003). This suggests a path from learned microscopic interaction laws to macroscopic PDE models within one architecture. Making this rigorous for a given fluid system requires a compatible kinetic lift, control of the neural approximation uniformly in the scaling limit, and convergence to the target macroscopic equation. For the spatially mollified Maxwell model considered here, a local fluid limit would additionally require removal of the spatial mollifier. Appendix J outlines these steps.

REPRODUCIBILITY STATEMENT

The exact architecture, gate conventions, coefficient classes, and proofs are given in Appendices A–J. The finite algorithm and all random draws are specified in Algorithm 1 and Appendix D; the optional equilibrium correction is proved in Appendix H. The source includes the editable vector architecture figure. This is a theoretical study; it reports neither a trained simulator nor benchmark experiments.

AI USE STATEMENT

OpenAI ChatGPT and the AI-assisted source-development workflow assisted with mathematical formulation and proof development, literature searches, drafting and revision, LaTeX preparation, and figure-code preparation. AI-generated critiques were used in iterative revisions; they are not independent human or formal verification. No empirical results are generated or reported in this work. The authors retain responsibility for validating the final claims, proofs, and citations.

REFERENCES

Luisa Andreis, Paolo Dai Pra, and Markus Fischer. McKean–Vlasov limit for interacting systems with simultaneous jumps. *Stochastic Analysis and Applications*, 36(6):960–995, 2018. doi:10.1080/07362994.2018.1486202.

Hans Babovsky and Reinhard Illner. A convergence proof for Nanbu's simulation method for the full Boltzmann equation. *SIAM Journal on Numerical Analysis*, 26(1):45–65, 1989.

Claude Bardos, François Golse, and C. David Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *Journal of Statistical Physics*, 63(1–2):323–344, 1991. doi: 10.1007/BF01026608.

G. A. Bird. Approach to translational equilibrium in a rigid sphere gas. *The Physics of Fluids*, 6(10):1518–1519, 1963. doi:10.1063/1.1710976.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv:2501.18322*, 2025.

Giulia Cavagnari, Giuseppe Savaré, and Giacomo Enrico Sodini. Extension of monotone operators and Lipschitz maps invariant for a group of isometries. *Canadian Journal of Mathematics*, 77(1):149–186, 2025. doi: 10.4153/S0008414X23000846.

Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. I. Models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022. doi:10.3934/krm.2022017.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 1556–1565, 2019.

Xavier Erny. Well-posedness and propagation of chaos for McKean–Vlasov equations with jumps and locally Lipschitz coefficients. *Stochastic Processes and their Applications*, 150:192–214, 2022.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

Lev Fedorov, Michaël E. Sander, Romuald Elie, Pierre Marion, and Mathieu Laurière. Clustering in deep stochastic transformers. *arXiv:2601.21942*, 2026. `https://arxiv.org/abs/2601.21942`.

Martin Friesen, Barbara Rüdiger, and Padmanabhan Sundar. The Enskog process for hard and soft potentials. *Nonlinear Differential Equations and Applications*, 26:20, 2019. doi:10.1007/s00030-019-0566-6.

Martin Friesen, Barbara Rüdiger, and Padmanabhan Sundar. On uniqueness and stability for the Boltzmann–Enskog equation. *Nonlinear Differential Equations and Applications*, 29:25, 2022. doi:10.1007/s00030-022-00755-6.

Takashi Furuya, Maarten V. de Hoop, and Matti Lassas. Transformers through the lens of support-preserving maps between measures. *arXiv:2509.25611*, 2025a.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, 2025b.

François Golse and Laure Saint-Raymond. The Navier–Stokes limit of the Boltzmann equation for bounded collision kernels. *Inventiones Mathematicae*, 155(1):81–161, 2004. doi: 10.1007/s00222-003-0316-5.

Carl Graham and Sylvie Méléard. Stochastic particle approximations for generalized Boltzmann models and convergence estimates. *Annals of Probability*, 25(1):115–132, 1997. doi:10.1214/aop/1024404281.

Daniel Heydecker. Kac's process with hard potentials and a moderate angular singularity. *Archive for Rational Mechanics and Analysis*, 244:699–759, 2022. doi:10.1007/s00205-022-01767-3.

Junteng Jia and Austin R. Benson. Neural jump stochastic differential equations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shi Jin, Lei Li, and Jian-Guo Liu. Random batch methods (RBM) for interacting particle systems. *Journal of Computational Physics*, 400:108877, 2020. doi:10.1016/j.jcp.2019.108877.

Michael Junk and Wen-An Yong. Rigorous Navier–Stokes limit of the lattice Boltzmann equation. *Asymptotic Analysis*, 35(2):165–185, 2003. doi: 10.3233/ASY-2003-563.

Mark Kac. Foundations of kinetic theory. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pp. 171–197. University of California Press, 1956.

Anastasis Kratsios. Universal regular conditional distributions via probabilistic transformers. *Constructive Approximation*, 57(3):1145–1212, 2023. doi:10.1007/s00365-023-09635-3.

J. Antonio Lara Benitez, Kareem Hegazy, Junyi Guo, Ivan Dokmanić, Michael W. Mahoney, and Maarten V. de Hoop. Neural equilibria for long-term prediction of nonlinear conservation laws. *arXiv:2501.06933*, 2025. Revised May 2026. https://arxiv.org/abs/2501.06933v3.

Kyle Lee, Corentin Delacour, Kevin Callahan-Coray, Kyle Jiang, Can Yaras, Samet Oymak, Tathagata Srimani, and Kerem Y. Camsari. Stochastic sparse attention for memory-bound inference. In *International Conference on Machine Learning*, 2026. https://arxiv.org/abs/2605.01910v2.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Kenichi Nanbu. Direct simulation scheme derived from the Boltzmann equation. I. Monocomponent gases. *Journal of the Physical Society of Japan*, 49(5):2042–2049, 1980. doi:10.1143/JPSJ.49.2042.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Laure Saint-Raymond. From the BGK model to the Navier–Stokes equations. *Annales scientifiques de l'École Normale Supérieure*, 36(2):271–317, 2003. doi: 10.1016/S0012-9593(03)00010-7.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 8459–8468, 2020.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *PMLR*, pp. 3515–3530, 2022. https://proceedings.mlr.press/v151/sander22a.html.

Alain-Sol Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX–1989*, volume 1464 of *Lecture Notes in Mathematics*, pp. 165–251. Springer, 1991.

Hiroshi Tanaka. Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 46:67–105, 1978.

Luke Tierney. A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998. doi: 10.1214/aoap/1027961031.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhanpeng Zeng, Yunyang Xiong, Sathya Ravi, Shailesh Acharya, Glenn M. Fung, and Vikas Singh. You only sample (almost) once: Linear cost self-attention via Bernoulli sampling. In *International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 12321–12332, 2021. https://proceedings.mlr.press/v139/zeng21a.html.

A NEURAL COEFFICIENTS WITH FIXED TOKEN EMBEDDING DIMENSION

We first specify the neural operations used by the theorems. This also separates two depths: internal attention layers compute the contextual coefficient functions, while stochastic blocks apply deterministic attention and jump updates. The new rate-only construction in Theorem A(iv) uses increasing stochastic-block depth. Under the small-time scaling of Section 3, block depth also resolves physical time. Increasing coefficient-network depth need not increase the number of particle-update steps. Physical time uses the schedule convention in Appendix D.1.

A.1 THE COEFFICIENT MODULE AND ITS ADMITTED OPERATIONS

On a compact input domain, a coefficient module consists of finite compositions of exponential query–key attention, residual updates, and shared tokenwise ReLU networks. Internal context tokens are feature vectors, not extra physical particles. A separate query token reads these feature contexts and undergoes the same in-context composition, but contributes no mass to them. A finite bank of queries is evaluated independently against that same sequence of feature contexts. Biases are implemented with a constant coordinate. Unnormalized heads integrate against a probability measure, so their empirical version always includes $1/N$.

Finite affine lifts, coordinate projections, clipping to an interval, nearest-point projection onto an explicitly specified closed convex output set, symmetry averaging, and random sampling are admitted operations. The probability laws of the marks are specified separately from the networks. The one-block sampler in Theorem A(iii) has a scalar uniform outcome seed independent of the Bernoulli acceptance draw. When a coupling proof uses a gate uniform, it means an independent $U \sim \mathrm{Unif}[0,1]$ implementing $A = \mathbf{1}_{\{U \le p_n(\lambda)\}}$; it is a representation of that same Bernoulli variable. All its internal complexity is in a finite coefficient network. For a prescribed physical jump map, evaluation of that map is an exact operation. None of these conventions asserts that its evaluation or training has an accuracy-independent cost.

Lemma A.1 (External-query form of transformer universality)**.** *Let $E \subset \mathbb{R}^d$ and $Q \subset \mathbb{R}^r$ be nonempty compact sets, and let $F : \mathcal{P}(E) \times Q \to \mathbb{R}^m$ be continuous for W_1 and Euclidean distance. For every $\varepsilon > 0$, a normalized transformer in-context module uniformly approximates F. Its token embedding dimension and head count can be bounded in terms of d, r, m alone, independently of ε and the number of context tokens. The query is not an additional atom of the context.*

Proof. Theorem 1 of Furuya et al. (2025b) applies to continuous maps on $\mathcal{P}(\Omega) \times \Omega$ for a compact Euclidean Ω, with embedding dimension at most $\dim(\Omega_{\mathrm{ambient}}) + 3m$ and at most m heads. We reduce our two input types to that setting. Choose $e_0 \in E$, $q_0 \in Q$, and the disjoint compact embeddings

$$c(y) = (y, 0, 0), \qquad u(q) = (0, q, 1), \qquad \Omega = c(E) \cup u(Q) \subset \mathbb{R}^{d+r+1}.$$

Define continuous maps $p : \Omega \to E$ and $a : \Omega \to Q$ by $p(c(y)) = y$, $p(u(q)) = e_0$, $a(c(y)) = q_0$, and $a(u(q)) = q$. The two branches are separated by their last coordinate. Then $\widetilde{F}(\rho, x) = F(p_{\#}\rho, a(x))$ is continuous on the entire $\mathcal{P}(\Omega) \times \Omega$. Apply the cited theorem and evaluate its in-context map at $(\rho, x) = (c_{\#}\mu, u(q))$. This gives $F(\mu, q)$ with error ε. In an in-context composition, the context evolves from its own pushforwards while the external query evolves by evaluation of the same maps. Thus $u(q)$ has never been added to ρ. The dimension bound is $d + r + 1 + 3m$, and the head bound is m. □

Lemma A.2 (Unnormalized implementation of a normalized head)**.** *On a compact set of context and query features, a normalized exponential attention head can be uniformly approximated by unnormalized exponential heads and a tokenwise ReLU readout. A fixed number of additional token coordinates suffices. Consequently the module in Lemma A.1 can be implemented with unnormalized primitives with embedding dimension and head count bounded solely in terms of d, r, m.*

Proof. Write a normalized head as $n(q, \rho)/d(q, \rho)$, where

$$n(q, \rho) = \int e^{s(q,y)} V(y)\, \rho(\,\mathrm{d}y), \qquad d(q, \rho) = \int e^{s(q,y)}\, \rho(\,\mathrm{d}y).$$

One unnormalized head with augmented values $(V, 1)$ computes both quantities. If $|s| \le M$ on the feature domain, then $e^{-M} \le d \le e^{M}$ for every probability context. The map $(n, d) \mapsto n/d$ is continuous on the resulting compact range; ReLU networks approximate it uniformly (Leshno et al., 1993). Store (n, d) in scratch coordinates, retain the incoming features for the residual update, and clear the scratch coordinates afterward. Multiple heads use separate scratch coordinates before their linear combination.

For completeness this simulation remains uniform through a deep coefficient network. Fix the finite normalized network supplied by Lemma A.1. At each internal layer its exact reachable features, over all contexts and external queries, lie in a compact set. Enlarge those sets by a fixed small neighborhood. Scores and all finite network maps are uniformly continuous there; the denominator retains a positive lower bound. The pushforward of a common input law by feature maps at sup distance η is at Wasserstein distance at most η. Inductively choose the quotient-approximation errors so that the cumulative context and query feature errors stay inside the enlarged sets and the final output error is below ε. Only finitely many internal layers are involved for this chosen accuracy. Scratch storage is reused, so it depends on the fixed per-layer embedding/head bounds, not on the number of layers or tokens. This proves the assertion. It does not preserve the sharp embedding constant of the normalized theorem, and it places no bound on the number of ReLU parameters required for a given accuracy. □

Proposition A.3 (Continuous pair coefficients)**.** *On compact state, time, and mark domains with fixed finite-dimensional embeddings, every continuous vector of pair coefficients is uniformly approximable by the preceding modules, with token embedding dimension and head count independent of accuracy and context size. Prescribed interval bounds, exchange symmetry, and closed convex output constraints can be preserved exactly. When imposed together, the output constraint is assumed invariant under exchange of output coordinates (as is E^2).*

Proof. Collect the finite-dimensional inputs as $q = (t, z, z_*, \xi)$ and apply the two lemmas. Clipping is nonexpansive and fixes the target interval. Projection onto a closed convex output set is nonexpansive and fixes the target outputs. Its exchange invariance ensures that symmetry averaging below also remains in that set. If σ denotes pair exchange (including an admitted measure-preserving exchange of mark coordinates) and τ exchanges the output coordinates, replace an approximate pair head H by $(H + \tau H \circ \sigma)/2$. The target is fixed by this operation, so uniform error does not increase. For scalar rates use the analogous average without τ. The parallel evaluation of two heads increases embedding storage by a fixed factor only. These operations enforce exchange and range, not conservation on a nonconvex collision shell. □

The shared-read implementation in (4) is included in this coefficient class. Apply Proposition A.3 to the vector $(\lambda, \Phi^1, \Phi^2)$, then split its final readout into the scalar rate and pair-output heads. The context features or internal registers can be shared between those heads. The finite output dimension remains fixed. The coefficient network may use more internal attention and MLP layers as the requested error decreases. This leaves the number of physical stochastic blocks unchanged.

Where the rate reads the empirical context. For a head $C(\mu, q) = \int s(q, y)V(y)\mu(\,\mathrm{d}y)$, assume $|s| \le B_s$, $|V| \le B_V$, and Lipschitz constants L_s for the score in (q, y) and L_V for the value map. Optimal coupling gives

$$|C(\mu, q) - C(\eta, \widetilde{q})| \le B_V L_s |q - \widetilde{q}| + (B_V L_s + B_s L_V) W_1(\mu, \eta). \tag{27}$$

Indeed, insert $s(\widetilde{q}, y)V(y)$ and integrate the remaining difference in the paired context variables. The constants are independent of N. Finite compositions retain a corresponding bound on their stated domains. No new attention approximation is introduced by replacing an integral with its empirical sum: they are the same function evaluated at different measures. In a varying-network limit we instead use

$$|\lambda_m(\mu^N, q) - \lambda(\mu, q)| \le |\lambda_m(\mu^N, q) - \lambda(\mu^N, q)| + |\lambda(\mu^N, q) - \lambda(\mu, q)|.$$

Only the target needs a common continuity or Lipschitz bound.

A.2 Coefficients on a noncompact context domain

Raw exponential scores of unbounded linear keys need not be integrable under an arbitrary finite-moment context. We do not define the physical coefficient modules that way. Let $\pi_R(x, v) = (x, \Pi_{\overline{B}_R} v)$ and apply attention to $(\pi_R)_{\#}\mu$, with periodic coordinates embedded in a fixed-dimensional Euclidean space. For $p > 1$,

$$W_1(\mu, (\pi_R)_{\#}\mu) \le \int_{|v|>R} |v|\, \mu(\,\mathrm{d}x,\, \mathrm{d}v) \le R^{1-p} \int |v|^p\, \mu(\,\mathrm{d}x,\, \mathrm{d}v). \tag{28}$$

Thus token clipping is an internal coefficient approximation, not a truncation of the particle state or the physical collision amplitude.

If a target multiplier a is W_1-Lipschitz and continuous in all finite-dimensional inputs, first fix a moment bound and bounded query range. Equation (28) makes replacing μ by $(\pi_R)_{\#}\mu$ uniformly accurate there. Apply Proposition A.3 on that compact token domain, then clip to $[0, A]$ and symmetrize. Increasing R and the query ranges gives a sequence converging uniformly on every bounded-state, bounded-p-moment set. All evaluated scores are finite. The token embedding dimension remains fixed while weights and internal depth may change with R and accuracy. Angular-mark continuity is needed for this existence statement. The limit theorem for an already supplied sequence of bounded approximants does not require a separate proof of their neural existence.

B The intrinsic static class and its realization

Theorem A joins a representation statement with an approximation statement. We prove those directions separately so that the geometry needed only by the neural construction is visible. Throughout this section E is nonempty compact, with metric $d = d_E$ and diameter D. Probability laws carry W_1, and joint input–output laws carry the Wasserstein metric for $d_{E^2}((z, y), (w, \widetilde{y})) = d_E(z, w) + d_E(y, \widetilde{y})$. The finite contexts in (11) have arbitrary positive weights; their normalized fibers are divided by the scalar atom mass. The block used for static realization is the law-level conditional rule (7). The finite single-pair layer has the additional selection probability recorded in Appendix D.1.

Theorem B.1 (Normalized fibers determine a continuous kernel)**.** *For a continuous $\widehat{F} : \mathcal{P}(E) \to \mathcal{P}(E^2)$, ancestry preservation and* (12) *are equivalent to a unique jointly continuous $K : \mathcal{P}(E) \times E \to \mathcal{P}(E)$ representing $\widehat{F}(\mu) = \mu(\,\mathrm{d}z)K(\mu, z;\, \mathrm{d}y)$.*

Proof. Let $\mathcal{I}$ be the finite-incidence set of (μ, z) with μ finitely supported and $\mu(\{z\}) > 0$. It is dense in $\mathcal{P}(E) \times E$: given μ, z, approximate μ by finite η_n and use $(1 - a_n)\eta_n + a_n\delta_z$ for $a_n \downarrow 0$. A representing continuous kernel agrees with the normalized fibers on $\mathcal{I}$. Compactness implies uniform continuity, hence the required common modulus.

Conversely the common modulus makes the fibers uniformly continuous on $\mathcal{I}$. For any sequence in $\mathcal{I}$ tending to (μ, z), the fibers are Cauchy in the complete space $(\mathcal{P}(E), W_1)$ and have a limit independent of the sequence. This defines a continuous K on all of $\mathcal{P}(E) \times E$ with that modulus. At finite contexts, ancestry preservation and summation over their atoms give $\widehat{F}(\mu) = \mu K(\mu)$. To pass to arbitrary contexts, fix $g \in C(E^2)$. The map

$$(\mu, z) \longmapsto \int g(z, y)K(\mu, z;\, \mathrm{d}y)$$

is jointly continuous: split its variation into the uniform variation of g in z and weak continuity of K. It is uniformly continuous on the compact domain. Integration against varying μ is therefore continuous. Approximate an arbitrary μ by finite contexts and pass the identity using continuity of both sides. Two continuous representing kernels agree on $\mathcal{I}$, so density gives uniqueness everywhere, including at a zero-mass ancestor. □

Why small atom weights matter. On $E = [0, 1]$, let $\rho_a = (1 - a)\delta_0 + a\delta_1$ and $s(a) = (1 + \sin(1/a))/2$ for $a > 0$. The ancestral joint laws $(1 - a)\delta_{(0,0)} + a\delta_{(1,s(a))}$ converge to $\delta_{(0,0)}$ as $a \downarrow 0$, whereas their normalized response at the atom 1 is $\delta_{s(a)}$ and has no limit. Thus joint-law convergence can hide conditional instability at an atom whose mass vanishes; condition (12) excludes this behavior.

Lemma B.2 (A continuous approximate sampler with one uniform seed)**.** *Suppose $E \subset \mathbb{R}^d$ is also convex and X is compact. Every continuous $K : X \to \mathcal{P}(E)$ admits continuous $H : X \times [0,1] \to E$ with $\sup_x W_1(H(x,\cdot)_{\#}\mathrm{Leb}, K(x))$ arbitrarily small.*

Proof. Choose an η-net $y_1, \dots, y_m$ in E and a continuous partition of unity $\rho_1, \dots, \rho_m$ subordinate to the η-balls. Put $p_i(x) = \int \rho_i \,\mathrm{d}K(x)$ and $c_j(x) = \sum_{i \le j} p_i(x)$. Transporting each y to y_i with probability $\rho_i(y)$ shows that $K_\eta(x) = \sum_i p_i(x)\delta_{y_i}$ is within η of $K(x)$. Let s_δ be continuous, nondecreasing, zero on $(-\infty, -\delta]$, and one on $[\delta, \infty)$. Define

$$H(x,u) = y_1 + \sum_{j<m} s_\delta(u - c_j(x))(y_{j+1} - y_j).$$

Since c_j is nondecreasing, the coefficients $s_\delta(u - c_j)$ are nonincreasing in j. Telescoping writes H as a convex combination of the y_i, so $H \in E$. Away from the $m-1$ intervals $|u - c_j(x)| < \delta$, this is the usual inverse-CDF sampler of $K_\eta(x)$. Their total Lebesgue measure is at most $2(m-1)\delta$. Coupling by the same uniform gives error at most $2D(m-1)\delta$. First choose η, then δ. □

Proof of Theorem A, equivalences (i)–(iii). Apply Theorem B.1. For a representing K, use Lemma B.2 with $X = \mathcal{P}(E) \times E$. Proposition A.3 approximates the continuous sampler $H(\mu, z, u)$, projecting its output into E. A common uniform seed couples the sampled laws, so the additional W_1 error is no larger than the uniform sampler error. Use the identity stream and the constant intensity h^{-1} in the linear gate. Every proposal is accepted, and its first output is $H_\eta(\mu, z, u)$. The proposed partner may be ignored. The symmetric pair rule $(H_\eta(\mu, z, u), H_\eta(\mu, z_*, u))$ realizes both marginals; no conservation is imposed. Its output dimension and seed dimension are fixed, so the embedding bound from the coefficient theorem is independent of sampler accuracy and of m.

For a prescribed homeomorphism $T_h : E \to E$ independent of μ, apply the same construction to $\widetilde{K}(\rho, x) = K((T_h^{-1})_{\#}\rho, T_h^{-1}x)$. Streaming followed by that sampler approximates the original kernel. For continuous representing kernels, the finite-incidence set is dense, so the supremum in d_{anc} equals the supremum on all of $\mathcal{P}(E) \times E$. Finally every finite block built from the continuous coefficients above has a continuous conditional kernel. For a test $\varphi \in C(E)$, integrate the continuous gate mixture against the partner and fixed mark laws; dominated convergence gives the assertion. Composition of finitely many law-level blocks preserves it. Suppose now a continuous ancestry-preserving tagged map is approximated in d_{anc}. Its finite fibers are a uniform limit, on the finite-incidence set, of uniformly continuous kernel restrictions. A uniform limit of uniformly continuous maps is uniformly continuous: for a given error first fix a uniformly close approximant, then use its modulus. Thus those fibers satisfy (12); Theorem B.1 supplies the representing kernel. This proves the reverse closure assertion without assuming a continuous representative in advance. □

B.1 Deterministic characteristics and the role of streaming

For $K(\mu, z) = \delta_{G(\mu,z)}$, $W_1(\delta_y, \delta_{\widetilde{y}}) = d(y, \widetilde{y})$, so the intrinsic condition reduces exactly to continuity of G. It does not require full Fréchet differentiability. One directional interpretation is available on an immutable-label extension. For a finitely supported labeled context, insert mass a at a fresh label and state z, rescale the old context by $1 - a$, and test only descendants of the fresh label. The change is exactly $a \int \varphi \,\mathrm{d}K((1-a)\mu + a\delta_z, z)$. Dividing by a and taking $a \downarrow 0$ yields $\int \varphi \,\mathrm{d}K(\mu, z)$, which becomes $\varphi(G(\mu, z))$ in the deterministic case. This is a labeled directional variation, not an unqualified derivative theorem for a density-restricted solution operator.

For a learned contextual stream the information supplied to the jump head is

$$(\mu, z) \longmapsto \big(S_h\mu, T_h(\mu, z)\big), \qquad S_h\mu = T_h(\mu, \cdot)_{\#}\mu.$$

Invertibility of $T_h(\mu, \cdot)$ separately for each μ is not sufficient to invert this joint map. On $\mathbb{R}^d$, $T(\mu, z) = z - \int y\mu(\,\mathrm{d}y)$ is a translation for each μ, yet all (δ_a, a) yield the same streamed pair $(\delta_0, 0)$. The identity member used in Theorem A avoids this obstruction. There is also a sufficient small-residual condition. If b is L_z-Lipschitz in state and L_μ-Lipschitz in law and $T_h(\mu, z) = z + hb(\mu, z)$, then for $h(L_z + L_\mu) < 1$,

$$[1 - h(L_z + L_\mu)]W_1(\mu, \eta) \le W_1(S_h\mu, S_h\eta). \tag{29}$$

To see this, use the $(1-hL_z)$ lower Lipschitz bound on the same-background map $T_h(\mu,\cdot)$ and bound the change from $T_h(\mu,\cdot)_\#\eta$ to $T_h(\eta,\cdot)_\#\eta$ by $hL_\mu W_1(\mu,\eta)$. Thus the context and then the tagged state are recoverable on the image. On compact domains the inverse is continuous. A projecting boundary correction need not preserve this injectivity; it is used for dynamic consistency below, not silently for this static claim.

B.2 A SCATTERING LAW BEYOND DETERMINISTIC STATE PUSHFORWARDS

The following elementary consequence of deterministic pointwise representation supplies the finite-atomic example in the introduction; compare the support-preserving formulation of Furuya et al. (2025a) and the pointwise representations of Cavagnari et al. (2025). It concerns state-pushforward representations without added random seeds or particle labels.

Proposition B.3 (Non-splitting and a two-beam scattering example)**.** *For a metric state space E, the probability measures supported on at most k points are closed under weak convergence. Consequently, if μ has at most k atoms and $\eta \in \mathcal{P}(E)$ has a finite first moment and more than k support points, then*

$$\inf_G W_1(G_\#\mu,\eta) > 0,$$

where the infimum ranges over deterministic maps on $\operatorname{supp}\mu$*. In particular, let $d \geq 2$, $r, h > 0$, and $\mu = (\delta_{re_1}+\delta_{-re_1})/2$ on velocity space. A single law-level hard-sphere reflection step, using a uniform partner from μ, uniform direction $n \in \mathbb{S}^{d-1}$ and acceptance probability $1-\exp[-h|(v-w)\cdot n|]$, has an output law with infinitely many support points. It cannot be approximated arbitrarily closely in W_1 by $G_\#\mu$.*

Proof. If a weak limit had $k+1$ distinct support points, select disjoint open neighborhoods of those points, each of positive limit mass. Portmanteau implies that every sufficiently late measure charges all $k+1$ neighborhoods, contradicting its support bound. If the stated Wasserstein infimum were zero, an approximating sequence of at most k-point measures would converge weakly to η, contradicting this closure.

For the two-beam example, the ordered incoming pair $(re_1, -re_1)$ has positive probability. Its reflected first velocity is

$$v'(n) = re_1 - 2rn_1 n.$$

Consider $n(\theta) = \cos\theta\, e_1 + \sin\theta\, e_2$ with $0 < \theta < \pi/2$. The outputs $v'(n(\theta))$ form a nonconstant continuous arc. The acceptance probability is positive at each such direction and remains positive on a neighborhood of it. Every neighborhood of each corresponding output therefore receives positive accepted probability, by continuity and full support of surface measure. The unconditional output law has infinitely many support points. All velocities remain in the ball of radius r, so it has a finite first moment and the preceding argument applies. □

B.3 CONTEXTUAL INTENSITIES AND FINITE-DEPTH UNIVERSALITY

We prove the new equivalence (iv) in Theorem A. The proof has two jobs. Sequential Bernoulli decisions must realize a context-dependent finite mixture, and later decisions must retain access to the initial context even though earlier jumps have changed it. Small disjoint copies of the state space provide this memory inside the original E. A two-point orientation mark then realizes the decisions through the single shared gate of a pair update. No immutable input channel or enlarged particle state is assumed.

Throughout this subsection $E \subset \mathbb{R}^d$ is nonempty, compact, and convex, $D = \operatorname{diam}(E)$, and all context measures carry W_1. The deterministic attention update in every block is the identity. A finite composition always refreshes its context by its own output law: for one-step kernels P_j, set

$$\mu_0 = \mu, \qquad \mu_{j+1} = \int_E P_j(\mu_j, x; \cdot)\mu_j(\,\mathrm{d}x), \qquad X_{j+1} \mid X_j \sim P_j(\mu_j, X_j; \cdot), \quad X_0 = z.$$

The resulting conditional kernel is $\operatorname{Law}(X_L)$ for the fixed initial context μ and state z.

Proposition B.4 (A one-block limitation for contextual rates)**.** *On $E = [0,1]$, let*

$$K_*(\mu, z) = \delta_{\frac{1}{2}+\frac{1}{2}(\int x\,\mu(\,\mathrm{d}x))^2}.$$

Every one-block conditional kernel with identity deterministic update, context-independent jump map, fixed mark law, and arbitrary contextual acceptance probabilities in $[0,1]$ *satisfies*

$$\sup_{\mu\in\mathcal{P}(E),\, z\in E} W_1(K_{h,\Theta}(\mu,z), K_*(\mu,z)) \geq \frac{1}{8}.$$

Lemma B.5 (A one-block obstruction)**.** *Let* $E = [0,1]$*. Fix any probability mark law* ν *and any measurable context-independent proposal map* $\phi : E^2 \times \Xi \to E$*. For arbitrary measurable acceptance probabilities* $a : \mathcal{P}(E) \times E^2 \times \Xi \to [0,1]$*, put*

$$P(\mu,z) = \int_{E\times\Xi} \left[(1-a(\mu,z,w,\xi))\delta_z + a(\mu,z,w,\xi)\delta_{\phi(z,w,\xi)}\right] \mu(\,\mathrm{d}w)\nu(\,\mathrm{d}\xi).$$

For $t(\mu) = \frac{1}{2} + \frac{1}{2}(\int x\,\mu(\,\mathrm{d}x))^2$*,*

$$\sup_{\mu,z} W_1(P(\mu,z), \delta_{t(\mu)}) \geq \frac{1}{8}.$$

Proof. It suffices to fix $z = 0$. Remove the acceptance decision and denote the proposal law by $Q(\mu) = \int \delta_{\phi(0,w,\xi)}\mu(\,\mathrm{d}w)\nu(\,\mathrm{d}\xi)$. This law depends affinely on μ. Since $t(\mu) \in [1/2, 1]$, every $y \in [0,1]$ satisfies $|y - t(\mu)| \leq t(\mu)$. Therefore replacing a rejected output 0 by its proposed output can only decrease its distance to the target point. In particular,

$$W_1(Q(\mu), \delta_{t(\mu)}) \leq W_1(P(\mu,0), \delta_{t(\mu)}). \tag{30}$$

Take $\mu_0 = \delta_0$, $\mu_1 = \delta_1$, and $\mu_{1/2} = (\mu_0 + \mu_1)/2$. Their target points are $1/2$, 1, and $5/8$, while $Q(\mu_{1/2}) = (Q(\mu_0) + Q(\mu_1))/2$. If e is the supremum error in the statement, convexity of W_1 and (30) give

$$\begin{aligned}\frac{1}{4} &= W_1\left(\tfrac{1}{2}\delta_{1/2} + \tfrac{1}{2}\delta_1, \delta_{5/8}\right) \\ &\leq \tfrac{1}{2}W_1(\delta_{1/2}, Q(\mu_0)) + \tfrac{1}{2}W_1(\delta_1, Q(\mu_1)) + W_1(Q(\mu_{1/2}), \delta_{5/8}) \leq 2e.\end{aligned}$$

This proves Proposition B.4. The argument does not require continuity of the proposal or the acceptance function. □

Lemma B.6 (Remembering the initial state inside E)**.** *If* E *is not a singleton, then for every integer* $m \geq 2$ *there are pairwise disjoint compact sets* $E_0, \ldots, E_{m-1} \subset E$*, affine homeomorphisms* $T_j : E \to E_j$ *with a common scaling factor, a Lipschitz recovery map* $\mathcal{R} : E \to E$*, and a Lipschitz function* $\chi_0 : E \to [0,1]$ *such that*

$$\mathcal{R}(T_j z) = z, \qquad \chi_0|_{E_0} = 1, \qquad \chi_0|_{E_j} = 0 \quad (j \geq 1).$$

For each $a \geq 1$ *a projected affine map* $U_a : E \to E$ *sends* $T_0 z$ *to* $T_a z$*. Given output points* $y_1, \ldots, y_m \in E$*, there is a Lipschitz map* $D_* : E \to E$ *satisfying*

$$D_*|_{E_a} = y_a \quad (1 \leq a < m), \qquad D_*|_{E_0} = y_m.$$

Proof. Work in the affine hull of E, whose dimension is positive. Choose $e \in \operatorname{ri} E$ and $r_* > 0$ for which the relative closed ball of radius r_* about e lies in E. Choose distinct centers $\zeta_0, \ldots, \zeta_{m-1}$ in the relative ball of radius $r_*/2$. For a sufficiently small $s > 0$, the sets

$$E_j = \zeta_j + s(E - e), \qquad T_j z = \zeta_j + s(z - e)$$

lie in E and are pairwise disjoint. Indeed, $E - e$ is bounded, and the finite set of centers has a positive minimum separation. This also handles lower-dimensional convex sets without changing their ambient state space.

Choose $a_* > 0$ smaller than one third of the minimum distance between the E_j, and set $\psi_j(x) = \max\{0, 1 - \operatorname{dist}(x, E_j)/a_*\}$. At most one $\psi_j(x)$ is nonzero. Let Π_E be Euclidean projection onto E and define

$$\mathcal{R}(x) = \sum_{j=0}^{m-1} \psi_j(x)\Pi_E\left(e + \frac{x - \zeta_j}{s}\right) + \left(1 - \sum_{j=0}^{m-1}\psi_j(x)\right)e.$$

This is an E-valued Lipschitz map by convexity and compactness, and it recovers z on each $T_j z$. Set $\chi_0 = \psi_0$ and $U_a(x) = \Pi_E(x + \zeta_a - \zeta_0)$; for $x = T_0 z$ its argument is $T_a z \in E$. Finally,

$$D_*(x) = y_m + \sum_{a=1}^{m-1} \psi_a(x)(y_a - y_m)$$

is a convex combination of the displayed output points and has all the required values. The construction uses only finite-dimensional maps. □

Lemma B.7 (Sequential mixture selection by exchange-symmetric pair gates)**.** *Suppose $m \geq 2$, $0 < \gamma < 1$, and continuous weights $\bar{p}_a : \mathcal{P}(E) \times E \to [\gamma/m, 1]$ sum to one. Set $n = \lceil \log_2(m/\gamma) \rceil$. With the encodings of Lemma B.6, a composition of $2 + (m-1)n$ law-level pair blocks realizes*

$$K_{\mathrm{mix}}(\mu, z) = \sum_{a=1}^{m} \bar{p}_a(\mu, z)\delta_{y_a}.$$

The deterministic updates are identities, the pair jump maps are context-independent, and the mark law is the fixed uniform law on $\{-1, +1\}$. Only the intensities depend on the evolving context. All intensities lie in $[0, h^{-1}]$ for linear acceptance with any fixed $h > 0$. In this exact construction the final decoder is D_*; it will be approximated by an MLP in the neural realization below.

Proof. The first block uses intensity h^{-1} and the pair map $(x, w) \mapsto (T_0 x, T_0 w)$. Every proposal is accepted, so the initial context becomes $(T_0)_{\#}\mu$ and the specified input becomes $T_0 z$. For $a < m$ define the remaining mass and its conditional selection probability

$$S_a(\mu, z) = \sum_{j=a}^{m} \bar{p}_j(\mu, z), \qquad r_a(\mu, z) = \frac{\bar{p}_a(\mu, z)}{S_a(\mu, z)}.$$

They are continuous and $1 - r_a = S_{a+1}/S_a \geq \gamma/m$. A two-point orientation mark selects which member of a proposed pair is eligible to move. That selection gives an individual member only half of the proposals. To obtain the desired probability after n sublayers, put

$$c_a(\mu, z) = 2\left[1 - (1 - r_a(\mu, z))^{1/n}\right].$$

Since $\gamma/m \geq 2^{-n}$, $0 \leq c_a \leq 1$, and $1 - (1 - c_a/2)^n = r_a$.

For current context $\rho \in \mathcal{P}(E)$ define

$$\alpha_a(\rho, x) = \chi_0(x) c_a(\mathcal{R}_{\#}\rho, \mathcal{R}(x)). \tag{31}$$

This is continuous on all of $\mathcal{P}(E) \times E$. For each of the n sublayers at stage a, draw $\sigma \in \{-1, +1\}$ with equal probability and use

$$\begin{aligned} \sigma = +1: &\quad \lambda_a(\rho, x, w, +1) = \alpha_a(\rho, x)/h, \quad \Phi_a(x, w, +1) = (U_a(x), w), \\ \sigma = -1: &\quad \lambda_a(\rho, x, w, -1) = \alpha_a(\rho, w)/h, \quad \Phi_a(x, w, -1) = (x, U_a(w)). \end{aligned} \tag{32}$$

Exchange of x, w, accompanied by $\sigma \mapsto -\sigma$, exchanges the outputs and leaves the scalar rate unchanged. This is precisely the marked exchange convention in Appendix D. The pair uses one Bernoulli gate; no two independent endpoint gates are assumed. Integrating over the partner and orientation gives the tagged rule

$$P_a(\rho, x) = \left(1 - \frac{\alpha_a(\rho, x)}{2}\right)\delta_x + \frac{\alpha_a(\rho, x)}{2}\delta_{U_a(x)}. \tag{33}$$

Every possible state remains in $E_0 \cup \cdots \cup E_{m-1}$. Accepted moves from E_0 send $T_0 z$ to $T_a z$, and selected copies have zero eligibility in subsequent stages. Both a move and a rejection preserve $\mathcal{R}(x)$. It follows, for every intermediate context ρ, that

$$\mathcal{R}_{\#}\rho = \mu. \tag{34}$$

Thus the inputs in (31) recover the original context and original state from the actual evolving context and state. While a particle remains in E_0, each of its n trials has acceptance probability

$c_a(\mu, z)/2$. Fresh marks and gates make the probability of remaining there through the stage equal to $(1 - c_a/2)^n = 1 - r_a$. Induction on a now gives, at the end of stage a,

$$\mathrm{Law}(X \mid \mu, z) = \sum_{j=1}^{a} \bar{p}_j(\mu, z)\delta_{T_j z} + S_{a+1}(\mu, z)\delta_{T_0 z}.$$

The induction starts with $S_1 = 1$ and uses $S_a r_a = \bar{p}_a$. After stage $m - 1$, a final always-accepted context-independent pair map $(x, w) \mapsto (D_*(x), D_*(w))$ sends the selected copies to y_a and the remaining copy to y_m. This proves the formula and the block count. □

Neural realization and stability of the evolving context. The recovery map is part of the definition of the target scalar functions α_a, not a separately supplied memory input to the network. Proposition A.3 approximates each $\alpha_a(\rho, x)$ uniformly on the compact context–state domain by an attention module; clipping gives $\widehat{\alpha}_a \in [0, 1]$. The same scalar module is evaluated at the endpoint selected by the mark, which preserves exchange symmetry exactly. Its input and output dimensions are fixed, regardless of the number of mixture categories. The maps T_0 and U_a are affine maps or projected affine maps, admitted exactly in Appendix A. An E-projected ReLU MLP uniformly approximates the continuous final decoder D_*. There is an explicit projected-MLP implementation of the marked branches. With $b_\sigma = (1 + \sigma)/2 \in \{0, 1\}$ and the centers ζ_j from Lemma B.6, the pair map is the projected affine map

$$\big(\Pi_E[x + b_\sigma(\zeta_a - \zeta_0)], \Pi_E[w + (1 - b_\sigma)(\zeta_a - \zeta_0)]\big).$$

The rate is obtained from the two shared scalar reads by

$$\frac{1}{h}\Big[\mathrm{ReLU}(\widehat{\alpha}_a(\rho, x) + b_\sigma - 1) + \mathrm{ReLU}(\widehat{\alpha}_a(\rho, w) - b_\sigma)\Big].$$

For $\sigma = +1$ this is $\widehat{\alpha}_a(\rho, x)/h$ and for $\sigma = -1$ it is $\widehat{\alpha}_a(\rho, w)/h$. Thus both range control and marked exchange symmetry hold exactly. Only the scalar rate reads the context.

Uniform approximation of the scalar rates must still control the context passed to later blocks. We give this step explicitly. Suppose all $\widehat{\alpha}_a$ have uniform error at most δ, and let ω be a common modulus for variation of the finitely many α_a in context at fixed state. Couple the exact and learned versions of (33) with the same orientation and gate draws. For each initial (μ, z), let e_j be the supremum, over those inputs, of the probability of any differing tagged acceptance decision in the first j selection sublayers. The initial encoding is exact, so $e_0 = 0$. Before a disagreement the states coincide. Integrating these couplings in the initial state also couples the two intermediate contexts, giving $W_1(\rho_j, \widehat{\rho}_j) \le De_j$. Consequently

$$e_{j+1} \le e_j + \tfrac{1}{2}\big[\delta + \omega(De_j)\big]. \tag{35}$$

For a fixed finite number of sublayers, induction and $\omega(s) \to 0$ as $s \to 0$ imply $e_j \to 0$ as $\delta \to 0$. If the final decoder error is δ_D, coupling its outputs costs at most $De_{(m-1)n} + \delta_D$. This proves uniform convergence of the composed conditional kernels, including the effect of approximate contexts; it does not assume that a learned approximation preserves (34) exactly.

Proof of Theorem A, the rate-only equivalence. If E is a singleton, its unique output law is already realized by the identity. Otherwise fix $0 < \varepsilon < D$ and a finite open $\varepsilon/4$ output net $y_1, \dots, y_m$. Choose a continuous partition of unity subordinate to these balls and integrate it against $K(\mu, z)$ to obtain continuous weights $p_a(\mu, z)$. The finite mixture $\sum_a p_a(\mu, z)\delta_{y_a}$ has uniform W_1 error at most $\varepsilon/4$. Set $\gamma = \varepsilon/(4D)$ and $\bar{p}_a = (1 - \gamma)p_a + \gamma/m$. This changes the output law by at most $\gamma D = \varepsilon/4$ and makes all weights at least γ/m.

Apply Lemmas B.6 and B.7. If $m = 1$, the constant output is an immediate special case. The exact mixture construction uses at most $2 + (m - 1)\lceil \log_2(4mD/\varepsilon) \rceil$ stochastic blocks. Then choose the scalar coefficient accuracy and final MLP accuracy so that (35) makes their additional output-law error less than $\varepsilon/2$. This gives the required approximation in the full uniform conditional metric, hence in d_{anc}. The original particle state remains in E, and the scalar coefficient modules have embedding dimension and head count independent of ε, N by Appendix A. The number of blocks and the parameters of these modules may depend on the target and tolerance.

Conversely, every finite composition used in (iv) has a jointly continuous conditional kernel. To check composition, if $P(\mu, z)$ and $Q(\mu, z)$ are continuous, then $F_P(\mu) = \int P(\mu, z)\mu(\,\mathrm{d}z)$ is continuous. For each continuous output test, integration of $Q(F_P(\mu), y)$ against $P(\mu, z;\,\mathrm{d}y)$ is continuous, by uniform continuity on the compact domain. Induction proves the claim. An approximation in d_{anc} therefore makes the finite normalized fibers a uniform limit of uniformly continuous restrictions. They have a common modulus, and Theorem B.1 yields (i) and (ii). Together with the one-block proof above, this completes all four equivalences. □

What the depth statement establishes. The obstruction and construction concern the same identity-transport subclass. The obstruction allows arbitrary fixed mark laws and arbitrary context-independent proposal maps; the construction already works with a two-point mark. A contextual deterministic transport would introduce an additional context-dependent coefficient, and is not part of this comparison. The block-count bound is constructive rather than optimal: the obstruction proves failure of one block, not a matching accuracy-dependent depth lower bound. The encoded maps are unconstrained by physical conservation laws. At fixed h their rates are bounded by h^{-1}; if the total duration is instead held fixed while the number of blocks grows, this construction does not have the depth-independent rate bound used in Theorem B.

The exact construction can also be applied in an even-N matching scheme. Each realized encoded empirical context retains its initial context under $\mathcal{R}_{\#}$, and the orientation of a specified particle is uniform at each sublayer. Its one-particle conditional law therefore has the mixture probabilities just proved. The joint law can retain correlations between particles paired in a sublayer. For a fixed learned network, repeated use of Proposition D.1 relates the empirical implementation to the law-level context recursion as $N \to \infty$; no finite-N independence of all outputs is claimed.

C A FINITE-RATE KINETIC EVOLUTION AND ITS DEPTH APPROXIMATION

We prove the well-posedness, repeated-block convergence, and endpoint regularity in Theorem B. Here h is the physical duration of the law-level block or matching round. The single-pair implementation uses the physical-time normalization in Appendix D.1. Let $E \subset \mathbb{R}^d$ be compact and convex. Restrict ambient C_b^1 functions to E as the smooth test class. Assume $b_t(\mu, z)$ is bounded and jointly Lipschitz in t, μ, z for $|t - s| + W_1(\mu, \eta) + |z - w|$. Every frozen-background flow is assumed to leave E invariant. Fix $\Lambda > 0$ and $J_t(\mu, z) \in \mathcal{P}(E)$ satisfying

$$W_1(J_t(\mu, z), J_s(\eta, w)) \le L_J(|t - s| + W_1(\mu, \eta) + |z - w|). \tag{36}$$

The generator is $\mathcal{L}_{t,\mu}\varphi = b_t(\mu, z) \cdot \nabla\varphi(z) + \Lambda[\langle\varphi, J_t(\mu, z)\rangle - \varphi(z)]$.

For the bounded pair coefficients of Theorem B, let $r_t(\mu, z) = \int_{E\times\Xi} \lambda_t(\mu, z, z_*, \xi)\mu(\,\mathrm{d}z_*)\nu(\,\mathrm{d}\xi)$. The probability law after one rate-Λ candidate is

$$\begin{aligned} J_t(\mu, z) = &\left(1 - \frac{r_t(\mu, z)}{\Lambda}\right)\delta_z \\ &+ \frac{1}{\Lambda}\int_{E\times\Xi} \lambda_t(\mu, z, z_*, \xi)\delta_{\Phi_t^1(\mu,z,z_*,\xi)}\mu(\,\mathrm{d}z_*)\nu(\,\mathrm{d}\xi). \end{aligned} \tag{37}$$

Its mass at z includes null events. Substitution shows that its generator is exactly (14). Coupling the partner laws and using the coefficient Lipschitz bounds verifies (36). Write $D = \mathrm{diam}(E)$. If both exact and approximate intensities lie in $[0, \Lambda]$, let δ_λ be their uniform difference and δ_Φ the uniform difference of their first output states. Then

$$W_1(J^\Theta(\mu, z), J(\mu, z)) \le \delta_\Phi + (D/\Lambda)\delta_\lambda.$$

Indeed, insert the exact outcome map in the approximate-rate integral and test against a 1-Lipschitz function; the change of outcome costs δ_Φ, and the remaining rate error multiplies $|\varphi(\Phi^1) - \varphi(z)| \le D$. Moreover $(1 - h\Lambda)\delta_z + h\Lambda J^\Theta$ equals the law of sampling the partner and mark and accepting with probability $h\lambda_\vartheta$. Thus the candidate-clock description recovers the original context block, not only its first-order generator. A probability solution of its nonlinear martingale problem has its own one-time laws in the coefficients. For a fixed starting law μ at s, $K_{s,t}(\mu, z)$ instead starts a tagged particle at any z in the already determined background evolution from μ.

Lemma C.1 (Self-consistency and conditional stability)**.** *Under these assumptions the nonlinear process and its law equation are unique. For a constant C_T depending only on the displayed bounds, E, and T,*

$$\sup_{t\le T} W_1(S_{0,t}\mu, S_{0,t}\eta) \le C_T W_1(\mu,\eta), \tag{38}$$

$$\sup_{t\le T} W_1(K_{0,t}(\mu,z), K_{0,t}(\eta,w)) \le C_T[W_1(\mu,\eta)+|z-w|]. \tag{39}$$

Proof. Freeze a continuous law path m. The frozen-flow invariance assumption also ensures invariance for this time-dependent background. Indeed, for each $z \in E$ and outward normal n to the closed convex set E at z, invariance of the frozen flow gives $n \cdot b_t(\mu, z) \le 0$. This holds with $\mu = m_t$ at every time. Equivalently, solve with a piecewise-constant approximation to m on successive intervals; each such flow stays in E, and Lipschitz flow stability and closedness of E give the same conclusion in the limit. Between the finitely many events of a rate-Λ Poisson clock solve that ODE with velocity $b_t(m_t, \cdot)$. At an event draw from $J_t(m_t, Z_{t-})$. Existence, uniqueness in law, and nonexplosion of this inhomogeneous process follow directly from the finite clock and the Lipschitz ODE. For two backgrounds m, n, use the same clock, optimally couple the initial states, and measurably couple the two event laws optimally. Measurable optimal couplings exist on compact Polish spaces; ε-optimal selections give the same estimate by passage to zero. If $D_t = \mathbb{E}|Z_t - \widetilde{Z}_t|$, the drift and event bounds imply

$$D_t \le D_0 + C\int_0^t [D_r + W_1(m_r, n_r)]\,\mathrm{d}r.$$

The output law is continuous in time in W_1: over an interval of length r, deterministic motion costs at most $\|b\|_\infty r$ and candidates cost at most $\Lambda D r$ in expectation. It has the prescribed initial law. Thus the frozen-path-to-law map leaves the complete space of continuous probability-law paths with that initial law invariant. Gronwall bounds the distance of its output laws by a constant times the integral of the distance of the background paths. Iterating that integral gives a factor $(C_T T)^k/k!$. Therefore some iterate of the frozen-path-to-law map is a contraction on the complete space of continuous law paths with fixed initial law. Its unique fixed point is also the unique fixed point of the original map. This constructs the nonlinear process. Coupling two fixed points and using $W_1(m_t, n_t) \le D_t$ proves (38); coupling arbitrary tagged starts in those backgrounds then proves (39).

For a fixed m, let $U^m_{r,t}$ push a signed measure forward by the drift flow from r to t, and set

$$Q^*_{r,m_r}\rho = \Lambda\left[\int J_r(m_r, y)\,\rho(\,\mathrm{d}y) - \rho\right].$$

A weak solution ρ_t of the frozen forward equation satisfies the variation-of-constants identity

$$\rho_t = U^m_{s,t}\rho_s + \int_s^t U^m_{r,t} Q^*_{r,m_r}\rho_r\ \mathrm{d}r,$$

where the integral is understood against continuous tests. One obtains this identity by using backward transport tests in the weak equation; for a Lipschitz drift, smooth approximations to those tests and flow stability justify the same identity. The propagator $U^m_{r,t}$ is a total-variation contraction, and $\|Q^*_{r,m_r}\rho\|_{\mathrm{TV}} \le 2\Lambda\|\rho\|_{\mathrm{TV}}$. For two solutions with the same initial law, iterating the identity k times therefore bounds their difference on $[s,T]$ by $2[2\Lambda(T-s)]^k/k!$, which tends to zero. Thus the frozen forward equation is unique. A nonlinear law-equation solution is consequently its own frozen-law fixed point and must equal the law already constructed. This is the usual bounded nonlinear Markov argument; see Graham & Méléard (1997, Section 4). □

Lemma C.2 (Local consistency, including learned streaming)**.** *Let $\mathsf{T}^\mu_{t,h}$ be the exact flow over duration h with velocity $b_t(\mu,\cdot)$ frozen in time and law. Stream by $T_{h,\theta}$ satisfying*

$$T_{h,\theta}(t,\mu,\cdot) : E \to E, \qquad \sup_{t,\mu,z} |T_{h,\theta}(t,\mu,z) - \mathsf{T}^\mu_{t,h} z| \le h\delta_b + C_S h^2.$$

Then with probability $h\Lambda \le 1$ draw from an E-valued kernel $J^\Theta_t(\bar\mu,\bar z)$, where $\bar\mu = T_{h,\theta}(t,\mu,\cdot)_{\#}\mu$ and $\sup_{t,\mu,z} W_1(J^\Theta_t(\mu,z), J_t(\mu,z)) \le \delta_J$. The one-step context and worst tagged endpoint errors against the exact evolution are at most $Ch^2 + Ch(\delta_b+\delta_J)$.

Proof. The exact context moves by at most $(\|b\|_\infty + \Lambda D)r$ in W_1 in time r. The probability of two or more candidates in duration h is at most $(\Lambda h)^2/2$. Couple the Bernoulli candidate with the Poisson count: disagreement in the zero/one event indicator has probability $O(h^2)$. These exceptional cases cost at most $O(Dh^2)$. With no event, freezing time and law in the drift costs $O(h^2)$ by the Lipschitz bounds. Conditional on one event, replace its actual time by the interval endpoint. The pre-event state, background, and event time change by $O(h)$, so (36) bounds its output-law error by $O(h)$. The remaining post-event transport also moves a state by at most $O(h)$. Multiplication by the event's $O(h)$ probability gives $O(h^2)$.

The assumed approximate-stream error changes both the state and its pushed-forward law by at most $h\delta_b + C_S h^2$. Insert the target kernel at these approximate inputs; replacing it by J^Θ costs at most $h\Lambda\delta_J$. The target's Lipschitz bound controls the remaining input differences. This gives the result without a Lipschitz estimate for J^Θ. All these comparisons are couplings, hence give both context and conditional W_1 bounds. □

Proof of Theorem B. Lemma C.1 gives well-posedness and conditional regularity for the candidate law (37). Applying its estimates on $[s, t]$ gives (16). Let δ_b, δ_J and C_S be the approximation quantities in Lemma C.2. At each layer insert the exact target block at the approximate current context before comparing with the target evolution. Target stability and that lemma imply, for the sum e_j of context and worst conditional endpoint errors,

$$e_{j+1} \le (1 + Ch)e_j + Ch^2 + Ch(\delta_b + \delta_J), \qquad e_0 = 0.$$

The context error is bounded first, then the tagged error in those contexts; the displayed combined inequality bounds both. Discrete Gronwall yields

$$\sup_{\mu,z} \max_{j \le L} W_1(K^{h,\Theta}_{0,t_j}(\mu, z), K_{0,t_j}(\mu, z)) \le C_T(h + \delta_b + \delta_J). \tag{40}$$

For pair coefficient errors $\delta_\lambda, \delta_\Phi$ defined above, $\delta_J \le \delta_\Phi + D\delta_\lambda/\Lambda$.

Proposition A.3 approximates b, λ, Φ on the compact state, time, and mark domains, with rate clipping to $[0, \Lambda]$ and projection of outputs into E^2. A domain-preserving deterministic update is $G_{h,\theta}(t, \mu, z) = \Pi_E[z + hb_\theta(t, \mu, z)]$. Since the exact frozen flow remains in E, nonexpansiveness of Π_E and the bounded Lipschitz Euler remainder give

$$|G_{h,\theta}(t, \mu, z) - \mathsf{T}^\mu_{t,h} z| \le h\|b_\theta - b\|_\infty + Ch^2.$$

The same estimate and $\mathsf{T}^\mu_{t,h} z = z + hb_t(\mu, z) + O(h^2)$ prove the limiting-displacement assertion in Theorem B(ii). Choose all coefficient errors tending to zero as $L \to \infty$. Equation (40) then proves the limit in (17). This proves existence of the coefficient sequence asserted in Theorem B. At fixed time, conditional regularity and Theorem B.1 put the endpoint joint law in the static class. □

C.1 Theorem B': recovering the local mechanism

Let $K_{s,t}$ be jointly continuous in s, t, μ, z, equal to δ_z when $s = t$, and satisfy (18). Suppose b has the preceding bounds and for every smooth test φ the following limit exists uniformly over admissible t, μ, z:

$$B_{t,\mu}\varphi(z) = \lim_{h \downarrow 0} \frac{\langle \varphi, K_{t,t+h}(\mu, z)\rangle - \varphi(z)}{h} - b_t(\mu, z) \cdot \nabla\varphi(z). \tag{41}$$

Assume it extends linearly to $C(E)$ and that $B1 = 0$. The field b is one whose existence is assumed in Theorem B'; it need not be the derivative of mean displacement. Once positivity supplies J, the intrinsic bound (21), divided by Λ, is equivalent by Kantorovich duality to its Wasserstein Lipschitz bound. The extension is part of the finite-jump hypothesis: a diffusion derivative does not generally extend as such a bounded functional.

Proposition C.3 (Positive derivative, finite event measure)**.** *For a fixed $\Lambda > 0$,* (20) *is equivalent to the existence of a unique $J_t(\mu, z) \in \mathcal{P}(E)$ such that*

$$B_{t,\mu}\varphi(z) = \Lambda \left[\int \varphi(y) J_t(\mu, z; \, \mathrm{d}y) - \varphi(z) \right].$$

If J satisfies (36)*, the given endpoint family is the unique family of Lemma C.1. Conversely that family has all these properties.*

Proof. For each t, μ, z, the functional $\ell(\varphi) = B_{t,\mu}\varphi(z) + \Lambda\varphi(z)$ is positive and $\ell(1) = \Lambda$. Thus $|\ell(\varphi)| \le \Lambda\|\varphi\|_\infty$; Riesz representation supplies the unique positive measure ΛJ. Its mass at z corresponds to null events. Changing the chosen budget changes only this null mass, not the off-diagonal event measure. The converse positivity follows immediately from integration of a nonnegative test.

To identify more than the tangent, subdivide $[s, t]$ and apply the restart identity at every subinterval. For a fixed smooth test, (41) writes each conditional increment as its length times $\mathcal{L}_{r,S_{s,r}\mu}\varphi$ plus a uniformly $o(h)$ residual. Write $\epsilon_\varphi(\delta)$ for the supremum of the absolute residual per unit time over admissible increments of length at most δ. Uniformity in (41) gives $\epsilon_\varphi(\delta) \to 0$. A partition with mesh δ therefore has total residual at most $(t - s)\epsilon_\varphi(\delta)$. The bound (36) and continuity of b make $(r, \rho, y) \mapsto \mathcal{L}_{r,\rho}\varphi(y)$ continuous for each fixed smooth test; joint continuity of K makes both $S_{s,r}\mu$ and the tagged law continuous in r. These facts justify passing the sum to its Riemann integral, giving

$$\langle \varphi, K_{s,t}(\mu, z)\rangle = \varphi(z) + \int_s^t \int \mathcal{L}_{r,S_{s,r}\mu}\varphi(y) K_{s,r}(\mu, z;\, \mathrm{d}y)\, \mathrm{d}r.$$

Integrate also in the initial μ to get the nonlinear forward equation. Its uniqueness, and then uniqueness of the tagged forward equation in that background, follow from Lemma C.1. Conversely the constructed process has restart by its Markov property in the deterministic nonlinear background. The bounded rate, bounded drift, and joint Lipschitz estimates give the uniform diagonal derivative, for example by the zero/one-event comparison in Lemma C.2. This proves the controlled intrinsic converse. □

Proposition C.3 proves Theorem B': positivity recovers the candidate law, restart propagates its derivative to each interval, and Lemma C.1 identifies the resulting evolution. This is a relative characterization under a specified generator convergence and regularity regime. It does not assert that static kernel convergence by itself provides a derivative, nor that every continuous kernel can be embedded in a finite-rate time family.

D THE FINITE PARTICLE ALGORITHM AND ITS GENERATOR

Algorithm 1 A parallel stream–jump round

Require: States $z_1, \dots, z_N \in E$, $N \ge 2$; step $h > 0$; parameters $\Theta = (\theta, \vartheta, \eta)$; mark law ν.
Ensure: Updated states $z_1^+, \dots, z_N^+$ and context $\mu^{N,+}$.

1: Form $\mu^N = N^{-1}\sum_i \delta_{z_i}$ and stream all states: $\bar{z}_i \leftarrow G_\theta(\mu^N, z_i)$.
2: Set $\bar{\mu}^N \leftarrow N^{-1}\sum_i \delta_{\bar{z}_i}$ and initialize $z_i^+ \leftarrow \bar{z}_i$ for every i.
3: Draw a uniform matching $\mathcal{M}$ of distinct indices; if N is odd, leave one uniformly chosen index unmatched.
4: **for all** $\{i, \ell\} \in \mathcal{M}$ **in parallel do**
5: Draw $\xi_{i\ell} \sim \nu$ and set $q_{i\ell} \leftarrow (\bar{z}_i, \bar{z}_\ell, \xi_{i\ell})$.
6: Evaluate $\lambda_{i\ell} \leftarrow \lambda_\vartheta(\bar{\mu}^N, q_{i\ell})$ and $(y_i, y_\ell) \leftarrow \Phi_\eta(\bar{\mu}^N, q_{i\ell})$.
7: Draw $A_{i\ell} \sim \mathrm{Bernoulli}(p_h(\lambda_{i\ell}))$.
8: **if** $A_{i\ell} = 1$ **then**
9: $(z_i^+, z_\ell^+) \leftarrow (y_i, y_\ell)$.
10: **end if**
11: **end for**
12: Return $z_1^+, \dots, z_N^+$ and $\mu^{N,+} = N^{-1}\sum_i \delta_{z_i^+}$.

All pair evaluations use the same streamed context; conditional on that context and the matching, the marks and Bernoulli draws are independent across pairs. Exchange symmetry of the marked pair rule makes the orientation of an unordered pair immaterial in law (Appendix D). Matching allows $\lfloor N/2 \rfloor$ proposals to be processed together (Jin et al., 2020); context evaluation has the separate cost of the chosen attention module.

The sampled pair–mark construction is the kinetic foundation supplied by Kac's collision process and the particle methods of Bird and Nanbu (Kac, 1956; Bird, 1963; Nanbu, 1980; Graham & Méléard,

1997). Rate attention makes the interaction rule context-dependent and learnable. Algorithm 1 updates both states of an accepted pair; Nanbu's original construction updates one selected state using a sampled partner.

We record the calculation that connects the implementable matching block to the law-level rule. This is where the full empirical law enters the intensity. The collision-simulation viewpoint also underlies convergence analyses for the full Boltzmann equation (Babovsky & Illner, 1989).

D.1 Individual output laws and schedule normalization

The figure makes one pair proposal, whereas Algorithm 1 makes a matching round. Their conditional output laws are different at fixed N. We record that difference before passing to the law-level rule in (7). Fix indexed input states $\mathbf{z} = (z_1, \dots, z_N) \in E^N$, their streamed states $\bar{z}_i$, and the streamed context $\bar{\mu}^N$. Suppress the start time and write $\lambda_{i\ell}(\xi) = \lambda_\vartheta(\bar{\mu}^N, \bar{z}_i, \bar{z}_\ell, \xi)$ and $\Phi^a_{i\ell}(\xi) = \Phi^a_\eta(\bar{\mu}^N, \bar{z}_i, \bar{z}_\ell, \xi)$ for $a \in \{1, 2\}$. Pair exchange symmetry lets the particle being observed be written first. For a test φ, the pair increment is $\Delta_\Phi \varphi_{i\ell} = \varphi(\Phi^1_{i\ell}) + \varphi(\Phi^2_{i\ell}) - \varphi(\bar{z}_i) - \varphi(\bar{z}_\ell)$. Define the first particle's output law for partner index ℓ by

$$\mathcal{R}^{i\ell}_{h,\Theta}(\mathbf{z}) = \int_\Xi \big[(1 - p_h(\lambda_{i\ell}(\xi)))\delta_{\bar{z}_i} + p_h(\lambda_{i\ell}(\xi))\delta_{\Phi^1_{i\ell}(\xi)}\big]\nu(\,\mathrm{d}\xi),$$

and its output law conditional on being selected by $\mathcal{R}^i_{h,\Theta}(\mathbf{z}) = (N-1)^{-1}\sum_{\ell \neq i} \mathcal{R}^{i\ell}_{h,\Theta}(\mathbf{z})$. The same formula with $\ell = i$ is used only to compare with integration against the empirical measure; the finite algorithm never chooses that index. For a fixed label i, its conditional output law in the single-pair layer is

$$\left(1 - \frac{2}{N}\right)\delta_{\bar{z}_i} + \frac{2}{N}\mathcal{R}^i_{h,\Theta}(\mathbf{z}). \tag{42}$$

In an even-N matching round it is $\mathcal{R}^i_{h,\Theta}(\mathbf{z})$; for odd N it is $N^{-1}\delta_{\bar{z}_i} + (1 - N^{-1})\mathcal{R}^i_{h,\Theta}(\mathbf{z})$. Averaging the respective laws over i gives $\mathbb{E}[\mu^{N,+} \mid \mathbf{z}]$. This expectation is a measure on E; $\mathrm{Law}(\mu^{N,+} \mid \mathbf{z})$ is instead a law on empirical measures and retains the randomness of the whole output cloud.

By direct substitution of the empirical measure in (7),

$$K_{h,\Theta}(\mu^N, z_i) = \frac{N-1}{N}\mathcal{R}^i_{h,\Theta}(\mathbf{z}) + \frac{1}{N}\mathcal{R}^{ii}_{h,\Theta}(\mathbf{z}). \tag{43}$$

If E has diameter D and $p_h(\lambda) \leq h\Lambda$, every row on the right is within $Dh\Lambda$ of $\delta_{\bar{z}_i}$ in W_1. Hence $W_1(\mathcal{R}^i_{h,\Theta}, K_{h,\Theta}(\mu^N, z_i)) \leq 2Dh\Lambda/N$. The identities follow because, conditional on its being selected, a particle has a uniform partner among the other $N-1$ indices. Its probability of being selected is $2/N$ for one pair, 1 for an even matching, and $1 - 1/N$ for an odd matching. They explain both the without-replacement correction and why the static one-tag theorem is not a one-pair whole-cloud universality claim.

The expected context is a one-particle marginal. Write Z^+_i for the random output of particle i. Let I be uniform on $\{1, \dots, N\}$ independently of the block's random draws. For every bounded measurable φ,

$$\mathbb{E}[\langle \varphi, \mu^{N,+}\rangle \mid \mathbf{z}] = \frac{1}{N}\sum_i \mathbb{E}[\varphi(Z^+_i) \mid \mathbf{z}] = \mathbb{E}[\varphi(Z^+_I) \mid \mathbf{z}].$$

Taking conditional expectation given μ^N proves (9), with $Z^{+,N} = Z^+_I$. At fixed input states the marginal of an individual label can depend on that label's input state. If the input particles are exchangeable, expectation over their initial states makes all label marginals equal. For even N, put $\overline{F}^N_{h,\Theta}(\mathbf{z}) = N^{-1}\sum_i \mathcal{R}^i_{h,\Theta}(\mathbf{z})$. It is the expected output of a matching round. The expected single-pair output, with the same streamed cloud, is exactly

$$\mathbb{E}[\mu^{N,+}_{\mathrm{pair}} \mid \mathbf{z}] = \left(1 - \frac{2}{N}\right)\bar{\mu}^N + \frac{2}{N}\overline{F}^N_{h,\Theta}(\mathbf{z}). \tag{44}$$

For both schedules, the joint block law is a probability measure on E^N; the measure above is its marginal after uniformly selecting one output index.

Proposition D.1 (A fixed matching round realizes the one-particle output law)**.** *Let E be compact with diameter D. Fix h and a shared exchange-symmetric pair rule with acceptance probabilities in $[0,1]$. Suppose its context kernel $K_{h,\Theta}$ is jointly continuous. For input clouds with $W_1(\mu^N,\mu)\to 0$ in probability, one matching round satisfies*

$$\mathbb{E}W_1\big(\mu^{N,+},F_{h,\Theta}(\mu)\big)\longrightarrow 0.$$

No independence of the input states is required. More precisely, for even N,

$$W_1\big(\mathbb{E}[\mu^{N,+}\mid \mathbf{z}],F_{h,\Theta}(\mu^N)\big)\le D/N, \tag{45}$$

$$\mathrm{Var}(\langle\varphi,\mu^{N,+}\rangle\mid\mathbf{z})\le C\|\varphi\|_\infty^2/N \quad \textit{for bounded measurable } \varphi. \tag{46}$$

For odd N the first bound is $2D/N$, and the variance bound holds with another numerical constant.

Proof. In (43), couple the common $(N-1)/N$ part identically and couple the remaining two probability rows arbitrarily. The row error is at most D/N. Averaging over labels gives (45). The unselected label in an odd matching contributes at most another D/N.

For the variance, condition first on the input cloud and matching. Different pairs use independent marks and gates, and each pair contributes at most $2\|\varphi\|_\infty/N$ in absolute value to the empirical test. The sum of these conditional variances is at most $2\|\varphi\|_\infty^2/N$. It remains to control the randomness of the conditional matching mean. Generate a uniform matching by a uniform random permutation and pair consecutive entries; for odd N leave the last entry unmatched. Swapping two entries changes the partners of at most four particles. The conditional matching mean therefore changes by at most $8\|\varphi\|_\infty/N$. Expose the permutation one entry at a time. For two possible values of the next entry, their uniform completions can be paired by the transposition of those values. The corresponding conditional means differ by the same bound. Hence each of the at most N increments of the exposure martingale has absolute value at most $8\|\varphi\|_\infty/N$. Orthogonality of its increments gives a matching-mean variance at most $64\|\varphi\|_\infty^2/N$. The conditional variance decomposition proves (46), including odd N.

Choose $z_0\in E$. The functions with Lipschitz constant at most one and $\varphi(z_0)=0$ form a compact set in the uniform norm. For any $\eta>0$ choose a finite uniform η-net $\varphi_1,\dots,\varphi_m$ from it. Kantorovich duality and (46) imply, uniformly in $\mathbf{z}$,

$$\mathbb{E}\Big[W_1\big(\mu^{N,+},\mathbb{E}[\mu^{N,+}\mid\mathbf{z}]\big)\mid\mathbf{z}\Big]\le 2\eta+\sum_{a=1}^m\sqrt{\mathrm{Var}(\langle\varphi_a,\mu^{N,+}\rangle\mid\mathbf{z})}.$$

First take $N\to\infty$, then $\eta\downarrow 0$. Joint continuity of $K_{h,\Theta}$ on the compact context–state domain makes $F_{h,\Theta}$ continuous. Thus $F_{h,\Theta}(\mu^N)$ converges in probability to $F_{h,\Theta}(\mu)$; bounded diameter upgrades this to mean W_1 convergence. Combine with (45) and the triangle inequality. □

For a single-pair update, the label coupling moves at most two atoms, so $W_1(\mu^{N,+},\bar\mu^N)\le 2D/N$ for every realization. Thus one such block has a vanishing jump-stage effect as $N\to\infty$. The following time normalization retains its effect when the blocks are repeated.

Physical-time normalization and repeated single-pair layers. The deterministic stream must use the same physical duration as the jump schedule. For a drift $b_\theta(t,\mu,z)$, this means choosing $G_\theta=T_{\Delta,\theta}$ to approximate $z+\Delta b_\theta(t,\mu,z)$, with a domain-preserving stream and the consistency bounds used below. Free flight is $T_\Delta(x,v)=(x+\Delta v,v)$. A single uniformly chosen pair gives $\Delta=2h/N$, whereas a full even matching gives $\Delta=h$. For completeness, the time normalizations can also be read directly from the generator. For one uniformly chosen unordered pair, the probability of choosing a given edge is $2/[N(N-1)]$. Put $\Delta=2h/N$ and stream over this duration. Let $\mathcal{A}_t^\varphi$ be the target weak generator in (52). Under the compact, bounded, Lipschitz coefficient assumptions used below, with uniform coefficient error ε and a consistent duration-Δ stream, a single-pair layer satisfies

$$\mathbb{E}[\langle\varphi,\mu^{N,+}-\mu^N\rangle\mid\mathbf{z}]=\Delta\mathcal{A}_t^\varphi(\mu^N)+O_\varphi\big(\Delta(h+N^{-1}+\varepsilon)\big), \tag{47}$$

$$\mathrm{Var}(\langle\varphi,\mu^{N,+}\rangle\mid\mathbf{z})\le C_\varphi\Delta/N. \tag{48}$$

Indeed, after streaming its collision expectation is

$$\frac{2}{N^2(N-1)} \sum_{i<\ell} \int p_h(\lambda_{i\ell}) \Delta_\Phi \varphi_{i\ell} \, \mathrm{d}\nu.$$

Replace $p_h(\lambda)$ by $h\lambda$. Division by Δ gives the usual unordered-pair generator average $[N(N-1)]^{-1} \sum_{i<\ell} \int \lambda_{i\ell} \Delta_\Phi \varphi_{i\ell} \, \mathrm{d}\nu$. Its difference from the product empirical average is $O_\varphi(N^{-1})$. For the exponential gate, $|p_h(\lambda) - h\lambda| \le h^2\Lambda^2/2$, giving $O_\varphi(\Delta h)$ before division. Coefficient replacement costs $O_\varphi(\Delta\varepsilon)$, while the consistent stream and the change of its context cost $O_\varphi(\Delta^2)$; here $\Delta \le h$. Only one accepted pair contributes a random increment. Its size is at most $4\|\varphi\|_\infty/N$ and its acceptance probability is at most $h\Lambda$, which proves (48). Thus the figure has the same normalized first-order generator when its stream uses $\Delta = 2h/N$. Matching-round convergence in this manuscript continues to count rounds of duration h.

Corollary D.2 (Repeated single-pair layers have the same kinetic limit)**.** *Assume the compact bounded Lipschitz target and exchange-symmetric coefficient approximations of Theorem D.3. Use one uniformly selected pair per layer, proposal parameter h_N, and physical duration $\Delta_N = 2h_N/N$. The preceding stream maps E into itself and is consistent over Δ_N, with a common bounded speed and a uniform local error $O(\Delta_N^2 + \Delta_N\varepsilon_N)$; the pair-coefficient errors are at most ε_N, with a common rate bound. Suppose $h_N \downarrow 0$, $N \to \infty$, $\varepsilon_N \to 0$, and $W_1(\mu_0^N, \mu_0) \to 0$ in probability. Then the empirical laws, held constant between successive physical layer times $j\Delta_N$, converge to the same McKean–Vlasov law μ_t:*

$$\mathbb{E} \sup_{0 \le t \le T} W_1(\mu_t^{N,h_N}, \mu_t) \longrightarrow 0.$$

This holds for either gate in the architecture, with the linear gate used only where its probability is at most one.

Proof. Sum (47) over at most $T/\Delta_N + 1$ layers. The accumulated consistency error is $C_{\varphi,T}(h_N + N^{-1} + \varepsilon_N + \Delta_N)$. By (48) and Doob's inequality, the squared martingale supremum has expectation at most $C_{\varphi,T}/N$. Conditionally on the current cloud, streaming changes its empirical law by at most $C\Delta_N$, while the expected jump-stage distance is at most $(2D/N)h_N\Lambda_0 = D\Lambda_0\Delta_N$. Summing after stopping times gives the increment bound $C(\delta + \Delta_N)$ used for tightness. Compact containment is automatic on compact E. Passing the approximate weak equation to subsequential limits, its vanishing martingale and consistency errors identify the unique nonlinear law, exactly as in the proof of Theorem D.3 below. The limit law curve is continuous; convergence is therefore uniform in physical time in probability, and bounded diameter gives convergence in expectation. This uses roughly $TN/(2h_N)$ single-pair layers over $[0, T]$. □

For the linear gate the term $O(\Delta_N h_N)$ from gate approximation is zero. The same proof permits $h_N = h > 0$ fixed with $h\Lambda_0 \le 1$: then only $N \to \infty$, coefficient consistency, and $\Delta_N = 2h/N \to 0$ are needed. With the exponential gate and fixed h, the resulting rate is instead $(1 - e^{-h\lambda})/h$; taking $h \downarrow 0$ recovers the target rate λ.

D.2 Exchange conventions and conditional moments

Write $\Phi = (\Phi^1, \Phi^2)$ and $\Delta_\Phi\varphi(z, w, \xi) = \varphi(\Phi^1) + \varphi(\Phi^2) - \varphi(z) - \varphi(w)$. We assume the marked pair-event law is unchanged by exchanging z, w and the two output coordinates. A sufficient pointwise convention is a measurable involution $\iota : \Xi \to \Xi$ with $\iota_\# \nu = \nu$ such that

$$\lambda_t(\mu, z, w, \xi) = \lambda_t(\mu, w, z, \iota\xi), \qquad \Phi_t^1(\mu, z, w, \xi) = \Phi_t^2(\mu, w, z, \iota\xi),$$

with the companion identity exchanging the two output components. The same conditions are imposed on each learned pair rule. This is exchange symmetry, separate from the time-reversal symmetry used for entropy. For even N, independently draw a uniform perfect matching at every layer. The draws of marks and uniforms are independent across its pairs conditional on the current cloud and matching. The law is exchangeable when its initial law is exchangeable.

Let $\mathcal{F}_j$ be the information through layer j. First consider compact states, bounded rates $\lambda \le \Lambda_0$, and a consistent stream as in Lemma C.2. Given the streamed cloud, put

$$G_{ij} = \int p_h(\lambda(\bar{\mu}^N, \bar{z}_i, \bar{z}_j, \xi))\Delta_\Phi\varphi(\bar{z}_i, \bar{z}_j, \xi)\nu(\,\mathrm{d}\xi).$$

Each edge belongs to a uniform matching with probability $1/(N-1)$. Therefore the conditional collision drift of $\langle\varphi, \mu^N\rangle$ is exactly

$$\frac{1}{N(N-1)}\sum_{i<j} G_{ij}. \tag{49}$$

Replacing this by $\frac{1}{2}\iint\int p_h(\lambda)\Delta_\Phi\varphi\,\nu\,\bar{\mu}^N\bar{\mu}^N$ costs $C_\varphi h/N$. The difference is the diagonal and the factor $N/(N-1)$. For odd N, an edge has inclusion probability $1/N$ after one uniformly chosen label is omitted; the same error bound follows.

Conditional on the matching, collision increments from its different pairs are independent. Their variance is bounded by $C_\varphi h/N$ because each normalized pair increment is bounded by $4\|\varphi\|_\infty/N$ and is nonzero with probability at most $h\Lambda_0$. The variance of the conditional matching mean is at most $C_\varphi h^2$: its absolute value is at most $C_\varphi h$, irrespective of the matching. The conditional variance decomposition gives

$$\mathbb{E}\left[|\eta^\varphi_{j+1}|^2 \mid \mathcal{F}_j\right] \le C_\varphi(h/N + h^2), \tag{50}$$

where η^φ_{j+1} is the centered full one-step increment. This deliberately nonoptimal bound already allows arbitrary joint $N \to \infty$, $h \to 0$.

For $\varphi \in C_b^2$ the stream Taylor estimate, (49), and the gate expansion give

$$\mathbb{E}[\langle\varphi, \mu^N_{j+1} - \mu^N_j\rangle \mid \mathcal{F}_j] = h\mathcal{A}^\varphi_{t_j}(\mu^N_j) + O_\varphi(h^2 + h/N + h\varepsilon), \tag{51}$$

where the target coefficients define

$$\mathcal{A}^\varphi_t(\mu) = \langle b_t(\mu,\cdot)\cdot\nabla\varphi, \mu\rangle + \frac{1}{2}\iiint \lambda_t(\mu, z, w, \xi)\Delta_\Phi\varphi(z, w, \xi)\nu(\,\mathrm{d}\xi)\mu(\,\mathrm{d}z)\mu(\,\mathrm{d}w). \tag{52}$$

Here ε is a uniform coefficient error on the stated compact domains; insert the target coefficients before changing state or law arguments. In particular no uniform Lipschitz bound on successive neural approximants is used. The stream maps must still satisfy their stated consistency and domain conditions.

Theorem D.3 (Compact joint implementation limit)**.** *Assume bounded, Lipschitz target pair coefficients on compact E, a consistent stream, and exchange-symmetric approximants with common bounds and uniform error $\varepsilon \to 0$. Assume $W_1(\mu_0^N, \mu_0) \to 0$ in probability. If the target nonlinear law equation is well posed as in Appendix C, then along every joint sequence $N \to \infty$, $h \downarrow 0$, $\varepsilon \to 0$,*

$$\sup_{t\le T} W_1(\mu_t^{N,h}, \mu_t) \longrightarrow 0 \quad \textit{in probability and in mean.}$$

For exchangeable initialization, fixed-order particle marginals at any fixed time converge to the corresponding product law.

Proof. Use piecewise-constant interpolation at physical layer times; continuous interpolation of the stream has uniformly $O(h)$ distance from it. Summing (51), and centering each increment, gives

$$\langle\varphi, \mu_t^{N,h}\rangle = \langle\varphi, \mu_0^N\rangle + \int_0^t \mathcal{A}^\varphi_r(\mu_r^{N,h})\,\mathrm{d}r + M_t^\varphi + R_t^\varphi,$$

with $\sup_t|R_t^\varphi| \le C_{\varphi,T}(h + N^{-1} + \varepsilon)$ and, by (50) and Doob's inequality, $\mathbb{E}\sup_t|M_t^\varphi|^2 \le C_{\varphi,T}(h + N^{-1})$. Expected W_1 variation of a block is at most Ch: couple by labels, use bounded stream speed, and bound the expected fraction of accepted particles by $h\Lambda_0$. The same conditional bound summed after a stopping time gives $C(\delta + h)$ over an interval of length δ. Compactness of $\mathcal{P}(E)$ and this stopping-time bound give tightness of the empirical curves in the Skorokhod space.

On a convergent subsequence use a Skorokhod representation. The generator functional is continuous in μ: couple the two partner laws and use bounded target regularity, integrating the mark. Its

value at a converging curve consequently converges at almost every time. Pass the martingale identity for a countable dense set of smooth tests. The vanishing martingales and remainders show that every limit curve solves the target nonlinear law equation. Its integrated right side is continuous in time, so the limit is continuous. Uniqueness identifies the same deterministic limit on every subsequence. Skorokhod convergence to a continuous curve is uniform; bounded diameter upgrades probability convergence to expectation. Finally the standard exchangeable empirical-measure characterization of chaos gives the marginal conclusion (Sznitman, 1991; Chaintron & Diez, 2022). This is the compactness–uniqueness method of Graham & Méléard (1997, Theorem 4.5); the preceding calculations verify the particular discrete scheme. □

E HARD-SPHERE AND HARD-POTENTIAL COLLISION DYNAMICS

We now place physical collision models inside the learned class. Work on $E = \mathbb{T}^d \times \mathbb{R}^d$, $d \geq 2$, with exact free streaming and the elastic reflection

$$v' = v - ((v-w)\cdot n)n, \qquad w' = w + ((v-w)\cdot n)n, \qquad n \sim \nu = \mathrm{Unif}(S^{d-1}). \tag{53}$$

The full pair map sends $((x,v),(y,w))$ to $((x,v'),(y,w'))$; Φ^1, Φ^2 continue to denote complete phase states. Existence and stability for spatially mollified kinetic equations have a substantial theory (Friesen et al., 2019; 2022). Here the prescribed geometry and a learned contextual frequency are verified together. Unbounded velocities still permit unbounded changes at an accepted reflection, even for a bounded intensity.

E.1 A BOUNDED-RATE PHYSICAL EXAMPLE WITH A FULLY LEARNED INTENSITY

For cutoff Maxwell molecules, the collision rate is independent of the magnitude of relative velocity; its angular law describes scattering (Tanaka, 1978). A simple reflection-form Maxwell model has target $\lambda^{\mathrm{M}}((x,v),(y,w),n) = \nu_0\beta(x-y)$, where $\nu_0 > 0$ and $\beta \in C^1(\mathbb{T}^d)$ is even, nonnegative, and normalized by $\int \beta = 1$. The network represents this *entire* bounded intensity through (4). Neither ν_0 nor β need be included as multiplicative factors in the rate head. More generally, it may learn any bounded continuous contextual rate with the regularity in Theorem E.1.

For example, an exact pair-query attention rule is

$$C(\mu,q) = e^{-1}\int e^{qg(u)}\frac{1+g(u)}{2}\,\mu(\,\mathrm{d}x,\,\mathrm{d}u), \qquad a(\mu,z,z_*) = \frac{1}{4} + \frac{1}{2}C(\mu,q), \tag{54}$$

where $g(v) = \mathrm{clip}(v_1,-1,1)$ and the scalar query is $q = (g(v)+g(w))/2$. Using $\lambda = a$ gives a fully specified bounded contextual intensity. It lies in $[1/4, 3/4]$ and is Lipschitz and exchange symmetric (Appendix F). This is a realizable interaction rule, illustrating what the learned class contains.

E.2 A JOINT NEURAL–PARTICLE–DEPTH THEOREM

A common formulation also permits increasing rates. Write

$$\lambda_\vartheta(t,\mu,z,z_*,n) = \langle v-w\rangle^\gamma a_\vartheta(t,\mu,z,z_*,n), \qquad 0 \leq \gamma \leq 1. \tag{55}$$

Here $z = (x,v)$, $z_* = (y,w)$, and $\langle u\rangle = (1+|u|^2)^{1/2}$. For $\gamma = 0$ the neural output is the complete intensity. For $\gamma > 0$, the fixed factor specifies only an allowed rate-growth envelope; the network learns the bounded coefficient, including its spatial, angular, and context dependence. For a target a, the nonlinear law solves

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\varphi,\mu_t\rangle = \langle v\cdot\nabla_x\varphi,\mu_t\rangle + \frac{1}{2}\int \lambda(t,\mu_t,z,z_*,n)\Delta\varphi\,\mu_t(\,\mathrm{d}z)\mu_t(\,\mathrm{d}z_*)\nu(\,\mathrm{d}n), \tag{56}$$

where $\Delta\varphi = \varphi(x,v') + \varphi(y,w') - \varphi(x,v) - \varphi(y,w)$.

Theorem E.1 (Shared learned intensities recover elastic kinetic evolution)**.** *Fix $0 \leq \gamma \leq 1$ and the reflection* (53)*; suppose $a \in [0,A]$ is exchange symmetric and jointly continuous in all inputs. On every finite horizon, assume uniform Lipschitz bounds in time, both states, and the context law in W_1, uniformly in the mark n. Fix $a_* > 0$, $M_* < \infty$ and*

$$\mathcal{C}_{a_*,M_*} = \left\{\mu \in \mathcal{P}(E) : \int e^{a_*|v|^2}\,\mu(\,\mathrm{d}x,\,\mathrm{d}v) \leq M_*\right\}.$$

There exist shared attention modules $a_{\vartheta_m} \in [0, A]$, with fixed token embedding dimension, such that Algorithm 1 with (55)*, independent initial samples, and acceptance $1 - e^{-h\lambda_\vartheta}$ satisfies*

$$\lim_{\substack{m\to\infty,\, N\to\infty \\ L\to\infty}} \sup_{\mu_0 \in \mathcal{C}_{a_*, M_*}} \mathbb{E} \sup_{0\le t\le T} W_1\Big(\mu_t^{N,T/L,\vartheta_m}, \mu_t\Big) = 0 \qquad (T < \infty) \tag{57}$$

along every joint sequence. The target process is unique among processes whose laws have all polynomial velocity moments locally bounded in time. Its conditional solution kernels are continuous in initial context and starting state, locally uniformly on bounded starts and finite horizons.

Appendix F.1 verifies this formulation using the growth and stability estimates in Appendix E. Bounded Maxwell rates use $\gamma = 0$ on unbounded velocity space. Hard spheres are included with $\gamma = 1$ and

$$a^{\mathrm{HS}}((x, v), (y, w), n) = \beta(x - y)\frac{|(v - w) \cdot n|}{\langle v - w\rangle}. \tag{58}$$

This coefficient is bounded, continuous, and globally Lipschitz in the pair states. Consequently the angular hard-sphere factor is represented by the neural output, while the fixed envelope allows linear growth at large relative speed. For the Maxwell and hard-sphere targets, β is fixed and smooth; their equations are spatially mollified Boltzmann models. Exact spatial locality requires a separate density limit.

From coefficient accuracy to physical prediction. The same module works for all initial contexts and sufficiently fine resolutions: for each $T, \varepsilon > 0$, there are $\vartheta_\varepsilon, N_\varepsilon, L_\varepsilon$ such that

$$\sup_{\mu_0 \in \mathcal{C}_{a_*, M_*}} \mathbb{E} \sup_{t\le T} W_1\Big(\mu_t^{N,T/L,\vartheta_\varepsilon}, \mu_t\Big) < \varepsilon \quad (N \ge N_\varepsilon,\ L \ge L_\varepsilon). \tag{59}$$

Internal network capacity resolves the intensity, L resolves physical time, and N resolves the population law. This is an approximation theorem for prescribed target coefficients; estimating the intensities from observations is a statistical learning problem.

The proof compares the learned and target coefficients at the *same* empirical context first:

$$|a_{\vartheta_m}(\mu^N, q) - a(\mu, q)| \le |a_{\vartheta_m}(\mu^N, q) - a(\mu^N, q)| + |a(\mu^N, q) - a(\mu, q)|.$$

Uniform approximation controls the first term and target law regularity controls the second. Pair conservation bounds polynomial moments of the correlated particles. These bounds make large-velocity errors negligible and yield possible nonlinear limits. An exponential-moment estimate for each limiting law then controls the velocity weights in a coupling comparison and gives uniqueness. Appendices E–F supply the complete estimates. Conditional approximation also controls the one-particle output law through

$$W_1(F_\Theta(\mu), F(\mu)) \le \int_E W_1(K_\Theta(\mu, z), K(\mu, z))\, \mu(\,\mathrm{d}z). \tag{60}$$

This appendix proves the fixed-coefficient probabilistic statements underlying Theorem E.1 and a specified hard-potential extension. Appendix F assembles these statements with neural approximation to prove its learned-intensity formulation. The proof has three logically separate parts. Polynomial moments control the actual, correlated particles and permit limits. A sharper generating estimate is proved for each nonlinear candidate, where products of its moments are legitimate. That estimate supplies uniqueness, so all subsequences have the same limit. We do not restart a correlated cloud as an independent sample.

E.3 Physical assumptions and the meaning of a solution

Let $E = \mathbb{T}^d \times \mathbb{R}^d$, $d \ge 2$, with normalized spatial volume and distance $d_E((x, v), (y, w)) = d_{\mathbb{T}^d}(x, y) + |v - w|$. In this appendix $M_p(\mu) = \int |v|^p \mu(\,\mathrm{d}x,\, \mathrm{d}v)$ denotes a raw moment. Fix an even, nonnegative $\beta \in C^1(\mathbb{T}^d)$ with $\int \beta = 1$ and a probability mark law ν on S^{d-1}. The physical pair map is

$$z' = (x, v + \delta(v, w, n)), \quad z'_* = (y, w - \delta(v, w, n)), \qquad \delta(v, w, n) = -((v - w) \cdot n)n.$$

The cross section $B(u,n)$ is measurable in n, exchange and reflection invariant, and, for some $0 \le \gamma \le 1$,

$$0 \le B(u,n) \le C_B\langle u\rangle^\gamma, \tag{61}$$

$$(|u| \vee |\widetilde{u}|)|B(u,n) - B(\widetilde{u},n)| \le C_B(\langle u\rangle^\gamma + \langle \widetilde{u}\rangle^\gamma)|u-\widetilde{u}|. \tag{62}$$

Here $\langle u\rangle = (1+|u|^2)^{1/2}$. Hard spheres use $B(u,n) = |u\cdot n|$ and normalized surface measure; both inequalities follow from the triangle inequality. These hypotheses specify the extension rather than covering every convention for an angular hard-potential kernel.

Let $a_t(\mu,z,z_*,n) \in [0,A]$ be measurable in n, exchange symmetric, and jointly Lipschitz in (t,z,z_*,μ) for the sum metric, uniformly in n on every finite horizon. Define

$$\lambda_t(\mu,z,z_*,n) = a_t(\mu,z,z_*,n)\beta(x-y)B(v-w,n). \tag{63}$$

Its total rate has the envelope $C(1+|v|^\gamma+|w|^\gamma)$, with C depending on $A, \|\beta\|_\infty, C_B$. Constants below may also depend on the common Lipschitz constants, d, T, and the initial bound. They do not depend on N or h.

For each initial law, a solution is a law P on right-continuous, left-limit kinetic trajectories such that, with $\mu_t = P_t$ and for every compactly supported smooth kinetic test φ,

$$\varphi(Z_t) - \varphi(Z_0) - \int_0^t \mathcal{L}_{r,\mu_r}\varphi(Z_r)\,\mathrm{d}r$$

is a martingale. Here

$$\mathcal{L}_{t,\mu}\varphi(x,v) = v\cdot\nabla_x\varphi(x,v) + \int\!\!\int \lambda_t(\mu,z,z_*,n)[\varphi(z')-\varphi(z)]\nu(\,\mathrm{d}n)\mu(\,\mathrm{d}z_*).$$

The initial class is $\mathcal{C}_{a_*,M_*}$ in Theorem E.1. Uniqueness is asserted in the class $\sup_{t\le T} M_p(\mu_t) < \infty$ for every $p < \infty$ and every finite T. The construction below belongs to this class. The linear tagged process at any specified starting state is defined in the resulting deterministic background law.

The finite scheme uses iid initial states, exact free streaming, fresh matching, and the saturated gate $1-e^{-h\lambda}$. Between layer times we use free flight with the velocity held at its most recent value, and perform the sampled reflection at the next layer time. Piecewise-constant interpolation changes the empirical curve by at most $h\sup_t M_1(\mu_t^{N,h})$ in W_1; the limits are the same. Write j for a layer and k for a marginal order. An empirical measure is random; a k-particle marginal is the law of k coordinates, not that empirical measure.

E.4 Polynomial bounds for the actual finite system

Lemma E.2 (Conservative moment increments)**.** *For $p \ge 4$, any outgoing velocities with $|v'|^2 + |w'|^2 = |v|^2 + |w|^2$ satisfy*

$$[|v'|^p + |w'|^p - |v|^p - |w|^p]_+ \le C_p(|v|^{p-2}|w|^2 + |w|^{p-2}|v|^2). \tag{64}$$

For the finite scheme, uniformly over the stated initial class,

$$\sup_{N,h\le 1} \mathbb{E}\sup_{t\le T} M_p(\mu_t^{N,h}) \le C_{p,T}. \tag{65}$$

The same estimate holds for the continuous-time binary system with pair rates $\lambda/(N-1)$.

Proof. Put $X = |v|^2$, $Y = |w|^2$. Convexity bounds the sum of outgoing pth powers by $(X+Y)^{p/2}$. For $X \ge Y$, the mean value theorem bounds $(X+Y)^{p/2} - X^{p/2} - Y^{p/2}$ by $C_p X^{p/2-1}Y$; exchange the variables in the other case. This proves (64).

Let $\mathcal{F}_0$ be the sigma-field generated by the initial particle states. Condition on $\mathcal{F}_0$; the empirical energy $c_{E,N} = M_2(\mu_0^N)$ is then constant for every realization. Since $p_h(\lambda) \le h\lambda$ and each pair is selected with probability at most $1/(N-1)$, the expected sum of the positive increases in M_p per unit time is bounded by

$$C_p\big(M_{p-2}c_{E,N} + M_{p-2+\gamma}c_{E,N} + M_{p-2}M_{\gamma+2}\big).$$

The harmless $N/(N-1) \le 2$ is absorbed into C_p. For $2 \le r \le p$, $M_r \le c_{E,N}^{(p-r)/(p-2)} M_p^{(r-2)/(p-2)}$. Set $b = (2-\gamma)/(p-2) \in (0,1]$, adding ones to avoid zero-energy cases. The preceding bound is at most

$$C_p(1+c_{E,N})^{1+b}(1+M_p)^{1-b}.$$

If $Y_j = \mathbb{E}[1+M_p(\mu_j^N) \mid \mathcal{F}_0]$, concavity gives $Y_{j+1} \le Y_j + hC_p(1+c_{E,N})^{1+b} Y_j^{1-b}$. For $0 < b \le 1$, $(Y + hcY^{1-b})^b \le Y^b + bch$. Hence

$$Y_j \le [Y_0^b + C_p bT(1+c_{E,N})^{1+b}]^{1/b}.$$

The analogous differential inequality gives the same bound in continuous time. The running supremum of M_p is no greater than its initial value plus the sum of its positive increments. Integrating the positive-increment bound and the last estimate gives the same type of bound for that supremum, polynomial in $c_{E,N}$ and $M_p(\mu_0^N)$. For any fixed power, Jensen controls empirical moment powers by a higher initial empirical moment. All such initial expectations are uniformly finite on $\mathcal{C}_{a_*,M_*}$. This proves (65). For fixed N, energy confines each velocity to its finite initial energy shell. Total collision rate is bounded there, establishing finite-N nonexplosion for the continuous binary system. □

Lemma E.3 (Tagged running moments and stopping-time variation)**.** *For each $p \ge 4$,*

$$\sup_{N,h} \mathbb{E} \sup_{t \le T} |V_t^{1,N,h}|^p \le C_{p,T}.$$

The tagged path laws and the empirical law curves are tight on each finite horizon. Their expected state or W_1 variation over any bounded stopping interval of length δ is at most $C_T(\delta + h)$.

Proof. This is not obtained by exchanging a supremum with an empirical average. For a particular tag, $|v'| \le |v| + |w|$ and the rate envelope give a conditional positive-variation bound

$$\mathbb{E}[(|V_{j+1}^i|^p - |V_j^i|^p)_+ \mid \mathcal{F}_j] \le C_p h[1 + |V_j^i|^{p+\gamma} + M_{p+\gamma}(\mu_j^N)].$$

Sum this inequality and use exchangeability at each fixed layer, then Lemma E.2 at a higher moment. The initial moment plus that sum bounds the expected running maximum of the tag. Similarly its expected jump length in a layer is at most $Ch[1 + |V_j^i|^{1+\gamma} + M_{1+\gamma}(\mu_j^N)]$, and free flight contributes $h|V_j^i|$. Summing after a bounded discrete stopping time over at most $\delta/h + 2$ layers, and bounding the summands by their running maxima, gives $C_T(\delta + h)$. Label coupling gives the empirical version. The tagged running bound gives compact containment. For the empirical curve, a bounded $p > 1$ moment sublevel is compact in $\mathcal{P}_1(E)$ because space is compact and velocity tails are uniformly integrable; (65) supplies containment there. These containment and stopping-time variation bounds give Skorokhod tightness (Ethier & Kurtz, 1986, Chapters 3–4). □

E.5 An exponential estimate for nonlinear laws

The next argument does not yet use uniqueness or a limit theorem. It applies to any nonlinear pair law with all polynomial moments locally bounded, free streaming, pair-energy conservation, and total marked rate at most $c_0(1 + |v|^2 + |w|^2)$. Every candidate constructed below has those properties. Testing with velocity powers is justified by truncation and a higher polynomial moment. In particular its energy $c_E = M_2(\mu_t)$ is constant.

Lemma E.4 (Closed generating inequality)**.** *Set $m_n(t) = M_{2n}(\mu_t)$ and $H_K(t,a) = \sum_{n=1}^K a^n m_n(t)/n!$. Then, for almost every time,*

$$\partial_t H_K \le \frac{c_0}{2} H_K^2 + c_0 H_K(\partial_a H_K - c_E). \tag{66}$$

The constant does not depend on K.

Proof. For $n \ge 2$, pair energy implies

$$\begin{aligned}[(X')^n + (Y')^n - X^n - Y^n]_+ &\le (X+Y)^n - X^n - Y^n \\ &= \sum_{r=1}^{n-1} \binom{n}{r} X^r Y^{n-r}, \qquad X = |v|^2,\ Y = |w|^2.\end{aligned}$$

Multiplying by the rate envelope and integrating the pair form against $\mu_t \otimes \mu_t$ gives

$$m_n' \le \frac{c_0}{2} \sum_{r=1}^{n-1} \binom{n}{r} m_r m_{n-r} + c_0 \sum_{r=1}^{n-1} \binom{n}{r} m_{r+1} m_{n-r}.$$

For $n = 1$ the derivative vanishes. Multiply by $a^n/n!$ and sum to K. The sums are the terms of degree at most K in H_K^2 and $H_K(\partial_a H_K - c_E)$, respectively. Their remaining coefficients are nonnegative, proving (66). For fixed n, the absolute collision integrand is bounded by $C_n(1+|v|^{2n+2}+|w|^{2n+2})$. Thus dominated convergence from compactly supported velocity tests validates the moment identities before summing the finite series. □

Theorem E.5 (Nonlinear Gaussian moments on every finite horizon)**.** *Under the preceding pair assumptions, if $\mu_0 \in \mathcal{C}_{a_*,M_*}$, then for each finite T there are $a_T > 0$ and $C_T < \infty$, uniform over that initial class, such that*

$$\sup_{t \le T} \int e^{a_T |v|^2} \mu_t(\,\mathrm{d}x,\,\mathrm{d}v) \le C_T. \tag{67}$$

No Gaussian density envelope is asserted.

Proof. If $c_E = 0$, the velocity is zero almost surely for all times. Otherwise choose $a_0 \in (0, a_*)$, write $q_0 = \int (e^{a_0|v|^2} - 1)\,\mathrm{d}\mu_0$, and solve

$$q' = \frac{c_0}{2} q^2 - c_0 c_E q, \qquad a' = -c_0 q, \qquad q(0) = q_0, \quad a(0) = a_0.$$

While the denominator is positive their explicit solutions are

$$q(t) = \frac{q_0 e^{-c_0 c_E t}}{1 - \frac{q_0}{2c_E}(1 - e^{-c_0 c_E t})}, \tag{68}$$

$$a(t) = a_0 + 2\log\left[1 - \frac{q_0}{2c_E}(1 - e^{-c_0 c_E t})\right]. \tag{69}$$

For fixed $c_E > 0$, $q_0/a_0 \to c_E$, hence $a(T)/a_0 \to e^{-c_0 c_E T} > 0$ as $a_0 \downarrow 0$. Choose a_0 so that $a(T) > 0$. Let $Y_K(t) = H_K(t, a(t))$. Direct substitution in (66) yields

$$Y_K' - q' \le c_0 (Y_K - q)\left[\tfrac{1}{2}(Y_K + q) - c_E + \partial_a H_K(t, a(t))\right].$$

For fixed K the bracket is integrable on $[0, T]$ by the polynomial moment assumption. Since $Y_K(0) \le q_0$, positive-part Gronwall gives $Y_K \le q$. Monotone convergence in K, and $a(t) \ge a(T)$, prove (67).

We verify the needed uniformity even when the energy tends to zero. The initial bound gives $0 \le c_E \le E_*$ and, for $a_0 \le a_*/2$, $q_0 \le a_0 c_E + C_* a_0^2$, uniformly over the class. Use this upper bound as the comparison initial value. The function $r_T(c_E) = (1 - e^{-c_0 c_E T})/c_E$ extends continuously to $c_E = 0$ with value $c_0 T$. The denominator in (68) is consequently positive for a sufficiently small common a_0. The logarithm in (69) and a uniform Taylor bound yield

$$a(T) \ge a_0 e^{-c_0 E_* T} - C_{*,T} a_0^2 > 0$$

for a sufficiently small common a_0. The same formulas bound q uniformly. Take a_T to be this positive lower bound. This completes the estimate with no circular exponential assumption on the candidate. □

The products in (66) are products of moments of a nonlinear law. They are not expectations of products of correlated empirical moments. This is why we apply the theorem to candidate nonlinear limits, not directly to $\mathbb{E} H_K(\mu_t^{N,h})$.

E.6 Uniqueness and regularity of the conditional process

Lemma E.6 (Integrated event coupling)**.** *For the physical coefficients* (63)*, let tildes denote a second state pair and background, at the same time and mark. With* $U = |v| + |w| + |\widetilde{v}| + |\widetilde{w}|$*,*

$$\int\int_0^\infty |\delta(v,w,n)\mathbf{1}_{r\le\lambda} - \delta(\widetilde{v},\widetilde{w},n)\mathbf{1}_{r\le\widetilde{\lambda}}|\,\mathrm{d}r\,\nu(\,\mathrm{d}n)$$
$$\le C(1+U)^{\gamma+1}[d_E(z,\widetilde{z}) + d_E(z_*,\widetilde{z}_*) + W_1(\mu,\eta)]. \tag{70}$$

Proof. Splitting the r-integral into its common and unmatched parts bounds it by

$$(\lambda\wedge\widetilde{\lambda})|\delta-\widetilde{\delta}| + |\lambda-\widetilde{\lambda}|(|\delta|\vee|\widetilde{\delta}|).$$

The reflection increment is Lipschitz in the two velocities and has magnitude at most $|v-w|$. For the rate difference, first change $a\beta$ with B fixed, then change B. The Lipschitz bounds on a,β and (61) control the first difference times the jump magnitude by $C(1+U)^{\gamma+1}$ times the input distance. The second is controlled by (62). The same estimates bound the common part, uniformly in the mark. Integration proves the claim. □

Lemma E.7 (Nonexplosion and nonlinear uniqueness)**.** *For every continuous background path with all velocity moments locally bounded, the associated linear tagged martingale problem is nonexplosive and unique. Two nonlinear solutions from the initial exponential class satisfy, for some increasing modulus* $\Omega_T(r)\downarrow 0$ *as* $r\downarrow 0$*,*

$$\sup_{t\le T} W_1(\mu_t,\eta_t) \le \Omega_T(W_1(\mu_0,\eta_0)).$$

In particular the nonlinear process is unique in the all-polynomial-moment class.

Proof. In a fixed background, stop the tagged velocity on leaving a ball. Its integrated rate is bounded there, so successive jumps and free flight give a unique stopped process. Pair energy bounds the positive tagged energy gain by $|w|^2$. Therefore

$$\mathcal{L}_{t,\mu_t}(1+|v|^2) \le C\int(1+|v|^\gamma+|w|^\gamma)|w|^2\mu_t(\,\mathrm{d}y,\,\mathrm{d}w) \le C_T(1+|v|^2).$$

The background moments through $\gamma+2$ make the constant finite. A stopped Dynkin estimate and Markov's inequality remove the velocity stops without explosion. On any bounded velocity region the number of accepted jumps has finite mean. Together these facts yield a nonexplosive finite-activity process on each finite interval, and stopped uniqueness gives uniqueness without stopping. The same localization constructs couplings of two such processes.

For nonlinear solutions Theorem E.5 supplies a common Gaussian moment. At each time couple their partner laws optimally, use the same angular mark and acceptance coordinate r, and apply (70). The four velocity marginals have common exponential moments, so their sum U does as well, by Hölder's inequality; no independence of the coupling is required. If $0\le D\le C(1+U)$ and $r=\mathbb{E}D\in(0,1]$, splitting at $U=R$ gives

$$\mathbb{E}[(1+U)^{\gamma+1}D] \le (1+R)^{\gamma+1}r + C_T e^{-cR^2}.$$

Choose $R^2 = c^{-1}(1+|\log r|)$. Since $(\gamma+1)/2\le 1$, the right side is bounded by $C_T\omega(r)$ with $\omega(r) = r(1-\log r)$ near zero. Extend ω to a nondecreasing concave function on $[0,\infty)$. The term involving $W_1(\mu_t,\eta_t)$ in (70) is bounded by the mean coupled distance times a common polynomial moment. Free streaming contributes at most the velocity distance. Consequently that mean distance $d(t)$ obeys

$$d(t) \le d(0) + C_T\int_0^t \omega(d(r))\,\mathrm{d}r.$$

Because $\int_{0+}\mathrm{d}r/\omega(r)=\infty$, Osgood's comparison proves uniqueness and continuous dependence. For small initial distances it gives a bound of the form $C_T d(0)^{\exp(-C_T T)}$. The logarithmic coupling mechanism is related to Friesen et al. (2022, Theorem 2.4 and Lemma 5.3); the event estimate above checks the additional full-context multiplier and the moment theorem supplies its solution class here. □

Lemma E.8 (Conditional endpoints at arbitrary starting states)**.** *In the backgrounds from Theorem E.5, the conditional kernels $K_{0,t}(\mu, z)$ are jointly continuous in μ, z, uniformly for $t \le T$ and bounded sets of starting velocities, with μ in the common initial exponential class.*

Proof. The tagged start need not be distributed as its background. We therefore establish its own tail bound. For $0 < s \le a_T/2$, conservation gives $e^{s|v'|^2} \le e^{s|v|^2} e^{s|w|^2}$. The background exponential bound and $e^{s|w|^2} - 1 \le s|w|^2 e^{s|w|^2}$ show

$$\mathcal{L}_{t,\mu_t} e^{s|v|^2} \le C_T s(1 + |v|^2) e^{s|v|^2}.$$

Use a decreasing $s(t) > 0$ with $s' = -C_T s$ and an additional scalar integrating factor. Stopped Dynkin's formula bounds the expectation uniformly; Fatou removes the stops. Initial velocities in a bounded set give a common bound. Couple two tagged processes in their already constructed backgrounds. Lemma E.7 controls the background distance, and the new tagged tails allow the same event-coupling/Osgood argument for arbitrary starting states. Its error tends to zero uniformly as the initial law and state errors vanish. □

E.7 Passing the interacting scheme to its nonlinear process

We supply the identification details because a formal drift calculation alone would not establish Theorem E.1. Path laws are a proof device here; no path-space universality is claimed.

The empirical martingale estimate. Let $\mathcal{A}_t^{\varphi}$ be (52) with free streaming and the physical pair rule. For a compactly supported C^2 test, the finite matching calculation gives

$$\langle \varphi, \mu_t^{N,h} \rangle = \langle \varphi, \mu_0^N \rangle + \int_0^t \mathcal{A}_r^{\varphi}(\mu_r^{N,h})\, \mathrm{d}r + M_t^{N,h,\varphi} + R_t^{N,h,\varphi}, \tag{71}$$

$$\mathbb{E} \sup_{t \le T} |R_t^{N,h,\varphi}| \le C_{T,\varphi}(h + N^{-1}), \qquad \mathbb{E} \sup_{t \le T} |M_t^{N,h,\varphi}|^2 \le C_{T,\varphi}(h + N^{-1}). \tag{72}$$

Indeed,

$$0 \le \lambda - \frac{1 - e^{-h\lambda}}{h} \le \frac{h}{2}\lambda^2.$$

This bounds the gate contribution by a polynomial moment. The change from sampling another label to sampling the full empirical law costs $C_{\varphi} N^{-1}(1 + M_{\gamma})$ per unit time. Streaming changes positions by $h|v|$, and its empirical law by at most hM_1 in W_1. The time/state/law Lipschitz bounds on a, the C^1 bound on β, and the smooth-test Taylor estimate bound the integrated freezing and Riemann errors by $C_{\varphi} h$ times fixed polynomial moments. Moment products are also controlled: Jensen bounds their powers by higher moments in (65). For variance, conditional independence across matched pairs gives $C_{\varphi} h/N$ times moments, and the squared conditional matching mean is at most $C_{\varphi} h^2$ times squared moments. Summation and Doob prove (72).

The full nonlinear martingale identity. Let $Q^{N,h} = N^{-1} \sum_i \delta_{Z^{i,N,h}}$ be the empirical law of whole trajectories. Tightness of its laws follows from tagged tightness and exchangeability: a compact path set that has large tagged probability has small expected empirical mass outside, and a sequence of such sets yields tightness in the space of path probability measures. Take a joint subsequential limit of $Q^{N,h}$ and the empirical curves, denoted Q and m. The expected variation estimate in Lemma E.3, on shrinking deterministic intervals, shows that Q has no jump at any specified time, almost surely. One can use open oscillation neighborhoods and the portmanteau theorem to pass this bound. Thus time evaluation of Q is continuous at every fixed member of a countable dense set, with probability one. The joint limit has $Q_t = m_t$ there; right continuity extends the identification. Higher running moments pass by truncation and Fatou, so Q and m inherit all locally bounded polynomial moments.

To pass a conditional, rather than just an averaged, identity, fix $t_1, \dots, t_\ell \le s < t$ and a bounded continuous H of those past states. On discrete approximating times let $H_i = H(Z_{t_1}^i, \dots, Z_{t_\ell}^i)$ and $D_j^i = \varphi(Z_{j+1}^i) - \mathbb{E}[\varphi(Z_{j+1}^i) \mid \mathcal{F}_j]$. The moment estimates and bounded test imply

$$\mathbb{E}|D_j^i|^2 \le C_{\varphi} h, \qquad \mathbb{E}\left|\mathbb{E}[D_j^i D_j^k \mid \mathcal{F}_j]\right| \le C_{\varphi}(h/N + h^2), \quad i \ne k.$$

For the second estimate, the tags are directly paired with probability $1/(N-1)$ and an accepted shared event has probability at most $h\lambda$. Otherwise simultaneous increments require two accepted pair events, independent conditional on the matching; their contribution is $O(h^2)$ after the polynomial bounds. The product of conditional means has the same bound. Therefore martingale orthogonality across layers gives

$$\mathbb{E}\left|\frac{1}{N}\sum_i H_i \sum_{j:\, s\le t_j<t} D_j^i\right|^2 \le C_{H,\varphi,T}(N^{-1}+h).$$

The analogous averaged drift remainder tends to zero by the estimates used in (72).

On bounded states and bounded-moment sets the generator tests are continuous in the state and law. For nonzero relative velocities this follows from (62) and the multiplier regularity. At zero relative velocity the collision increment vanishes, so the integrated test is continuous even if B itself is not continuous there. Measurability in n causes no problem: pointwise-in-n convergence of the integrand, with a common bound on this truncated set, permits dominated convergence in the fixed mark law. Pass to the limit at almost every time of path continuity, then remove the state/moment truncations with the inherited higher moments. This gives, almost surely,

$$\int H(z_{t_1},\dots,z_{t_\ell})\left[\varphi(z_t)-\varphi(z_s)-\int_s^t \mathcal{L}_{r,m_r}\varphi(z_r)\,\mathrm{d}r\right] Q(\,\mathrm{d}z_\cdot) = 0.$$

Countable dense smooth test families, a countable algebra of past tests, and rational times, followed by the monotone-class argument and right continuity, give the nonlinear martingale problem with $m_r = Q_r$. This is the martingale method of Graham & Méléard (1997, Theorem 4.5) with the unbounded integrability checks made above.

Only now apply Theorem E.5 and Lemma E.7. Polynomial testing of this candidate is valid; pair energy is conserved and the generating lemma supplies its Gaussian moments. Nonlinear uniqueness makes Q the same deterministic path law on every subsequence. Its marginal curve is continuous in W_1 by the integrated jump and streaming variation bounds. Thus empirical curves converge uniformly in time in probability in W_1.

Expectation and uniform initial populations. For $X_{N,h} = \sup_{t\le T} W_1(\mu_t^{N,h}, \mu_t)$, comparison through a fixed zero-velocity state gives

$$X_{N,h} \le C_d + \sup_{t\le T} M_1(\mu_t^{N,h}) + \sup_{t\le T} M_1(\mu_t).$$

For a fixed $p \ge 4$, Jensen, (65), and (67) imply $\sup_{\mu_0,N,h} \mathbb{E}X_{N,h}^p < \infty$. This is uniform integrability, so convergence in probability upgrades to convergence of expectations.

The class $\mathcal{C}_{a_*,M_*}$ is compact in W_1: it is closed by lower semicontinuity and has uniformly integrable first-moment tails. Its iid empirical initialization converges in mean W_1 uniformly over the class. To check this directly, restrict to a large compact velocity region, divide it into M cells of diameter at most η, and use Cauchy–Schwarz and binomial variances: the expected sum of cell-frequency errors is at most $\sqrt{M/N}$. The transport cost on a region of diameter D_R is at most $2\eta + D_R\sqrt{M/N}$; the initial tail is uniformly small. Choose the region, then η, then N. If the asserted supremum over initial laws failed, a violating sequence would have a W_1-convergent subsequence of initial laws. The compactness and identification argument above has uniform constants and gives the nonlinear limit from that limiting initial law. Lemma E.7 gives uniform-in-time convergence of the varying target solutions to the same law, contradicting the violation. This proves the fixed-coefficient particle–depth limit used in (57). Exchangeability gives fixed-order marginal chaos, and Lemma E.8 supplies the conditional endpoint conclusion. This proves the fixed-coefficient statements and their extension under (61)–(62); neural transfer is established next.

F Shared coefficients and solution-operator approximation

The coefficients are the objects learned once and then reused. This appendix states what the approximation results imply about that reuse; it adds no training guarantee.

Certificate for the selective example. In (54), $|g| \le 1$ and $\mathrm{Lip}(g) \le 1$, so $|q| \le 1$. The integrand $e^{-1}e^{qg(u)}(1+g(u))/2$ lies in $[0,1]$. Its derivative with respect to q has magnitude at most one. As a function of the scalar key $g(u)$ its Lipschitz constant is at most $3/2$, by differentiating $e^{-1}e^{qk}(1+k)/2$ on $[-1,1]^2$. Consequently

$$|a(\mu,z,z_*) - a(\eta,\widetilde{z},\widetilde{z}_*)| \le \frac{1}{4}(|v-\widetilde{v}| + |w-\widetilde{w}|) + \frac{3}{4}W_1(\mu,\eta).$$

The multiplier lies in $[1/4, 3/4]$ and is symmetric in v, w. Clipping g is an exact finite ReLU map, the query and value maps are affine after clipping, and the exponential score is an admitted attention primitive. Thus this is an exact finite module. For example, contexts can retain positive mass at the same pair $v_1 = 1/2$, $w_1 = -1/2$ and move the remaining mass from $u_1 = -1$ to $u_1 = 1$. The query remains zero but the integral, and hence the rate of that proposed pair, changes.

Theorem F.1 (Shared modules on the controlled compact class)**.** *Suppose b, J are jointly continuous in all inputs and satisfy the assumptions of Appendix C with common constants. Then there are shared coefficient modules, with time as an input and fixed token embedding dimension, whose context endpoint kernels satisfy*

$$\sup_{\mu,z} \max_{j \le T/h} W_1(K^{h,\Theta}_{0,t_j}(\mu,z), K_{0,t_j}(\mu,z)) \le C_T(h + \delta_b + \delta_J),$$

where $\delta_b, \delta_J \to 0$ along the neural approximation sequence. For bounded regular pair coefficients with the same uniformity and common range, the empirical implementation has the corresponding uniform joint coefficient–particle–depth limit, provided its empirical initialization converges uniformly in mean W_1 to the requested initial law. In particular, iid initialization has this property on compact E.

Proof. Apply Proposition A.3 to the drift and the continuous sampler of Lemma B.2, treating time as a query input. Use the projected Euler stream and apply Lemma C.2 with common constants. Discrete Gronwall gives the uniform endpoint bound. The conditional-to-context inequality (60) follows by coupling the two output laws separately for each common ancestor and then integrating that coupling. For a pair coefficient sequence, insert target coefficients at the approximate state and empirical law, as following (27). Uniform errors and bounds verify the moment and martingale estimates of Theorem D.3. For iid initialization, a finite η-net of E with M cells gives a uniform empirical initialization error at most $2\eta + D\sqrt{M/N}$, where $D = \mathrm{diam}(E)$, by binomial variances and Cauchy–Schwarz. Taking $N \to \infty$ and then $\eta \downarrow 0$ verifies the stated initialization condition. If convergence were not uniform, compactness would give a convergent subsequence of initial laws; the same proof identifies the unique limiting target. Target stability yields a contradiction. □

Theorem F.2 (Neural transfer in the physical class)**.** *Hold β, B, and the reflection fixed as in Appendix E. Let $a_m \in [0, A]$ be exchange-symmetric coefficient modules converging to a uniformly in the mark n on each set*

$$\{0 \le t \le T,\ |v| + |w| \le R,\ M_p(\mu) \le R\}, \qquad p > 1,$$

for every finite R, T. Assume the targets satisfy the physical Lipschitz hypotheses with common constants (joint continuity in the mark is not needed for this transfer assertion). Then

$$\sup_{\mu_0 \in \mathcal{C}_{a_*,M_*}} \mathbb{E} \sup_{t \le T} W_1(\mu^{N,h,m}_t, \mu_t) \longrightarrow 0 \quad \text{as } m, N, h^{-1} \to \infty$$

along every joint sequence. The approximants need no common Lipschitz constant. Under additional joint continuity in the angular mark, such neural modules exist by bounded token preprocessing as in Appendix A.2.

Proof. All finite particle moment bounds use only the common envelope of a_m, and so are uniform in m. The one-step variance estimates have the same property. On bounded states and bounded-moment sets, write $a_m(\mu^N) - a(\mu)$ as $[a_m(\mu^N) - a(\mu^N)] + [a(\mu^N) - a(\mu)]$. The first term vanishes uniformly by hypothesis; the target Lipschitz constant controls the second. Streaming and time-freezing errors are estimated after this insertion, not by differentiating a_m. On the complement, use the common rate envelope and higher running moments. For example the event $M_p(\mu^N) > R$ has uniformly small contribution to any lower-order polynomial integrand by Markov, Cauchy–Schwarz, and a

sufficiently high moment bound. Large partner or tagged velocities are handled in the same way. Thus the drift remainders vanish without an m-uniform rate of convergence, while the martingales still vanish. Repeat the identification and expectation argument of Appendix E.7. This proves all uniformities. The existence clause follows from (28), the fixed-dimensional coefficient theorem, clipping, and exchange averaging; it uses mark continuity. It is logically separate from transfer for a supplied approximation sequence. □

F.1 Fully learned rates and polynomial envelopes

The main physical theorem is a specialization of the preceding limit and transfer results. It includes a bounded intensity learned in its entirety and uses a generic growth factor for unbounded rates. This section verifies that specialization; it changes neither the reflection map nor the analytic estimates in Appendix E.

Proposition F.3 (Verification of Theorem E.1)**.** *The conclusions of Theorem E.1 follow from the fixed-coefficient physical limit and Theorem F.2 by taking the reference spatial factor to be* 1 *and the reference cross section to be*

$$B_\gamma(u,n) = \langle u\rangle^\gamma, \qquad 0 \le \gamma \le 1.$$

In particular the bounded case learns the full intensity. For any fixed even nonnegative $\beta \in C^1(\mathbb{T}^d)$ *and any* $\nu_0 > 0$*, the Maxwell target* $\lambda^{\mathrm{M}} = \nu_0\beta(x-y)$ *satisfies the hypotheses with* $\gamma = 0$*. The hard-sphere target* $\lambda^{\mathrm{HS}} = \beta(x-y)|(v-w)\cdot n|$ *satisfies them with* $\gamma = 1$ *and* a^{HS} *as in* (58)*, after choosing a sufficiently large common output bound.*

Proof. The constant spatial reference has integral one on the normalized torus. The function B_γ is even in u and is preserved by elastic reflection, since reflection preserves $|u|$. It satisfies the growth bound in Appendix E. We check its weighted increment bound explicitly. Let $M = \max(|u|, |\widetilde{u}|)$ and $d = |u - \widetilde{u}|$. If $d \ge M/2$, then

$$M|\langle u\rangle^\gamma - \langle \widetilde{u}\rangle^\gamma| \le 2(\langle u\rangle^\gamma + \langle \widetilde{u}\rangle^\gamma)d.$$

If $d < M/2$, every point on the segment between the two arguments has norm at least $M/2$. For $0 < \gamma \le 1$, the gradient of $\langle u\rangle^\gamma$ has norm at most $\gamma\langle u\rangle^{\gamma-1}$. The mean-value bound, treating $M \le 1$ separately, gives

$$M|\langle u\rangle^\gamma - \langle \widetilde{u}\rangle^\gamma| \le C_\gamma(\langle u\rangle^\gamma + \langle \widetilde{u}\rangle^\gamma)d.$$

For $\gamma = 0$ the left side vanishes. These are exactly the cross-section conditions required by the physical proof. Its contextual multiplier is now a, so the target has the existence, uniqueness, and conditional regularity already proved there.

Joint target continuity includes the angular mark. The noncompact neural construction in Appendix A.2, applied on successively larger moment and state sublevels, gives bounded exchange-symmetric a_{ϑ_m} converging uniformly on every set used in Theorem F.2. A diagonal over integer state, moment, and time bounds yields one sequence for all finite horizons. Its token embedding dimension remains fixed by Proposition A.3. The transfer theorem gives (57), including its joint quantifiers and initial-law uniformity.

For Maxwell molecules the target multiplier is $a^{\mathrm{M}} = \nu_0\beta(x-y)$. It is bounded and globally Lipschitz in the state variables, independent of time, the context, and the direction. The neural intensity is a_ϑ itself. This gives one cutoff Maxwell angular law in the reflection convention; no assertion about every angular Maxwell kernel is needed.

For hard spheres write $H(u) = u/\langle u\rangle$. Its derivative is

$$DH(u) = \langle u\rangle^{-1}\mathrm{Id} - \langle u\rangle^{-3}u \otimes u,$$

whose operator norm is at most one. Hence $|H(u)\cdot n|$ is bounded by one and 1-Lipschitz in u, uniformly in $n \in S^{d-1}$. Multiplication by the fixed bounded Lipschitz spatial function $\beta(x-y)$ proves the required global state regularity of a^{HS}. It is continuous in the mark and invariant under exchange of the two states. Thus it belongs to the target class. The map a_ϑ approximates this normalized spatial and angular dependence; multiplying by $\langle v - w\rangle$ recovers the intended unbounded intensity in the limit. □

Equation (59) is the threshold form of the joint convergence result. Given a tolerance, there are common thresholds for m, N, L beyond which its error is smaller than that tolerance. Otherwise a violating sequence with all three indices tending to infinity would contradict (57). Fix one module beyond its threshold and then vary N, L beyond theirs. This proves the stated shared-module assertion.

For conditional context kernels, the tagged Gaussian estimate and localized event coupling likewise turn local uniform coefficient error into a vanishing forcing term in the Osgood comparison. Integration over ancestors gives (60). These arguments concern approximation of a prescribed coefficient family. No statistical procedure for identifying that family from data is asserted.

G Supporting extensions and boundaries of the characterization

The next results retain the nonconservative part of the framework. They are separate interfaces, not additional conclusions of the hard-sphere theorem.

G.1 An intrinsic noncompact finite-activity tangent

Proposition G.1 (Persistence and no loss of event mass to infinity)**.** *Let $E = \mathbb{R}^d$ or $\mathbb{T}^d \times \mathbb{R}^d$. Fix time, background, and state z. Let B be linear on $C_0(E) + \mathbb{R}1$, with $B1 = 0$, and fix a finite $a \geq 0$. For $0 \leq \chi_R \leq 1$ in $C_c(E)$ increasing to one, a positive measure q of mass a represents $B\varphi(z) = \int [\varphi(y) - \varphi(z)] q(\,\mathrm{d}y)$ if and only if*

$$B\varphi(z) + a\varphi(z) \geq 0 \quad (\varphi \geq 0), \qquad B\chi_R(z) + a\chi_R(z) \longrightarrow a.$$

The representing measure is unique. Locally uniform tightness and continuity of these scalar functionals give narrow continuity in the parameters; first-moment continuity further needs locally uniform first-moment tails.

Proof. $C_0(E) + \mathbb{R}1$ is $C(E_\infty)$ on the one-point compactification. The positive functional $\ell(\varphi) = B\varphi(z) + a\varphi(z)$ has mass a, so Riesz represents it by a unique measure on E_∞. The cutoff condition says by monotone convergence that all its mass is on E. This proves existence and the converse. Parameter continuity first gives convergence against C_0; locally uniform tightness and total mass convergence upgrade it to narrow convergence. The asserted first-moment upgrade is the usual truncation argument. This is a tangent representation, not a nonlinear nonexplosion or uniqueness theorem. □

G.2 A nonconservative polynomial-rate class

High rates can also be controlled by small event amplitudes, rather than by conservation. We record the sufficient class from the general framework. It does not include hard-sphere amplitudes. Let $E = \mathbb{R}^d$, $\nu(\Xi) = 1$, $q \geq 0$, and $p > \max\{2, q\}$. Suppose

$$0 \leq \lambda_t(\mu, z, w, \xi) \leq C_\lambda(1 + |z|^q + |w|^q), \quad \ell = \log(1 + \lambda), \quad \Delta^a = \frac{g^a}{1 + \lambda}, \quad a = 1, 2. \tag{73}$$

Assume g^a are bounded by G, and g^a, ℓ are globally Lipschitz in time, both states, and W_1 context, uniformly in the mark. The rate and pair map $(z + \Delta^1, w + \Delta^2)$ satisfy the same marked exchange symmetry as in Appendix D; in particular, exchanging the pair leaves the rate unchanged and exchanges the outputs. The drift is globally Lipschitz in state and law, uniformly Lipschitz in time, with $\sup_{t \leq T} |b_t(\delta_0, 0)| < \infty$.

Lemma G.2 (Rate-weighted amplitude control)**.** *For two coefficient inputs x, y and $J^a_x(u) = \Delta^a_x \mathbf{1}_{u \leq \lambda_x}$,*

$$\int_0^\infty |J^a_x(u) - J^a_y(u)| \,\mathrm{d}u \leq |g^a_x - g^a_y| + 2G|\ell_x - \ell_y|.$$

Moreover $\lambda|\Delta^a| \leq G$ and $\lambda|\Delta^a|^2 \leq G^2$. The integrated bound remains valid after restriction to $u \leq R$.

Proof. For $\lambda_x \geq \lambda_y$, the integral is $\lambda_y|g_x/(1+\lambda_x) - g_y/(1+\lambda_y)| + (\lambda_x - \lambda_y)|g_x|/(1+\lambda_x)$. Insert $g_y/(1+\lambda_x)$ and bound by $|g_x - g_y| + 2G(\lambda_x - \lambda_y)/(1+\lambda_x)$. The ratio is at most $\log(1+\lambda_x) - \log(1+\lambda_y)$. The reverse ordering is identical. The other bounds are immediate, and restriction decreases a nonnegative integral. □

Theorem G.3 (A rate-compensated nonconservative limit)**.** *Under* (73)*, every initial* $\mu_0 \in \mathcal{P}_p(\mathbb{R}^d)$ *generates a unique nonexplosive nonlinear process, with*

$$\mathbb{E}\sup_{t\leq T}\langle Z_t\rangle^p \leq C_T(1 + \langle\langle z\rangle^p, \mu_0\rangle), \qquad W_1(\mu_t, \eta_t) \leq e^{C_\tau t} W_1(\mu_0, \eta_0).$$

The saturated stream–matching scheme with Euler drift and iid initialization has the same qualitative joint $N \to \infty$, $h \to 0$ *empirical limit uniformly in time on each finite horizon, in probability in* W_1. *The finite continuous-time systems are nonexplosive and also propagate fixed-time chaos.*

Proof. For $V_p(z) = \langle z\rangle^p$ and $|a| \leq G$, the mean value theorem gives $[V_p(z+a) - V_p(z)]_+ \leq C_{p,G}\langle z\rangle^{p-1}|a|$. Lemma G.2 bounds the rate-integrated positive variation by $C_{p,G}\langle z\rangle^{p-1}$. The drift has growth $C(1 + |z| + \langle|z|, \mu\rangle)$; Young's inequality gives a moment bound of linear Gronwall type after expectation. Stopping at state and jump-count levels, then summing positive jump variation and absolute drift variation, proves the claimed running-moment bound uniformly in the stops. The same averaged argument for particles, followed by the tagged-coordinate argument, is uniform in N. Euler drift and $p_h \leq h\lambda$ give the identical discrete bound.

For construction truncate the acceptance coordinate to $u \leq R$, keeping the original amplitude $g/(1+\lambda)$. Finite clocks and a frozen-background Picard argument construct the truncated process; Lemma G.2 makes its coupling constant independent of R. The unrepresented jump variation above R is at most $G\mathbf{1}_{\lambda>R}$. The polynomial rate envelope and the moment bound imply $\int_0^T \mathbb{E}\mathbf{1}_{\lambda_t>R}\,\mathrm{d}t \leq C_T/R$. Synchronous coupling and Gronwall make the cutoff solutions Cauchy in expected uniform path distance. To realize the limit, select a subsequence with successive expected coupling costs at most 2^{-j} and glue these couplings on the Polish Skorokhod path space. Their uniform distances are almost surely summable. The resulting paths converge uniformly to a right-continuous path with left limits; uniform convergence also implies Skorokhod convergence, so this construction has a measurable path law. Its limit solves the original martingale problem by the integrated increment bound; the same coupling gives uniqueness and stability. Since $q < p$, its expected integrated accepted rate is finite. For fixed N the analogous total-rate expectation is finite, proving nonexplosion.

For the discrete limit, the gate discrepancy against a Lipschitz test is controlled by

$$\left(\lambda - \frac{1 - e^{-h\lambda}}{h}\right)|\Delta^a| \leq \frac{hG}{2}\lambda.$$

Its expectation tends to zero using the qth moment. The integrated first and second jump moments are bounded by G, G^2; conditional matching variance and the without-replacement bias thus have the bounds of Appendix D. The integrated observable is Lipschitz in its inputs: for a C_b^2 test, comparing its two increments gives

$$\|\nabla\varphi\|_\infty|J_x - J_y| + \mathrm{Lip}(\nabla\varphi)|z - w|\,|J_y|,$$

whose acceptance-coordinate integral is controlled by Lemma G.2. Euler consistency is integrable because $p > 2$. Running moments give compact containment and uniform integrability; the expected jump variation on a stopping interval is at most $G(\delta + h)$. The drift variation is bounded by the same running moments. Tightness, the empirical martingale identity, and uniqueness identify the limit exactly as in Theorem D.3, using moment truncation instead of compactness of E. Uniform-in-time weak convergence and first-moment uniform integrability give the stated W_1 conclusion. The argument without time steps proves the continuous-particle limit. This is a verified sufficient interface, not an optimal replacement of general McKean–Vlasov jump theory (Andreis et al., 2018; Erny, 2022). □

G.3 What the present closure does not assert

A bound on the polynomial growth of a rate alone neither gives nonexplosion nor supplies a stable nonlinear evolution. The two noncompact mechanisms above are different: rate-compensated

amplitudes in Theorem G.3, and conservative physical reflections in Theorem E.1. We do not transfer the latter's stability to arbitrary conservative maps. Similarly, an unbounded sequence of activity rates can generate a diffusion. For example, jumps $\pm\sqrt{h}$ each with probability $1/2$ at successive time steps have generator $[\varphi(z+\sqrt{h})+\varphi(z-\sqrt{h})-2\varphi(z)]/(2h) \to \varphi''(z)/2$. On a circle the same construction is a conservative angular small-jump limit. A finite-rate derivative characterization must therefore state its activity control.

For Boltzmann, fixed spatial localization makes the phase-space law an admissible empirical context. The functional $f \mapsto f(x,\cdot)$ for exact locality is not continuous in ordinary phase-space Wasserstein topology (Graham & Méléard, 1997, Section 1). Removing localization needs additional density estimates, not an informal substitution $\beta = \delta_0$. Noncutoff kernels require their own small-jump theory; see, for the spatially homogeneous case, Heydecker (2022). Neither extension is a premise of our results.

Equation (22) gives eventwise momentum and energy conservation. The correction below supplies a separate equilibrium guarantee for an optional collision block. Its acceptance changes the transition law, so it is not part of the algorithm in Theorems B–E.1. Free transport preserves a spatially uniform kinetic equilibrium but has adjoint given by backward transport. Ordinary detailed balance of the collision block is consequently distinct from the time-reversal properties of a complete stream–collision layer.

H An optional parallel detailed-balance correction

We specify the finite-particle structural option in Section 4. All statements concern the jump stage at fixed time and fixed network weights. The correction is an auxiliary-variable Metropolis step (Tierney, 1998); it uses the proposed pre- and post-jump clouds and one additional binary decision.

Let $\mathbf{z} = ((x_i, v_i))_{i=1}^N \in E^N$ and define $\mu_{\mathbf{z}}^N = N^{-1}\sum_i \delta_{(x_i,v_i)}$. The subscript specifies the indexed configuration at which the empirical context is evaluated. Let Π_N be the product of uniform positions on the torus and uniform surface measure on a nondegenerate velocity shell at fixed total momentum and energy. The same argument applies to uniform positions and independent velocities from one Maxwellian. Draw a matching M independently of $\mathbf{z}$. For its pairs use the maps (22), with matrix inputs restricted to the pair invariants $I = (c, |g|)$, and assume $\Omega_\eta(I, -\xi) = -\Omega_\eta(I,\xi)$ and $(-\mathrm{Id})_{\#}\nu = \nu$. These restrictions make each fixed-mark pair map preserve Π_N and give its inverse at the negative mark. The intensity can still read the entire context.

H.1 A single-pair reverse-context symmetrization

For one uniformly selected unordered pair $e = \{i,\ell\}$, let $\mathbf{z}_e^\star = \mathcal{T}_{e,\xi}\mathbf{z}$ denote the configuration obtained by applying the fixed-mark pair map, before drawing its gate. Only the two selected coordinates change. For a nonnegative raw rate $\widetilde{\lambda}_\vartheta$, set

$$\begin{aligned}\lambda_N(\mathbf{z}, e, \xi) = \tfrac{1}{2}\big[&\widetilde{\lambda}_\vartheta(\mu_{\mathbf{z}}^N, z_i, z_\ell, \xi) \\ &+ \widetilde{\lambda}_\vartheta(\mu_{\mathbf{z}_e^\star}^N, z_i^\star, z_\ell^\star, -\xi)\big].\end{aligned} \tag{74}$$

Assume pair exchange symmetry, as for the other finite implementations. The marked map $(\mathbf{z},\xi) \mapsto (\mathbf{z}_e^\star, -\xi)$ is an involution preserving $\Pi_N(\,\mathrm{d}\mathbf{z})\nu(\,\mathrm{d}\xi)$: applying the negative mark restores both selected states, and the spectators never change. Thus $\lambda_N(\mathbf{z}_e^\star, e, -\xi) = \lambda_N(\mathbf{z}, e, \xi)$. The same identity holds after applying either admissible gate p_h. The accepted flux is consequently symmetric, and the rejected flux lies on the diagonal. Averaging over the state-independent pair selection proves reversibility of this single-pair collision block with respect to Π_N. In particular, the reverse context in (74) is the proposed context, irrespective of whether the gate subsequently accepts.

For a quantitative comparison, suppose additionally that all states and pair outputs lie in a compact set of diameter D, and that $\widetilde{\lambda}_\vartheta$ is L_μ-Lipschitz in its context, uniformly in the other inputs. Let λ^{fix} denote (23) evaluated at the empirical context: its reverse term uses $\mu_{\mathbf{z}}^N$ instead of $\mu_{\mathbf{z}_e^\star}^N$. Label coupling gives

$$W_1(\mu_{\mathbf{z}_e^\star}^N, \mu_{\mathbf{z}}^N) \le \frac{d_E(z_i, z_i^\star) + d_E(z_\ell, z_\ell^\star)}{N} \le \frac{2D}{N}, \qquad |\lambda_N - \lambda^{\mathrm{fix}}| \le \frac{L_\mu D}{N}.$$

Since $|p_h(a) - p_h(b)| \le h|a-b|$ for both gate conventions (on the admissible range for the linear gate), the difference of their single-pair empirical drifts, divided by the physical duration $\Delta = 2h/N$, is at most $L_\mu D^2 \operatorname{Lip}(\varphi)/N$ for a Lipschitz test φ. Indeed the empirical increment has magnitude at most $2D \operatorname{Lip}(\varphi)/N$ and only its gate probability changes. Under the bounded compact hypotheses of Corollary D.2, this vanishing drift change and the same moment and variance bounds preserve the kinetic limit. On unbounded state spaces the first, displacement-dependent bound remains valid; a uniform $O(N^{-1})$ assertion requires additional moment control.

H.2 The parallel proposal correction

For each edge $e = \{i, \ell\} \in M$, set $p_e(\mathbf{z}, \xi_e) = p_h(\lambda_\vartheta(\mu_{\mathbf{z}}^N, z_i, z_\ell, \xi_e))$. Draw $\xi_e \sim \nu$ and a Bernoulli outcome $a_e \in \{0, 1\}$ with this probability, independently across edges conditional on $\mathbf{z}, M$. Write $\boldsymbol{a} = (a_e)_{e\in M}$ and $\boldsymbol{\xi} = (\xi_e)_{e\in M}$. Relative to $\nu^{\otimes|M|}$ and counting measure on the binary outcomes, their joint density is

$$q_\Theta(\boldsymbol{a}, \boldsymbol{\xi} \mid \mathbf{z}, M) = \prod_{e\in M} p_e(\mathbf{z}, \xi_e)^{a_e}[1 - p_e(\mathbf{z}, \xi_e)]^{1-a_e}. \tag{75}$$

A factor with exponent zero is one, including when its base is zero. Actual auxiliary draws have positive density almost surely. Let $\mathcal{T}_{\boldsymbol{a},\boldsymbol{\xi}}\mathbf{z}$ apply the pair maps on precisely the edges with $a_e = 1$ and leave all other coordinates fixed, and write $\mathbf{z}' = \mathcal{T}_{\boldsymbol{a},\boldsymbol{\xi}}\mathbf{z}$. Accept this whole proposal with the additional probability

$$\alpha(\mathbf{z}, \boldsymbol{a}, \boldsymbol{\xi}) = 1 \wedge \frac{q_\Theta(\boldsymbol{a}, -\boldsymbol{\xi} \mid \mathbf{z}', M)}{q_\Theta(\boldsymbol{a}, \boldsymbol{\xi} \mid \mathbf{z}, M)}. \tag{76}$$

Otherwise retain $\mathbf{z}$. The reverse evaluation uses $\mu_{\mathbf{z}'}^N$ and includes the probabilities of rejected pairs as well as accepted ones. For a fixed M we abbreviate the densities by $q_{\mathbf{z}}$ and $q_{\mathbf{z}'}$ below.

Proposition H.1 (Reversibility of the corrected matching block)**.** *The algorithm which draws $(M, \boldsymbol{\xi}, \boldsymbol{a})$, proposes $\mathbf{z}'$, and accepts it with* (76) *is reversible with respect to Π_N. The matching distribution must be independent of $\mathbf{z}$. A rejection returns $\mathbf{z}$. The proposal, forward-rate evaluation, and reverse-rate evaluation can each be carried out in parallel across the matched pairs.*

Proof. Fix M. Each pair map preserves its center of mass and relative speed, and rotates its relative direction by an orthogonal matrix. Conditional on all other velocities and these pair invariants, Π_N has uniform relative direction. Thus every fixed-mark pair map preserves Π_N. Disjoint pair maps commute, and the negative marks are their inverses. Consequently the map

$$\mathcal{R} : (\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{a}) \longmapsto (\mathcal{T}_{\boldsymbol{a},\boldsymbol{\xi}}\mathbf{z}, -\boldsymbol{\xi}, \boldsymbol{a})$$

is an involution preserving the reference measure $\Pi_N(\,\mathrm{d}\mathbf{z})\nu^{\otimes|M|}(\,\mathrm{d}\boldsymbol{\xi})$ times counting measure in $\boldsymbol{a}$. The accepted joint density relative to this reference is

$$q_{\mathbf{z}}(\boldsymbol{a}, \boldsymbol{\xi})\alpha(\mathbf{z}, \boldsymbol{a}, \boldsymbol{\xi}) = \min\{q_{\mathbf{z}}(\boldsymbol{a}, \boldsymbol{\xi}), q_{\mathbf{z}'}(\boldsymbol{a}, -\boldsymbol{\xi})\}.$$

It is invariant under $\mathcal{R}$. Pushing forward by $(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{a}) \mapsto (\mathbf{z}, \mathbf{z}')$ therefore gives a symmetric accepted flux. The rejection flux is supported on $\mathbf{z}' = \mathbf{z}$ and is symmetric as well. This proves detailed balance at fixed M; averaging over a state-independent matching distribution preserves it.

All reverse probabilities, including the rejection factors for pairs with $a_e = 0$, are evaluated at $\mu_{\mathbf{z}'}^N$. Those factors can change when other pairs move, which is why they must be retained. Only the proposed pre- and post-collision clouds are needed; no sequential replay or context refresh between pairs occurs. This also establishes the stated parallel implementation. □

For a prescribed equilibrium with nonconstant density $w(\mathbf{z})$ relative to the above invariant reference, the ratio additionally contains $w(\mathbf{z}')/w(\mathbf{z})$ on its positive-density region, with the usual zero-flux convention. The same minimum-flux proof applies. If one instead uses a jump map that does not preserve the reference measure, its Jacobian or reference Radon–Nikodym factor must also enter. None of these factors can be inferred solely from preservation of total energy.

The correction guarantees equilibrium preservation at every finite particle number. It can reject many otherwise accepted pair proposals at once; neither its efficiency nor an unchanged kinetic generator is claimed. For physical free streaming, composition with this collision kernel still preserves the stated spatially uniform equilibrium. A learned stream has that property only after its own invariant-measure condition has been verified.

H.3 Microscopic reversibility and the kinetic entropy identity

We give the additional, conditional entropy statement used in Section 4. It concerns a sufficiently regular density solution of a specified symmetric collision law. The general approximation theorems do not assume these extra symmetries. The Maxwell specialization uses $d_v = 3$. Let $E = \mathbb{T}^{d_v} \times \mathbb{R}^{d_v}$ with normalized spatial volume and let $M(x, v) = M(v)$ be a positive Maxwellian of unit total mass. For a density f write $\mu_f(\mathrm{d}z) = f(z)\,\mathrm{d}z$ and $H(f \mid M) = \int_E f \log(f/M)\,\mathrm{d}z$. A coefficient with argument $[f]$ is evaluated at this measure. With the network parameters fixed, $\Phi_\xi[f](z, z_*)$ denotes the full pair map $\Phi_\eta(\mu_f, z, z_*, \xi)$, and $\lambda_t[f]$ the corresponding possibly time-dependent contextual intensity. Write $z = (x, v)$ and $z_* = (y, w)$. For each fixed time and density f, let $(z', z'_*) = \Phi_\xi[f](z, z_*)$, where $z' = (x, v')$ and $z'_* = (y, w')$, so positions are unchanged. Assume the map

$$\mathscr{R}_f : (z, z_*, \xi) \longmapsto (z', z'_*, -\xi)$$

is an involution preserving $\mathrm{d}z\,\mathrm{d}z_*\nu(\mathrm{d}\xi)$. Assume also pair exchange symmetry, $M(v)M(w) = M(v')M(w')$, and the forward–reverse rate identity

$$\lambda_t[f](z, z_*, \xi) = \lambda_t[f](z', z'_*, -\xi). \tag{77}$$

The invariant-input rotations (22) with $\Omega_\eta(I, -\xi) = -\Omega_\eta(I, \xi)$, a mark law invariant under sign reversal, and rate symmetrization (23) supply these conditions. The continuum population f is the same argument on both sides of (77). The pair map fixes center and relative speed and rotates uniform angular measure. In these coordinates it preserves volume; allowing its rotation matrix to depend on the fixed invariants does not change the angular Jacobian. The same proof applies with any measure-preserving involution of the mark space in place of $\xi \mapsto -\xi$. For the Maxwell model the involution is $R \mapsto R^{-1}$ on $SO(3)$, which preserves Haar measure. Its position-only intensity is unchanged by this reversal.

Proposition H.2 (Entropy dissipation for a reversible learned collision law)**.** *Suppose $f_t > 0$ is a differentiable density solution with free flight and weak collision equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi f_t = \int v \cdot \nabla_x \varphi\, f_t + \frac{1}{2}\int \lambda_t[f_t] f_t f_{t,*}[\varphi(z') + \varphi(z'_*) - \varphi(z) - \varphi(z_*)]\,\mathrm{d}z\,\mathrm{d}z_*\nu(\mathrm{d}\xi).$$

Assume the collision logarithms and transport terms below are integrable, differentiation of $H(f_t \mid M)$ is legitimate, and integration by parts in x has no boundary contribution. Under the preceding symmetries,

$$\frac{\mathrm{d}}{\mathrm{d}t}H(f_t \mid M) = -\mathcal{D}(f_t) = -\frac{1}{4}\int \lambda_t[f_t](f_t f_{t,*} - f'_t f'_{t,*}) \log \frac{f_t f_{t,*}}{f'_t f'_{t,*}}\,\mathrm{d}z\,\mathrm{d}z_*\nu(\mathrm{d}\xi) \le 0.$$

In particular M is stationary for the collision equation and free flight.

Proof. Conservation of mass removes the constant in the derivative of $f \log(f/M)$. The transport contribution is $\int v \cdot \nabla_x f = 0$ on the torus, since M is independent of x. At a fixed time abbreviate $a = f(z)f(z_*)$ and $b = f(z')f(z'_*)$. Pair exchange and $MM_* = M'M'_*$ make the collision contribution

$$I = \frac{1}{2}\int \lambda_t[f] a \log(b/a)\,\mathrm{d}z\,\mathrm{d}z_*\nu(\mathrm{d}\xi).$$

Change variables by the measure-preserving involution $\mathscr{R}_f$. The two products exchange, and the rate is unchanged, so also $I = -\frac{1}{2}\int \lambda_t[f] b \log(b/a)$. Averaging these identities gives $I = -\frac{1}{4}\int \lambda_t[f](a - b)\log(a/b)$. For $a, b > 0$, $(a - b)\log(a/b) \ge 0$. This proves the assertion. The density dependence of $\lambda_t[f]$ is evaluated at this time; differentiating the entropy functional requires no derivative of the rate with respect to f. At $f = M$, the incoming and outgoing products agree, giving stationarity by the same change of variables. □

This is the classical Boltzmann symmetrization applied to a learned rate satisfying the displayed identities; it is not inferred from propagation of chaos. For singular initial measures, or weak solutions for which the logarithmic test is unavailable, an entropy inequality requires its own approximation and lower-semicontinuity argument. No such extension is included in this proposition.

Spatially varying solutions and the choice of reference. Nothing in Proposition H.2 requires f_t to be spatially uniform. Its Boltzmann entropy $\mathcal{H}(f) = \int_E f \log f$ satisfies $\mathrm{d}\mathcal{H}(f_t)/\mathrm{d}t = -\mathcal{D}(f_t)$ by the same proof with test $\log f_t$. For a spatially uniform Maxwellian $M(v)$, the quantity $\log M(v)$ is a constant plus a linear and a quadratic function of v. Its integral is conserved, explaining why the same dissipation holds for $H(f_t \mid M)$.

For comparison, let $\widetilde{M}(x, v) > 0$ be a fixed smooth reference density and assume the following tests are integrable. The weak equation gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} H(f_t \mid \widetilde{M}) = & -\mathcal{D}(f_t) - \int_E f_t v \cdot \nabla_x \log \widetilde{M} \, \mathrm{d}z \\ & - \frac{1}{2} \int_{E^2 \times \Xi} \lambda_t[f_t] f_t f_{t,*} \log \frac{\widetilde{M}' \widetilde{M}'_*}{\widetilde{M}\widetilde{M}_*} \, \mathrm{d}z \, \mathrm{d}z_* \nu(\mathrm{d}\xi).\end{aligned} \tag{78}$$

Indeed, subtract the weak-equation derivative of $\int f_t \log \widetilde{M}$ from that of $\mathcal{H}(f_t)$. Thus the same relative-entropy identity holds for any reference satisfying both $v \cdot \nabla_x \widetilde{M} = 0$ and $\widetilde{M}\widetilde{M}_* = \widetilde{M}'\widetilde{M}'_*$. A general local Maxwellian with position-dependent mean or temperature has neither property for the present transport and nonlocal pair interactions. For spatially local elastic collisions at one common position, the product identity does hold for a local Maxwellian, but its transport term still requires separate treatment. This distinguishes an inhomogeneous solution from an inhomogeneous equilibrium reference.

H.4 The finite-particle entropy statement and the Maxwell construction

Let P_N be a Markov collision kernel preserving the probability equilibrium Π_N. For a probability law $\rho_N \ll \Pi_N$, define $H(\rho_N \mid \Pi_N) = \int \log(\mathrm{d}\rho_N / \mathrm{d}\Pi_N) \, \mathrm{d}\rho_N$, and set it to $+\infty$ otherwise. It is the entropy of the joint law on E^N. For the elastic blocks below, Π_N is the product of uniform positions and either uniform surface measure on a nondegenerate velocity shell of fixed total momentum and energy, or independent velocities with one common Maxwellian law. These are the equilibria specified in Appendix H.

Corollary H.3 (Entropy contraction for an invariant particle block)**.** *For any probability law ρ_N with $H(\rho_N \mid \Pi_N) < \infty$,*

$$H(\rho_N P_N \mid \Pi_N) \le H(\rho_N \mid \Pi_N).$$

This applies to the corrected collision block of Proposition H.1. It also applies, without that correction, to the Maxwell collision rounds of Theorem I.1 with intensity $r\beta(x - y)$, and to their neural substitutions whose admissible gate probabilities depend only on the unchanged pair positions and the frequency input. This finite-round claim holds at every step size for which the gates are probabilities. Composing either collision block with exact free streaming preserves its entropy inequality for the stated uniform-position equilibrium.

Proof. Use the stationary joint law $\Pi_N(\mathrm{d}\mathbf{z}) P_N(\mathbf{z}, \mathrm{d}\mathbf{y})$ on pairs of configurations in E^N. If $r = \mathrm{d}\rho_N / \mathrm{d}\Pi_N$, the output density is $\widetilde{r}(\mathbf{y}) = \mathbb{E}[r(\mathbf{z}) \mid \mathbf{y}]$ under this joint law. Conditional Jensen for $s \log s$ yields $\int \widetilde{r} \log \widetilde{r} \, \mathrm{d}\Pi_N \le \int r \log r \, \mathrm{d}\Pi_N$. The corrected kernel is invariant by Proposition H.1. For the Maxwell round, fix the positions and a state-independent matching. The probability p_e of accepting an edge e is unchanged by its collision and independent of its Haar rotation. Conditional on the spectator velocities and the pair's center and relative speed, both choices of Π_N have uniform relative direction. A fixed rotation preserves this conditional law, and its inverse has the same Haar law. Consequently the Haar-averaged pair kernel Q_e is reversible, as is $(1 - p_e)\mathrm{Id} + p_e Q_e$. Kernels on disjoint edges commute, so their product is reversible; averaging over the matching retains this property. Exact free streaming is an invertible measure-preserving map for uniform positions and the same velocity equilibrium. It preserves relative entropy, which proves the claim for the composed block. This last composition need not satisfy ordinary detailed balance: backward transport is its transport adjoint. □

The position-only restriction is used for this direct invariance proof. Uniform approximation to $r\beta(x - y)$ by a contextual neural intensity alone does not give exact finite-particle equilibrium preservation: the empirical context changes under the reverse collision. Such gates instead require the separate conditions or corrections in Appendix H.

I THE FINITE-FREQUENCY MAXWELL–BOLTZMANN REALIZATION

We give the precise physical-space result used in Section 4. Take $E = \mathbb{T}^3 \times \mathbb{R}^3$, with the product cost specified there, and assume (24)–(26). Fix a velocity mean $u \in \mathbb{R}^3$, centered energy $c_E > 0$, and $M_4 \geq 2(\beta_{\max}/\beta_{\min})c_E^2$. Define

$$\mathcal{K} = \left\{ f \in \mathcal{P}_2(E) : \int v \, \mathrm{d}f = u, \quad \int |v-u|^2 \, \mathrm{d}f = c_E, \quad \int |v-u|^4 \, \mathrm{d}f \leq M_4 \right\}. \tag{79}$$

The spatial law may be singular and arbitrarily correlated with velocity. Write $b_0 = \beta_{\min}$, $b_1 = \beta_{\max}$, $B = b_1/b_0$, and $M = M_4$. All convergence statements below are uniform on fixed bounded frequency intervals; no infinite-frequency limit or collision-compression claim is made.

Theorem I.1 (Maxwell realization on the physical phase space)**.** *For every finite $r \geq 0$ and $f \in \mathcal{K}$, the generator* (26) *defines a unique nonlinear process in law. The population law remains in the W_2-compact class $\mathcal{K}$, and the tagged kernels $K_t^{r,\beta}(f,z)$ are jointly W_2-continuous in (r,t,f,z) on $[0,R_*] \times [0,T] \times \mathcal{K}$ and bounded tagged-velocity sets, for every $R_*, T < \infty$.*

Let $K_{t,h}^{r,\beta}(f,z)$ and $S_{t,h}^{r,\beta} f$ be the tagged and population laws of the stream–collision recursion defined below. At every fixed $t \geq 0$, as $h \downarrow 0$, they converge in W_2 to $K_t^{r,\beta}(f,z)$ and $S_t^{r,\beta} f$, uniformly over $r \in [0,R_*]$, $f \in \mathcal{K}$, and bounded tagged velocities.

The same limits hold jointly as $h \downarrow 0$ and the uniform coefficient error tends to zero when $r\beta(x-y)$ is replaced by nonnegative neural rates $\lambda_\vartheta(r,x,y) = \lambda_\vartheta(r,y,x)$, uniformly in $r \in [0,R_*]$ and $x,y \in \mathbb{T}^3$. Such approximants exist. Their continuous-time laws converge as the coefficient error tends to zero as well. The original fourth-moment threshold need not be invariant for these neural rates; a common finite-horizon bound suffices.

The following moment estimates, Lemma I.2, and Appendix I.4 prove the theorem.

I.1 THE PHYSICAL AND DISCRETE LAWS

Let $E = \mathbb{T}^3 \times \mathbb{R}^3$, with the product distance in the main text. For $R \in SO(3)$ put

$$c = \frac{v+w}{2}, \qquad g = \frac{v-w}{2}, \qquad \mathcal{V}_R^1(v,w) = c + Rg, \qquad \mathcal{V}_R^2(v,w) = c - Rg. \tag{80}$$

These symbols denote the two outgoing velocities; the full pair map retains the positions as in (24). The generator is (26). In its tagged process the background is $S_s^{r,\beta} f$, independently of the chosen tagged start z; hence

$$K_t^{r,\beta}(f,z) = \mathrm{Law}(Z_t \mid Z_0 = z; \text{background } S_s^{r,\beta} f), \qquad S_t^{r,\beta} f = \int K_t^{r,\beta}(f,z) f(\,\mathrm{d}z).$$

For a step $h > 0$, first stream both input states by $(x,v) \mapsto (x+hv, v)$, draw an independent partner from the streamed population, and accept the Haar collision with probability

$$p_{r,h}(x-y) = 1 - e^{-rh\beta(x-y)}, \tag{81}$$

where x, y are the streamed positions. This defines the population recursion and, using the same backgrounds, its separate tagged chain. We write $S_{t,h}^{r,\beta} f$ and $K_{t,h}^{r,\beta}(f,z)$ for their laws after n steps when $t = nh$. A last step of length less than h can be used for an arbitrary reporting time.

Pair exchange leaves the rate and mark law invariant. Each accepted pair conserves its momentum and kinetic energy exactly, while free flight leaves its velocities unchanged. Symmetrizing the weak equation therefore preserves the population mean u and centered energy c_E; the same argument applies to each population step. We may subtract u from velocities and translate positions by $-tu$, since β depends only on position differences. The moment calculations below use this co-moving frame.

I.2 Fourth moments and compactness

For a Haar rotation, direct spherical averaging gives

$$\mathbb{E}_R \mathcal{V}_R^1(v,w) = \tfrac{1}{2}(v+w), \qquad \mathbb{E}_R|\mathcal{V}_R^1(v,w)|^2 = \tfrac{1}{2}(|v|^2+|w|^2), \tag{82}$$

$$\mathbb{E}_R|\mathcal{V}_R^1(v,w)|^4 = \tfrac{1}{3}(|v|^2+|w|^2)^2 - \tfrac{1}{3}(v\cdot w)^2. \tag{83}$$

Indeed, $\mathbb{E}_R(c\cdot Rg)^2 = |c|^2|g|^2/3$. Adding the two fourth moments and discarding a nonpositive term yields

$$\mathbb{E}_R\left[|v'|^4+|w'|^4-|v|^4-|w|^4\right] \le -\tfrac{1}{3}(|v|^4+|w|^4)+\tfrac{4}{3}|v|^2|w|^2. \tag{84}$$

Write $M_4(f) = \int |v-u|^4 f(\,\mathrm{d}x,\,\mathrm{d}v)$. The pair-symmetrized weak equation has a factor $1/2$. Applying the lower rate bound to the negative term in (84) and the upper bound to its positive term gives

$$\frac{\mathrm{d}}{\mathrm{d}t}M_4(f_t) \le -\frac{rb_0}{3}M_4(f_t) + \frac{2rb_1}{3}c_E^2. \tag{85}$$

Streaming does not change this moment. Thus $M_4(f_t) \le \max\{M_4(f_0), 2Bc_E^2\} \le M$, proving the invariance of $\mathcal{K}$ together with conservation of its first two velocity moments.

For $rh > 0$ set $p_0 = 1-e^{-rhb_0}$ and $p_1 = 1-e^{-rhb_1}$. Concavity of $a \mapsto 1-e^{-a}$, with value zero at zero, gives $p_1/p_0 \le B$. Consequently the identical pair calculation gives

$$M_4(f_{j+1}) \le (1-p_0/3)M_4(f_j) + (2Bp_0/3)c_E^2. \tag{86}$$

The saturated discrete populations also remain in $\mathcal{K}$. For $r = 0$ they undergo free flight only. The linear-gate comparison scheme used below has the same estimate with $p_0 = hrb_0$, provided $hrb_1 \le 1$.

A separate tag does not conserve its own energy, but its background has the preceding bounds. By (83), its fourth-moment collision drift, at velocity v in the co-moving frame, is at most

$$-\frac{2rb_0}{3}|v|^4 + \frac{2rb_1c_E}{3}|v|^2 + \frac{rb_1M}{3} \le -\frac{rb_0}{3}|v|^4 + \frac{r}{3}\left(b_1M + \frac{b_1^2}{b_0}c_E^2\right).$$

The discrete version follows by replacing rb_0, rb_1 by p_0, p_1. It follows, in both cases, that

$$\sup_{t\ge 0}\mathbb{E}|V_t-u|^4 \le \max\{|v_0-u|^4, BM + B^2c_E^2\}, \tag{87}$$

with the supremum over step indices for a discrete tag. A corresponding second-moment bound follows by Cauchy–Schwarz. These bounds justify removing velocity cutoffs in the preceding weak moment identities by stopping and uniform integrability.

Finally, $\mathcal{K}$ is compact for W_2. Tightness follows from the bounded spatial domain and the fourth-moment bound, and

$$\sup_{f\in\mathcal{K}}\int_{|v-u|>L}|v-u|^2\,\mathrm{d}f \le M/L^2$$

gives uniform integrability of the squared velocity. The mean and second-moment constraints are closed under W_2 convergence; the fourth-moment sublevel is closed by lower semicontinuity.

I.3 Existence, uniqueness, and continuity

For a finite r, dominate the collision rate by rb_1. A backward collision tree adds an independent partner lineage at every candidate clock event, accepts with probability $\beta(x-y)/b_1$, and uses a fresh Haar mark. Each active lineage branches at the dominating rate; a Yule process bounds the number of leaves, which is almost surely finite before a finite horizon. Initialize leaves independently from f, use exact free flight along every edge, and evaluate the marked tree forward. This gives the population mild equation. For a separate tag, fix the starting state of its distinguished ancestral lineage at z and initialize all partner leaves from f.

Iterating the mild equation gives precisely these finite-tree terms. For any other solution, the unexplored remainder has mass at most the probability that the dominating tree exceeds the chosen cutoff, which tends to zero. Thus the population law is unique. Its deterministic background then

determines a unique nonexplosive tagged process, since its candidate rate is bounded. This is the bounded-kernel collision-tree construction of Graham & Méléard (1997); bounded rates make the tree finite even though the velocities are unbounded. There is also an a priori integrability bound: on a tree with n leaves, pair energy conservation bounds the squared centered root velocity by the sum of the squared centered leaf velocities, so its fourth power is at most n times their fourth-power sum. The Yule leaf count has a finite second moment on $[0, T]$, independently of the initial leaf states. Thus the initial fourth moment already gives a finite fourth moment on every finite horizon, for both a random root and a fixed tagged start. The same argument applies to the candidate Bernoulli trees below. It justifies the weak polynomial identities by truncation before applying the sharper bounds (85)–(87).

Lemma I.2 (Tagged continuity and time consistency)**.** *For every $R_*, T, L < \infty$, the map $(r, t, f, z) \mapsto K_t^{r,\beta}(f, z)$ is jointly W_2-continuous for $r \in [0, R_*]$, $t \in [0, T]$, $f \in \mathcal{K}$, and $z = (x, v)$ with $|v - u| \le L$. At each fixed reporting time t,*

$$\sup_{\substack{r \le R_*,\ f \in \mathcal{K} \\ x \in \mathbb{T}^3,\ |v-u| \le L}} W_2\big(K_{t,h}^{r,\beta}(f,(x,v)), K_t^{r,\beta}(f,(x,v))\big) \longrightarrow 0,$$

$$\sup_{r \le R_*,\ f \in \mathcal{K}} W_2\big(S_{t,h}^{r,\beta} f, S_t^{r,\beta} f\big) \longrightarrow 0 \qquad (h \downarrow 0).$$

Proof. If $R_* = 0$, the assertions follow from exact free flight. Otherwise use candidate clocks of rate $\Lambda = R_* b_1$ and acceptance probability $r\beta/\Lambda$. On $[0, T]$ the expected number of tree leaves is at most $e^{\Lambda T}$. Stop the exploration at J nodes and restrict its leaf velocities to a ball. The output on this finite tree is continuous in its input states, frequency, and edge lengths except at acceptance thresholds. Each gate uniform is independent of the states already constructed at its node, so equality with a threshold has probability zero. Candidate times have continuous distributions as well. Coupling the finitely many input leaves along a convergent sequence of laws and tagged starts therefore proves convergence in probability of the truncated outputs.

The probability of exceeding J nodes tends to zero uniformly. The initial fourth moments uniformly control the probability of a large leaf velocity. Equations (85) and (87) upgrade convergence in probability to convergence in squared mean under the coupling: if a discarded event has probability q, its squared endpoint-distance contribution is at most $Cq^{1/2}$ by Cauchy–Schwarz. This proves the asserted W_2 continuity. For a population output, the root start is sampled from f, and its fourth moment is bounded by M.

For time consistency, first compare with the linear gate $hr\beta$, for $h\Lambda \le 1$. It can be sampled by a candidate Bernoulli variable of parameter $h\Lambda$, followed by acceptance $r\beta/\Lambda$. A partner tree is explored only at a candidate; its expected number of leaves is at most $(1 + h\Lambda)^{\lceil T/h \rceil}$. For a fixed cutoff J, the Bernoulli candidate times converge to the Poisson tree times. The probability of two relevant Poisson candidates in one bin is $O_{\Lambda,T,J}(h)$. On the remaining finite tree, exact free flights and the continuous position-dependent rates converge as the candidate times approach their grid locations. This includes the stream-before-collision convention: the difference between the physical candidate time and its grid endpoint tends to zero on every retained edge. Thus the linear-gate endpoint converges in probability.

The saturated gate can use the same candidate Bernoulli variable, with conditional acceptance $(1 - e^{-hr\beta})/(h\Lambda)$. Since

$$0 \le hr\beta - (1 - e^{-hr\beta}) \le \tfrac{1}{2}(h\Lambda)^2,$$

the two conditional acceptance probabilities differ by at most $h\Lambda/2$. The uniformly bounded expected number of explored candidates makes their total disagreement probability $O_{\Lambda,T}(h)$. Both schemes satisfy the fourth-moment bounds, so truncation and the preceding Cauchy–Schwarz estimate give W_2 convergence.

Uniformity follows by contradiction. Any violating sequence has a subsequence with convergent frequencies, initial laws in the compact class $\mathcal{K}$, and bounded tagged starts. The same finite-tree coupling applies along that subsequence. For populations use a coupled random initial root, with the uniform fourth-moment bound. This proves both uniform limits. □

I.4 SYMMETRIC NEURAL INTENSITIES

Fix $R_*, T < \infty$. Let $\lambda_m(r, x, y)$ be nonnegative, symmetric pair-local neural intensities satisfying

$$\delta_m := \sup_{r \le R_*,\, x,y \in \mathbb{T}^3} |\lambda_m(r, x, y) - r\beta(x - y)| \longrightarrow 0.$$

They have a common finite upper bound Λ_*. Such approximants exist: embed the torus by periodic coordinates, approximate the continuous function on the resulting compact pair-input domain, clip to a common nonnegative interval containing the target, and average under x, y exchange. Proposition A.3 supplies the admitted neural operations. No context read or velocity truncation is needed for this position-only coefficient.

Retain exact free flight and the elastic Haar map. To control moments without a positive lower bound on λ_m, use pair energy conservation and the pointwise inequality

$$\big[|v' - u|^4 + |w' - u|^4 - |v - u|^4 - |w - u|^4\big]_+ \le 2|v - u|^2|w - u|^2.$$

Exchange symmetry conserves u, c_E and, after integrating the pair increment with its factor $1/2$, gives

$$\sup_{t \le T} M_4(f_t^m) \le M + \Lambda_* T c_E^2. \tag{88}$$

The same bound holds for the discrete populations at $jh \le T$, since $1 - e^{-h\lambda_m} \le h\Lambda_*$. For a separate tag, $|v' - u|^4 \le 2(|v - u|^4 + |w - u|^4)$ gives a fourth-moment drift bounded above by $C\Lambda_*(|v - u|^4 + M_4(f_t^m))$. Gronwall, or its discrete counterpart, therefore gives a uniform bound on $[0, T]$ for bounded tagged starts. The neural laws need not preserve the original fourth-moment threshold M; the bound just proved is what the convergence argument uses.

Couple the target and neural trees through clocks of a common rate $\Lambda_* \ge R_* b_1$. Up to the first differing acceptance, their states agree, and each candidate has disagreement probability at most δ_m/Λ_*. On any fixed tree cutoff the disagreement probability tends to zero, while the omitted-tree probability is uniformly small. For the discrete schemes use a candidate Bernoulli variable with parameter $h\Lambda_*$; the inequality

$$|(1 - e^{-ha}) - (1 - e^{-hb})| \le h|a - b| \qquad (a, b \ge 0)$$

gives the same conditional mismatch bound. Their expected explored tree sizes are bounded independently of small h and m. Equation (88) and the tagged fourth-moment bound then upgrade these couplings to W_2 convergence, uniformly over $r \le R_*$, $f \in \mathcal{K}$, and bounded tagged starts. Combining with Lemma I.2 proves the last clause of Theorem I.1, jointly as $h \downarrow 0$ and $\delta_m \downarrow 0$. The argument also proves convergence of the continuous neural laws as $\delta_m \downarrow 0$.

Thus Theorem I.1 is proved directly on the noncompact physical phase space by bounded-rate collision trees and fourth-moment control. The compact-convex-state universality theorem is not needed for this physical realization.

J KINETIC LIFTING AS A ROUTE TO FLUID MODELS

Kinetic evolution also supplies a representation of macroscopic fields. For density, momentum, and energy $U(x) = (\rho(x), m(x), \mathcal{E}(x))$ with positive density and internal energy, normalize total mass to one and choose a local Maxwellian $M_U(x, v)$ with those moments. Lifting and projection then take the form

$$U_0 \;\overset{\mathscr{E}}{\longmapsto}\; M_{U_0} \;\overset{S_t^{\mathrm{Kn}}}{\longmapsto}\; f_t^{\mathrm{Kn}} \;\overset{\mathscr{M}}{\longmapsto}\; U_t^{\mathrm{Kn}}, \qquad \mathscr{M} f = \int_{\mathbb{R}^d} (1, v, |v|^2/2) f(\cdot, v)\, \mathrm{d}v. \tag{89}$$

Here Kn is the Knudsen number controlling the collision scale. Kinetic moment scalings formally yield Euler and Navier–Stokes equations (Bardos et al., 1991); the Navier–Stokes limit is rigorous for suitable bounded collision kernels (Golse and Saint-Raymond, 2004). Relaxation and lattice kinetic models provide other fluid limits (Saint-Raymond, 2003; Junk and Yong, 2003).

This gives the enlarged representation freedom a macroscopic purpose. One can design a kinetic model around the fluid observables to be predicted. When the corresponding martingale problem is

well posed, exact equality of generators on its determining test class preserves the whole kinetic law. Different generators can also have the same macroscopic limit, but this requires a hydrodynamic-limit argument as well as compatible equilibria and transport coefficients. This motivates learning effective collision mechanisms subject to the required physical constraints. A fluid theorem for rate attention would additionally require locality, hydrodynamic scaling, and approximation estimates uniform as $\mathrm{Kn} \downarrow 0$. These are future objectives; Theorems A and B, together with Appendix I, establish the underlying representation and kinetic evolution results.