AS.110.503 : Undergraduate Research in Mathematics

Reconstructing Trajectories for Interacting Particle Systems from Unlabeled Data

James Guo (sguo45)
Advisor: Prof. Lu, Fei

May 12, 2026

Abstract

We study stochastic interacting particle systems with pairwise interaction potentials and external confinement when observations consist only of *unlabeled* empirical measures at discrete times—snapshots of particle positions without consistent identities across frames. The exposition connects optimal transport and Wasserstein geometry to the particle dynamics; develops learning via labeled simulation pipelines and a *trajectory-free self-test* objective built from discrete energy increments together with dissipation and diffusion functionals designed so that correct (Φ, V) shrink the residual when compared with observed energy changes. Learned drifts feed *prediction-informed* assignment costs between successive clouds. Trajectory reconstruction uses Hungarian matching and entropic regularization (Sinkhorn algorithm); comparisons include alternating match–estimate loops and *oracle* baselines where the simulator's true radial coefficients (or full oracle-kernel matching costs) are supplied.

Contents

Throughout the investigation of this topic this semester, I would really acknowledge the help and guidances offered by Prof. Lu, Fei, who have guided me into the theme of undergraduate research, constructing experimental models to verify theory, and the critical thinking views in analyzing how experimental setups can cause failures. I would greatly appreciate his support for the past year and instructions in leading me forward.

I Introduction

This thesis connects three mathematical threads: **(i)** Wasserstein gradient flows on the space of probability measures [JKO98; Vil09; PC19]; **(ii)** learning interaction potentials from *unlabeled* ensemble snapshots via a trajectory-free self-test loss in the spirit of recent work on interacting particle systems without trajectory labels [Lu+19; LL22]; and **(iii) trajectory reconstruction**—recovering latent particle labels across time—via optimal transport and assignment algorithms, including *prediction-informed* matching built from learned potentials. The unifying idea is that interacting particle systems can be read as finite-N realizations of gradient-flow structure in Wasserstein space; optimal transport both defines that geometry and supplies the tools to match unlabeled snapshots once a model of the drift is available.

Interactive systems are common in applications in physics, biology, and social dynamics. We seek mathematical models of such systems, discrete approximations for simulation, and—in the inverse direction—methods to infer underlying laws from data. This document focuses on stochastic interacting particles with pairwise and external potentials, and on what can be done when observations are unlabeled.

I.1 Motivation

At microscopic scales, neutral molecules are routinely modeled by empirical pairwise potentials fit to data or quantum calculations. Many chemical models specify such laws together with particle initial conditions.

Example I.1.1. Lennard–Jones interactions [Jon24].
Given particles at $x, y \in \mathbb{R}^n$ with separation $r = \|x - y\|$, the Lennard–Jones radial force magnitude is

$$F(r) := 24\epsilon \left[\left(\frac{\sigma}{r}\right)^6 - \left(\frac{\sigma}{r}\right)^{12} \right],$$

where $\epsilon > 0$ sets the well depth and σ sets the equilibrium separation. The force on x is $\mathbf{F}_x = F(r)\,(y - x)/\|y - x\|$ and on y is $\mathbf{F}_y = -\mathbf{F}_x$. With many particles, sum pairwise contributions; deterministic dynamics use $\ddot{x} \propto \mathbf{F}$, and stochastic particle models add diffusion as in Section III. ◊

Coarse-grained descriptions replace microscopic degrees of freedom by mesoscale agents governed by effective interactions and noise; the same mathematical abstraction applies whenever many identical units are tracked only through their empirical distribution.

When trajectories are fully labelled, tractable estimators exist for restricted hypothesis classes [WSL25]. This thesis emphasises the harder regime where labels between observation times are unknown.

Remark I.1.2. Singular Lennard–Jones Nature in Simulation.
The Lennard–Jones pair potential $\Phi_{\mathrm{LJ}}(r) = 24\epsilon\big((\sigma/r)^{12} - (\sigma/r)^6\big)$ diverges to ∞ as $r \searrow 0^+$: since the repulsive r^{-12} term dominates, so $\|\nabla\Phi_{\mathrm{LJ}}(r)\|$ grows without bound along approaching pairs. When particles are concentrated close to each other with in any time gap $dt > 0$, an explicit Euler–Maruyama could induce a drastic repulsion, and thus make the simulation unrealistic. ⌟

Figure I.1 is an example of the Lennard–Jones interactions ($d = 2$, $N = 30$, with $F_{\text{cap}} = 20$ and $dt = 0.002$). The trajectories are very irregular due to the drastic repulsion forces.

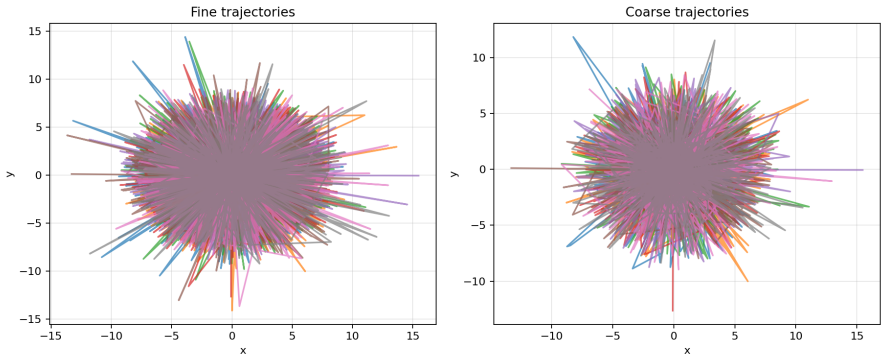

Figure I.1. Lennard–Jones merge-then-separate ensemble (`exp_20_lj_merge_then_separate_capped`, `model_lj`): `fig01` trajectory map showing crowded, rapidly turning paths under capped stiff pair forces.

I.2 The Inverse Problem

One assumption in the previous section is that we know the trajectory of each particle at a discrete set of data points, which does not necessarily have to be the case. When all the particles look the exact same, we rather only observe the "distribution" of the particles at a discrete time, but not the actual trajectories.

Example I.2.1. Unlabeled Particles Observed in Cell Biology [SST+19].
A concrete example appears in cell biology. In developmental or reprogramming experiments, one often measures gene expression profiles for many cells at a sequence of fixed sampling times; each cell is naturally represented as a point in a high-dimensional expression space, sampled over discrete time steps through an empirical distribution over that space. Because measurements are destructive or cells are not individually tracked between time points, the data give snapshots of the population but not the trajectories of the cells.
Such observations are inherently unlabeled in the same abstract sense as identical interacting particles: one could know how many units occupy each region of state space at each time, but not which unit corresponds to which across time. ◊

The thesis is organized around the following three questions:

(i) How do interacting particle systems (IPS) relate to gradient flows in Wasserstein space, and how does the empirical measure μ_t^N approximate a McKean–Vlasov limit?

(ii) How can one learn interaction potentials (Φ, V) from unlabeled ensemble snapshots, without knowing which particle at time t corresponds to which at time $t + \Delta t$?

(iii) How can one **reconstruct trajectories**—recover permutations (labels) across successive snapshots—using learned potentials together with optimal transport and assignment?

The **primary original contribution** emphasized here is **(3)**: trajectory reconstruction via *prediction-informed label matching*. Learning potentials in **(2)** (least squares on labeled simulation data and/or the trajectory-free self-test loss) is developed as enabling infrastructure: once $(\hat{\Phi}, \hat{V})$ approximate the drift, one-step Euler predictions yield cost matrices that often outperform naive Euclidean matching between consecutive clouds, especially when the observation interval Δt is large [Kuh55; Mun57; Cut13; SST+19].

Hereby, we formalize the *unlabeled snapshot observation problem* attached to question **(3)** above, since trajectory recovery is ill-posed without extra structure:

Write particle positions as $X_t = (X_t^1, \dots, X_t^N) \in (\mathbb{R}^d)^N$. Suppose that at each observation time we record only the empirical measure

$$\mu_t = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i},$$

up to relabeling of indices within each time slice—equivalently the multiset $\{X_t^1, \dots, X_t^N\}$ without a consistent correspondence across different t. Recover hidden labels / trajectories $\{X_t^i\}_{t,i}$, i.e. determine permutations linking successive snapshots whenever dynamics identify particles continuously.

I.3 Naïve Approaches

Video-like reasoning suggests each particle moves to its spatial nearest neighbour at the next frame—a *stable matching* under preferences ranked by inverse distance. Classical Gale–Shapley implementations build such a matching from pairwise comparisons using $\mathcal{O}(N^2)$ work per bipartite instance; chaining one matching between each consecutive pair of clouds over T frames yields $\mathcal{O}(TN^2)$ comparisons overall—yet stability does not imply physical fidelity when Δt is large or velocities are high:

Example I.3.1. Monotone tracks versus greedy spatial chaining.
Fix four distinguishable particles on parallel vertical lines. Suppose their true heights at integer times $t \in \{0, 1, 2\}$ are $y_k(t) = (k-1) + t$ for $k = 1, \dots, 4$ (constant horizontal offsets between particles). Thus each physical trajectory is a straight line of unit slope; unordered snapshots at successive times only reveal multisets $\{0, 1, 2, 3\}$, $\{1, 2, 3, 4\}$, and $\{2, 3, 4, 5\}$.

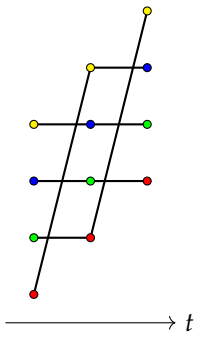

Figure I.2. Each filled dot marks a true particle location; heavier segments illustrate a greedy spatial chain between consecutive snapshots that joins spatially close points without enforcing trajectory continuity.

The diagram overlays these monotone tracks (filled dots) with a naive "nearest-neighbor" linkage between consecutive observation times (thicker segments): although each marginal snapshot looks like an ascending vertical stack, locally minimizing Euclidean displacement across frames can stitch together pairs that belong to different continuous trajectories whenever crossings or uneven speeds break simple vertical ordering—exactly the ambiguity absent from color labels. ◊

Minimizing total Euclidean displacement via the Hungarian algorithm ($\mathcal{O}(TN^3)$ per dense cost matrix) upgrades stable matching to optimal transport between clouds, yet still ignores dynamics encoded in (Φ, V) beyond instantaneous geometry.

On a higher level, given an algorithm that learns the interaction laws A with labeled data, we can develop the following naïve approaches (together with some constraints):

- **Brute-force enumeration.** Jointly choosing permutations $\pi_0, \ldots, \pi_{T-1}$ yields $(N!)^T$ label histories; evaluating a learner A on each history attains global optimum in principle but is completely infeasible for moderate N, T.
- **Matching-Learning Loop**. It is possible to iteratively make some naïve matching, learn the underlying laws based on this assumed matching, and make updates to the matching, which goes on and on. However, if the initial matching is off by a lot, this algorithm gets stuck with an incorrect matching and poorly learned laws under the system.

Therefore, beyond deriving a trajectory-free loss for learning, this thesis stresses **how learned dynamics improve trajectory reconstruction** when matching is posed as optimal transport between empirical measures.

I.4 Research Scope and Contributions

Recent work infers interaction laws from *labeled* trajectory data [Lu+19; LL22; Mil+23]. Snapshot-based observations without temporal identity arise in optimal-transport analyses of single-cell data and related trajectory-inference theory [SST+19; Lav+24; Vil09; PC19]. This thesis connects Wasserstein geometry [JKO98], learning from unlabeled ensembles, and assignment-based label recovery [Kuh55; Cut13].

Contributions are ordered to reflect the narrative above; the **capstone contribution** is trajectory matching:

- **Trajectory reconstruction (main contribution).** Implement and analyze *prediction-informed* label matching: build $\hat{X}^i_{t+\Delta t}$ from $(\hat{\Phi}, \hat{V})$ via explicit Euler, form costs $\|\hat{X}^i_{t+\Delta t} - X^{\text{obs},j}_{t+\Delta t}\|^2$, and solve assignment (Hungarian) or entropic OT (Sinkhorn). Compare to naive Euclidean matching between consecutive snapshots on simulated IPS with known ground-truth permutations.
- **Learning potentials (supporting).** Two-stage Legendre least squares on labeled fine-scale simulation [Sze75; Tre19; HK70; KP92]; trajectory-free self-test loss from energy, dissipation, and diffusion terms [JKO98].

- **Expository thread.** Self-contained background on optimal transport, W_2, Otto calculus, and Wasserstein gradient flows (heat flow, interaction energy, McKean–Vlasov free energy, JKO scheme), tied to the finite-N particle picture [Vil09; PC19; Szn91].

Synthetic experiments use known (Φ, V) so that matching accuracy and potential recovery can be quantified. Training may use labeled fine paths; evaluation stresses *coarse, randomly permuted* snapshots [SST+19].

I.5 Outline of the Writing

The remainder of the thesis is organized as follows:

- **Chapter II** develops optimal transport, the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, Otto calculus, Wasserstein gradient flows (including the JKO scheme and links to McKean–Vlasov free energy), and—on finite supports—the entropic regularization of Kantorovich transport that yields the Sinkhorn matrix-scaling iteration [Vil09; PC19; JKO98; Cut13; Sin64].
- **Chapter III** states the IPS model, the empirical measure μ_t^N, its evolution and mean-field limit, numerical tools (Brownian motion and SDEs, Euler–Maruyama, kernel caps, Legendre bases for radial kernels), and a subsection on *parametric* learning when the interaction kernel and external potential are expressed in the oracle mean-field coordinates [Øks03; KP92; Szn91; Sze75].
- **Chapter IV** formalizes unlabeled ensemble observations, the trajectory-free self-test loss and its energy-dissipation interpretation, brief least-squares estimation, and contrasts between the different approaches to solve this problem [Lu+19; LL22; HK70].
- **Chapter V** is the **main methodological chapter**: the label assignment problem, Kantorovich matching, Sinkhorn and Hungarian algorithms, **prediction-informed** costs from learned drifts, and numerical comparisons of reconstruction accuracy [Kuh55; Mun57; Cut13], as well as the experimental results.
- **Chapter VI** summarizes contributions, the logical circle linking Wasserstein geometry, learning, and matching, and future directions [SST+19; Lav+24; ZMM20].

II Optimal Transport and Wasserstein Space

This chapter follows the outline of a Wasserstein gradient-flow viewpoint [Vil09; PC19; JKO98]: optimal transport defines distances and geometry on probability measures; gradient flows in that geometry reproduce familiar PDEs and motivate the free-energy structure underlying interacting particle models.

II.1 Optimal transport, couplings, and Wasserstein distances

The **Monge problem** seeks T with $T_\# \mu = \nu$ (pushforward measure) minimizing $\int_{\mathbb{R}^d} c(x, T(x))\, d\mu(x)$ for lower semicontinuous $c \geq 0$ [Vil09]. Transport maps need not exist; the **Kantorovich relaxation** minimizes $\int_{\mathbb{R}^d \times \mathbb{R}^d} c\, d\pi$ over couplings $\pi \in \Pi(\mu, \nu)$ (joint laws with marginals μ, ν) [Vil09; PC19].

Definition II.1.1. Wasserstein-p distance.
Fix $p \in [1, \infty)$ and write $\mathcal{P}_p(\mathbb{R}^d)$ for Borel probabilities with finite p-th moment. The **Wasserstein-p distance** for the Euclidean metric is

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\pi(x, y) \right)^{1/p}, \qquad \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d). \tag{1}$$

For $p = 2$, W_2^2 is the minimal mean-square displacement over $\Pi(\mu, \nu)$. On $\mathcal{P}_2(\mathbb{R}^d)$, convergence in W_2 is equivalent to weak convergence of measures together with convergence of second moments—not to weak convergence alone [Vil09; PC19]. The **gluing lemma** concatenates couplings along an intermediate measure and yields the triangle inequality, so $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is a metric space [Vil09].

Example II.1.2. For intuition, in one dimension if $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$ and $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$,

$$W_2^2(\mu_0, \mu_1) = (m_0 - m_1)^2 + (\sigma_0 - \sigma_1)^2,$$

a closed form verified by coupling the natural coordinate-wise quantiles (equivalently via optimal monotone transport on $\mathbb{R}$) [PC19]. ◊

Thus W_2 captures both *location* (means) and *dispersion* (standard deviations) in one dimension.

For Gaussian measures on $\mathbb{R}^d$ with means m_0, m_1 and covariances Σ_0, Σ_1, the value W_2^2 admits the explicit Bures / Gelbrich formula with a matrix square-root term involving Σ_0, Σ_1 [PC19].

II.2 Riemannian Structure and Gradient Flow

Heuristically, $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ carries a formal Riemannian structure (Otto calculus): tangent directions correspond to velocity fields solving the *continuity equation* $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ in the sense of distributions [Vil09]. The Wasserstein gradient of a functional $\mathcal{F}[\rho]$ describes the direction of steepest decrease of $\mathcal{F}$ in W_2 [JKO98; PC19]. This Riemannian picture is a guiding calculus; optimal-map existence (Theorem II.2.1) and the JKO scheme stated later are standard rigorous facts cited from the literature.

Theorem II.2.1. Brenier's Theorem [Vil09].
Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and suppose μ is absolutely continuous with respect to Lebesgue measure. For the quadratic cost $c(x, y) = \frac{1}{2}\|x - y\|^2$, the optimal Kantorovich coupling between μ and ν is induced by a transport map $T = \nabla\varphi$ pushing μ forward to ν, where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex; T is uniquely determined μ-almost everywhere.

Remark II.2.2. Constant-speed W_2 Geodesics.
Under the hypotheses of Theorem II.2.1, if $\mu_1 = \nu$ and T is the optimal map from $\mu_0 = \mu$ to μ_1, then

$$\mu_t = \big((1 - t)\mathrm{Id} + tT\big)_{\#}\mu_0, \qquad t \in [0, 1],$$

is a constant-speed geodesic for W_2: trajectories interpolate linearly in Lagrangian coordinates along T [Vil09; PC19]. Discrete empirical matching across observation times is the finite counterpart of this

transport picture. ⌟

Definition II.2.3. Wasserstein Gradient Flow.
Given a functional $\mathcal{F}$ on probability densities on $\mathcal{P}_2(\mathbb{R}^d)$, a *Wasserstein gradient flow* is a narrowly continuous curve $(\rho_t)_{t\geq 0}$ satisfying the continuity equation $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ together with the gradient-flow velocity $v_t = -\nabla \frac{\delta \mathcal{F}}{\delta \rho_t}$ (under sufficient regularity). Equivalently we can write the nonlinear continuity–flux closure as

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho_t} \right). \tag{2}$$

⌟

Remark II.2.4. Heat flow, interaction energies, and McKean–Vlasov models.
Equation (2) specializes to classical macroscopic PDEs [JKO98; Vil09]. To align macroscopic diffusion with the particle SDE (6), fix a scalar noise level $\sigma \geq 0$ and use the entropic contribution

$$\mathcal{H}_\sigma[\rho] = \frac{\sigma^2}{2} \int_{\mathbb{R}^d} \rho \log \rho dx.$$

On $\mathbb{R}^d$ without boundary (or with no-flux data on bounded domains, which replaces Δ by the Neumann generator), the formal W_2 gradient flow of $\mathcal{H}_\sigma$ is $\partial_t \rho = \frac{\sigma^2}{2} \Delta \rho$. **Aggregation-type** equations stem from interaction energies

$$\mathcal{W}[\rho] = \frac{1}{2} \iint \Phi(x-y)\rho(x)\rho(y) dx dy.$$

McKean–Vlasov models combine $\mathcal{V}[\rho] = \int V \, d\rho$, pairwise interactions, and diffusion. ⌟

Two canonical specializations of the gradient flow equation (2) are worth recording explicitly, as they directly motivate the structure of the particle SDE (6) and the trajectory-free loss of Chapter IV.

- **Pure diffusion**. Taking $\mathcal{F} = \mathcal{H}_\sigma[\rho] = \frac{\sigma^2}{2} \int \rho \log \rho \, dx$, the functional derivative is $\frac{\delta \mathcal{H}_\sigma}{\delta \rho} = \frac{\sigma^2}{2}(\log \rho + 1)$, whose gradient is $\frac{\sigma^2}{2} \frac{\nabla \rho}{\rho}$. Substituting into (2) gives the heat equation $\partial_t \rho = \frac{\sigma^2}{2} \Delta \rho$. Given a smooth test function φ and applying integration by parts twice, the boundary terms vanish either by decay of ρ at infinity on $\mathbb{R}^d$.

- **Diffusion with external confinement**. Adding the external potential energy $\mathcal{V}[\rho] = \int V \, d\rho$ to $\mathcal{H}_\sigma$, the functional derivative gains the term ∇V, and the gradient flow becomes

  $$\partial_t \rho = \frac{\sigma^2}{2} \Delta \rho + \nabla \cdot (\rho \nabla V).$$

  This is the Fokker–Planck equation for a particle diffusing in the potential V: the first term spreads mass through diffusion while the second drives mass toward the minima of V.

These two cases show that the diffusion term $\sigma \, dW_t^i$ and the potential-driven drift $\nabla V(X_t^i)$ in the particle SDE (6) are both forced by the gradient flow structure of the free energy $\mathcal{F}$, and that the energy–dissipation balance they satisfy is precisely what the trajectory-free self-test loss in Chapter IV exploits to learn interaction potentials from unlabeled snapshots.

Remark II.2.5. Jordan–Kinderlehrer–Otto scheme.
To discretize (2) one employs the **JKO** variational scheme [JKO98; PC19]: given $\rho_k \in \mathcal{P}_2(\mathbb{R}^d)$ and $\tau > 0$,

choose

$$\rho_{k+1} \in \operatorname*{arg\,min}_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho_k) \right\}.$$

Each minimization plays the role of an implicit Euler step on $\mathcal{P}_2(\mathbb{R}^d)$, penalizing W_2 displacement from ρ_k. Jordan–Kinderlehrer–Otto prove convergence (for the motivating examples of [JKO98]) under structural hypotheses on $\mathcal{F}$—lower semicontinuity, convexity along generalized geodesics, sufficient growth/coercivity, and finite second moments for ρ_0 [JKO98; PC19]. ⌟

Remark II.2.6. Particles versus continuum limits.
Finite-N interacting diffusions embed into this picture: empirical measures μ_t^N approximate McKean–Vlasov limits consistent with (2) as $N \to \infty$ [Szn91]. The same energy–transport balance motivates the variational viewpoint adopted later for trajectory-free learning. ⌟

II.3 Entropic regularization and Sinkhorn scaling

The Kantorovich problem can be restricted to finite supports, which leads to the Sinkhorn scaling.

Consider $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^N b_j \delta_{y_j}$ as finite configuration with nonnegative weights $a_i, b_j \geq 0$ such that $\sum_{i=1}^N a_i = \sum_{j=1}^N b_j$, then $\pi \in \Pi(\mu, \nu)$ corresponds to $\gamma \in \mathbb{R}_{\geq 0}^{N \times N}$ with $\gamma \mathbf{1} = \mathbf{a}$, $\gamma^\intercal \mathbf{1} = \mathbf{b}$, and cost $\sum_{ij} C_{ij}\gamma_{ij}$ for $C_{ij} = c(x_i, y_j)$. Equal-weight empirical measures have $\mathbf{a} = \mathbf{b} = \frac{1}{N}\mathbf{1}$. We thus can write this in terms of a linear program:

$$\begin{aligned}
&\min_{\gamma \in \mathbb{R}^{N \times N}} \sum_{i,j=1}^N C_{ij}\gamma_{ij}, \\
&\text{s.t.} \sum_{j}^N \gamma_{ij} = a_i, \qquad \text{for } i \in [N], \\
&\qquad \sum_{i}^N \gamma_{ij} = b_j, \qquad \text{for } i \in [N], \\
&\qquad \gamma_{ij} \geq 0, \qquad \text{for } i, j \in [N].
\end{aligned}$$

Thus, in minimizing $\langle C, \gamma \rangle$ over $\Pi(\mu, \nu)$ is itself a linear program, and there exists algorithm simplex algorithm ($\mathcal{O}(2^{N^2})$) and interior point method ($\mathcal{O}(N^6)$), but are rather expensive, and so we should consider about entropic regularization [PC19].

Entropy penalty. Parallel to the macroscopic entropy functional $\mathcal{H}_\sigma[\rho] = \frac{\sigma^2}{2} \int \rho \log \rho \, dx$ whose formal W_2 gradient flow produces $\frac{\sigma^2}{2}\Delta\rho$ on $\mathbb{R}^d$ (Remark surrounding (2)), we can consider the Shannon entropy term of the *coupling*, where

$$H(\gamma) := -\sum_{i,j} \gamma_{ij} \log \gamma_{ij} \qquad (0 \log 0 := 0),$$

and weighted by a temperature $\varepsilon > 0$. The **entropic Kantorovich problem**—equivalently the discrete Schrödinger bridge at one time scale [Cut13; PC19]—turns the optimization into

$$\min_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \Big(C_{ij}\,\gamma_{ij} + \varepsilon\,\gamma_{ij}\log\gamma_{ij} \Big). \tag{3}$$

Since $\gamma_{ij} \mapsto \gamma_{ij}\log\gamma_{ij}$ is strictly convex on $\mathbb{R}_{>0}$, we have the uniqueness of the minimizer $\gamma^\star$ whenever $\Pi(\mu,\nu)$ has nonempty relative interior (excluding discrete dimensions, suffices to have positive marginals). As $\varepsilon \searrow 0^+$, $\gamma^\star$ concentrates on low-cost entries and selects an optimal vertex [Cut13; PC19].

Lagrangian derivation of the Gibbs kernel. Now, assume $a_i, b_j > 0$ for all i, j. Consider multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^N$ for the marginal constraints and write the Lagrangian (so we don't need to keep the constraints) as

$$\mathcal{L}(\gamma; \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{i,j=1}^{N} \big(C_{ij}\gamma_{ij} + \varepsilon\gamma_{ij}\log\gamma_{ij} \big) + \sum_{i=1}^{N} \lambda_i \big(a_i - (\gamma\mathbf{1})_i \big) + \sum_{j=1}^{N} \mu_j \big(b_j - (\gamma^\intercal\mathbf{1})_j \big).$$

At any interior critical point, $\partial\mathcal{L}/\partial\gamma_{ij} = C_{ij} + \varepsilon(\log\gamma_{ij} + 1) - \lambda_i - \mu_j = 0$, hence

$$\gamma_{ij} = \exp\left(\frac{\lambda_i + \mu_j - C_{ij}}{\varepsilon} - 1 \right) = \exp\left(\frac{\lambda_i + \mu_i}{\varepsilon} - 1 \right) \cdot \underbrace{\exp\left(\frac{-C_{ij}}{\varepsilon} \right)}_{K_{ij}}.$$

We define the Gibbs kernel $K_{ij} := \exp(-C_{ij}/\varepsilon)$ and absorb the displayed exponential prefactor into positive scalings of the multipliers, setting

$$u_i := \exp\left(\frac{\lambda_i}{\varepsilon} - \frac{1}{2} \right), \qquad v_j := \exp\left(\frac{\mu_j}{\varepsilon} - \frac{1}{2} \right),$$

so $u_i v_j = \exp\big((\lambda_i + \mu_j)/\varepsilon - 1\big)$ and hence $\gamma_{ij} = u_i\,K_{ij}\,v_j$. Folding all of it into u_i or v_j would be equally valid and is absorbed into the Sinkhorn gauge $\mathbf{u} \mapsto t\mathbf{u}$, $\mathbf{v} \mapsto t^{-1}\mathbf{v}$. Hence, we have the product structure for Sinkhorn scaling [Cut13; PC19] as

$$\gamma_{ij} = u_i\,K_{ij}\,v_j. \tag{4}$$

The minimizer $\gamma^\star \in \Pi(\mu,\nu)$ is uniquely determined by (4) with the marginal constraints [Cut13; PC19].

Sinkhorn–Knopp iteration. We can write (4) as $\gamma = D_{\mathbf{u}} K D_{\mathbf{v}}$, the constraints $\gamma\mathbf{1} = \mathbf{a}$ and $\gamma^\intercal\mathbf{1} = \mathbf{b}$ as $\mathbf{u} \odot (K\mathbf{v}) = \mathbf{a}$ and $\mathbf{v} \odot (K^\intercal\mathbf{u}) = \mathbf{b}$ (as Hadamard product). Equivalently,

$$u_i = \frac{a_i}{(K\mathbf{v})_i}, \qquad v_j = \frac{b_j}{(K^\intercal\mathbf{u})_j}. \tag{5}$$

Starting from arbitrary $\mathbf{u}^{(0)} \in \mathbb{R}^N_{>0}$ and $\mathbf{v}^{(0)} \in \mathbb{R}^N_{>0}$, we can update via

$$u_i^{(m+1)} = \frac{a_i}{(K\mathbf{v}^{(m)})_i} \qquad \text{and} \qquad v_j^{(m+1)} = \frac{b_j}{(K^\intercal\mathbf{u}^{(m+1)})_j},$$

which constitute the *Sinkhorn–Knopp* (iterative proportional fitting) algorithm [Sin64; PC19; Cut13]. Each half-step projects onto one marginal while holding the other scaling fixed. For strictly positive K, a future statement (Theorem V.3.1) would guarantee the existence and uniqueness of $(\mathbf{u}^\star, \mathbf{v}^\star)$ solving (5) up to $\mathbf{u} \mapsto t\mathbf{u}$, $\mathbf{v} \mapsto t^{-1}\mathbf{v}$, and the iterates converge to that solution [Sin64; PC19], with iteration costs $\mathcal{O}(N^2)$.

III Interacting Particle Systems and the Empirical Distribution

III.1 Brownian motion, Itô integrals, and the Fokker–Planck equation

Stochastic differential equations (SDEs) provide the continuous-time language for noisy interacting systems; Brownian motion supplies the driving noise. Definition III.1.1 below from [Øks03] defines the Brownian motion by specifying finite-dimensional distributions first; Proposition III.1.2 records continuity as a consequence together with other standard properties about the Brownian motion.

Definition III.1.1. Brownian Motion.
A stochastic process $\{B_t\}_{t\geq 0}$ (where $B_t \in \mathbb{R}^n$) is a **Brownian motion** if given any finite sequence $0 \leq t_1 \leq t_2 \leq \cdots \leq t_k$ with measure $\nu_{t_1,\cdots,t_k}$ on $\mathbb{R}^{nk}$:

$$\nu_{t_1,\cdots,t_k}(F_1,\cdots,F_k) = \int_{F_1\times\cdots\times F_k} p(t_1,x,x_1)p(t_2-t_1,x_1,x_2)\cdots p(t_k-t_{k-1},x_{k-1},x_k)dx_1\cdots dx_k,$$

where F_i's are Borel sets in $\mathbb{R}^n$ and p is defined for all $x \in \mathbb{R}^n$ as:

$$p(t,x,y) = (2\pi t)^{-n/2}\cdot\exp\left(-\frac{|x-y|^2}{2t}\right) \qquad \text{for } y\in\mathbb{R}^n \text{ and } t>0,$$

and the extension $t \to 0^+$ recovers the Dirac mass δ_x as an initial condition for the transition kernel (interpreted in the weak sense; there is no integrable pointwise density at t=0).
The Brownian motion satisfies that:

$$\mathbb{P}^x(B_{t_1}\in F_1,\cdots,B_{t_k}\in F_k) = \nu_{t_1,\cdots,t_k}(F_1,\cdots,F_k).$$

Here $\mathbb{P}^x$ denotes the law of the process started from $B_0 = x$ almost surely (standard Brownian motion corresponds to $x = 0$). ⌟

Finite-dimensional distributions defined above are consistent under marginalisation (Gaussian transitions). Kolmogorov's extension theorem yields a probability space $(\Omega,\mathcal{F},\mathbb{P}^x)$ and a process $\{B_t\}_{t\geq 0}$ with those finite-dimensional laws and initial point $B_0 = x$ a.s.

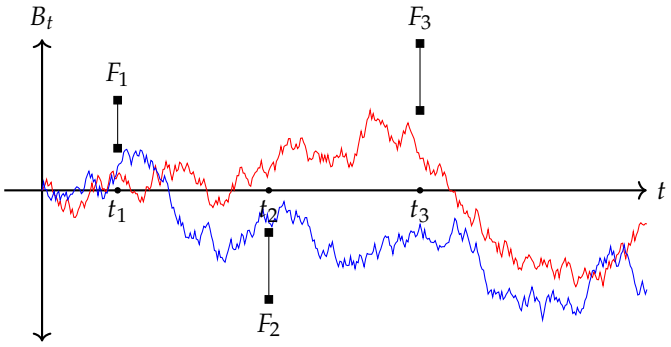

Figure III.1. Illustration of Brownian Motion in 1D with $k = 3$.

Proposition III.1.2. Basic properties of Brownian motion.
Standard consequences include:

(i) B_t is a Gaussian process, *i.e.*, for all $0 \le t_1 \le \cdots \le t_k$, the vector $Z = (B_{t_1}, \cdots, B_{t_k}) \in \mathbb{R}^{nk}$ is jointly Gaussian (equivalently, its entries admit a multivariate normal distribution).

(ii) B_t has independent increments: for $0 < t_1 < \cdots < t_k$, the vectors $B_{t_1} - B_0$, $B_{t_2} - B_{t_1}$, …, $B_{t_k} - B_{t_{k-1}}$ are mutually independent (under $\mathbb{P}^x$ one has $B_0 = x$ a.s.).

(iii) $t \mapsto B_t(\omega)$ is continuous for almost all $\omega \in \Omega$.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge0}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions, and let B_t be an $(\mathcal{F}_t)$-Brownian motion in $\mathbb{R}^n$ as in Definition above. For an $(\mathcal{F}_t)$-predictable matrix-valued process H_t satisfying a square-integrability condition (e.g. $\mathbb{E}\int_0^T \|H_t\|^2 dt < \infty$), the Itô integral $\int_0^T H_t\, dB_t$ is defined as the L^2-limit of simple integrands and satisfies the Itô isometry in the scalar case [Øks03]. This construction extends componentwise when H_t is vector- or matrix-valued, which is the setting for multi-dimensional SDEs driven by independent Brownian motions.

A key foundational result ensures that Itô SDEs of the form studied here are well-posed under standard regularity assumptions. We state this as a theorem, as follows:

Theorem III.1.3. Existence and Uniqueness of Strong Solutions [Øks03, Thm. 5.2.1].
Consider the time-homogeneous Itô SDE in $\mathbb{R}^m$,

$$dY_t = b(Y_t)dt + \sigma(Y_t)\, dW_t,$$

where W_t is a d-dimensional Brownian motion. Suppose the drift $b : \mathbb{R}^m \to \mathbb{R}^m$ and diffusion $\sigma : \mathbb{R}^m \to \mathbb{R}^{m\times d}$ satisfy the following:

- (**Lipschitz Continuity**) There exists $L > 0$ such that for all $x, y \in \mathbb{R}^m$,

$$\|b(x) - b(y)\| + \|\sigma(x) - \sigma(y)\| \le L\|x - y\|.$$

- (**Linear Growth**) There exists $K > 0$ such that for all $x \in \mathbb{R}^m$,

$$\|b(x)\| + \|\sigma(x)\| \le K(1 + \|x\|).$$

Then, for any initial condition $Y_0 \in \mathbb{R}^m$, the SDE admits a *unique strong solution* $(Y_t)_{t\ge0}$ up to any finite time horizon.

These regularity conditions constitute the standard hypothesis under which we simulate interacting particle systems throughout this thesis. They guarantee that the simulated stochastic processes are well-defined and numerically tractable. More general existence and uniqueness resultsallowing for local Lipschitz assumptions or Lyapunov-type conditionsare available but will not be used here.

Theorem III.1.4. SDE and Fokker–Planck Equation.
Let Y_t be a solution to the SDE

$$dY_t = b(Y_t)dt + \sigma(Y_t)\, dW_t,$$

where $b : \mathbb{R}^m \to \mathbb{R}^m$ is the drift, $\sigma : \mathbb{R}^m \to \mathbb{R}^{m\times d}$ is the diffusion coefficient, and W_t is a d-dimensional Brownian motion. Suppose Y_t admits a smooth density $p(t, y)$. Then p formally satisfies the forward Kolmogorov (Fokker–Planck) equation:

$$\frac{\partial p}{\partial t}(t, y) = -\nabla_y \cdot \big(b(y)\, p(t, y)\big) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial y_i \partial y_j}\big[(\sigma\sigma^\intercal)_{ij}(y)\, p(t, y)\big],$$

with initial condition $p(0, \cdot)$ given by the law of Y_0 [Ris96; Øks03].

Here, an Itô SDE governs not only the evolution of individual sample trajectories but also the time evolution of their probability law. The Fokker–Planck equation describes how the probability density p for the process Y_t changes over time, and is especially relevant in contexts where observations are unlabeled and only empirical distributions are available.

Definition III.1.5. N**-particle system.**
An N**-particle interacting system** in $\mathbb{R}^d$ consists of N indistinguishable particles with positions $X_t^i \in \mathbb{R}^d$, $i = 1, \dots, N$, evolving according to the coupled stochastic differential equations

$$dX_t^i = -\left(\frac{1}{N}\sum_{j\neq i} \nabla\Phi(X_t^i - X_t^j) + \nabla V(X_t^i)\right) dt + \sigma\, dW_t^i, \qquad i = 1, \dots, N, \tag{6}$$

where

- $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ is the *pairwise interaction potential*, governing interactions between particles,
- $V\colon \mathbb{R}^d \to \mathbb{R}$ is an *external* (confining) potential,
- $\sigma \geq 0$ is the noise amplitude,
- $(W_t^i)_{t\geq 0}$ are independent Brownian motions in $\mathbb{R}^d$,
- the prefactor $1/N$ is the standard mean-field scaling under which bounded, sufficiently regular Φ produce $\mathcal{O}(1)$ pairwise drift per particle as $N \to \infty$ [Szn91]; for singular kernels (Remark I.1.2) this heuristic "$\mathcal{O}(1)$ force" picture breaks unless one caps or softens the singularity, yet simulations still employ (6) with numerical stabilizers. ⌟

Notes on conventions. Equation (6) writes the drift as *minus* the gradient stack $\frac{1}{N}\sum_{j\neq i} \nabla\Phi(X_t^i - X_t^j) + \nabla V(X_t^i)$; some texts instead define bare forces $-\nabla\Phi$ and fold signs differently. Here conservative attraction/repulsion along the separation vector is encoded entirely in Φ' (radially), while the overall drift carries the leading minus sign consistent with overdamped descent references used elsewhere.

Also, we assume throughout that Φ is *symmetric* in the sense that $\Phi(-x) = \Phi(x)$ for all $x \in \mathbb{R}^d$, so that the interaction force on particle i arising from j equals $-\frac{1}{N}\nabla\Phi(X_t^i - X_t^j)$ and Newton's third law holds at the level of conservative forces. In applications one often takes Φ *radial*, i.e. $\Phi(x) = \phi(\|x\|)$ for a scalar function ϕ on $[0,\infty)$; then $\nabla\Phi(x) = \phi'(\|x\|)\, x/\|x\|$ for $x \neq 0$. The external potential V need not be radial; a standard choice in experiments is a quadratic well $V(x) = \delta\|x\|^2$.

What distinguishes the normal learning problem is that we do not have the trajectory for the particles. We only have the empirical measure of the particle configuration at time t.

Definition III.1.6. Empirical Measure Configuration.
At time t, the **empirical measure configuration** of the particle system is given by

$$\mu_t^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i},$$

where $\delta_{X_t^i}$ denotes the Dirac measure (canonical delta) at the position X_t^i of particle i. ⌟

Remark III.1.7. Stacking $Y_t = (X_t^1, \ldots, X_t^N)$ reduces the N-particle system to a single SDE of the form $dY_t = b(Y_t)dt + \Sigma d\mathcal{W}_t$, to which the Fokker-Planck theorem applies when Φ, V have bounded Hessians; for singular kernels such as LennardJones, the drift is not globally Lipschitz and the equation holds only formally (see Remark I.1.2). ⌟

Under appropriate regularity on (Φ, V), as $N \nearrow \infty$ the sequence $(\mu_t^N)_{t\geq 0}$ converges (in law) to a deterministic measure-valued flow solving a McKean–Vlasov or continuity equation; as the framework of *propagation of chaos* [Szn91; McK66]. Inside this thesis, we use finite-N systems for simulation and learning; the mean-field picture motivates the $1/N$ scaling and the interpretation of data as samples from an evolving distribution.

III.2 Euler–Maruyama Discretization

Then, we will explore into the procedure of "discretizing" the continuous relations of the particle system into a "discrete" time-step simulation, known as the **Euler's method** in ODEs. However, what turns different would be the "noise term" and we would account for that in the simulation of the system.

Definition III.2.1. Euler–Maruyama Discretization.
For numerical simulation, we discretize (6) using a step size $\Delta t > 0$. The **Euler–Maruyama scheme** updates each particle as

$$X_{t+\Delta t}^i = X_t^i - \left(\frac{1}{N}\sum_{j\neq i}\nabla\Phi(X_t^i - X_t^j) + \nabla V(X_t^i)\right)\Delta t + \sigma\sqrt{\Delta t}\,\xi_t^i,$$

where $\xi_t^i \sim \mathcal{N}(0, I_d)$ are independent standard normal vectors across i and t [KP92; Mar55]. In code, Δt may be chosen as the fine integration step dt or as a coarser step used for subsampled observations. ⌟

Theorem III.2.2. Convergence of the Euler–Maruyama Scheme.
Under standard Lipschitz assumptions on the drift and diffusion coefficients, the Euler–Maruyama scheme's **strong** (pathwise) error at a fixed time horizon has root-mean-square order $\mathcal{O}(\Delta t^{1/2})$ in the step size Δt, while **weak** errors for smooth test functionals are $\mathcal{O}(\Delta t)$ [KP92, Theorem 14.1.1].

In this thesis, moderate (or no) noise and small (if not no noise) dt are chosen so that discretization bias is secondary compared to statistical error in learning.

Remark III.2.3. Stabilization via Force and Kernel Caps.
Pairwise potentials such as Lennard–Jones at the origin (see Remark I.1.2) are singular. In simulation, it is common to clip the *values* of Φ and V to finite intervals—for reporting or for learned models—while retaining smooth analytic gradients for the reference dynamics when specified [FS02]. This thesis adopts the "cap scalar kernel" convention, where

$$\Phi_{\text{eff}} = \text{clip}(\Phi) \qquad \text{and} \qquad V_{\text{eff}} = \text{clip}(V)$$

are used for displayed energies (see Section IV for experimental implementation). ⌟

III.3 Parametric Learning in the Oracle Basis

In simulations, Φ and V are specified in a closed analytic family: finite radial harmonics for ϕ and quadratic wells $V(x) = \delta\|x\|^2$. Learning reduces to estimating low-dimensional coefficient vectors rather than searching over unstructured hypothesis spaces.

Definition III.3.1. Oracle-Linear Parametrization.
Fix radial features $\psi : [0, r_{\max}] \to \mathbb{R}^{K_\Phi}$ and external template $q : \mathbb{R}^d \to \mathbb{R}$ (typically $q(x) = \|x\|^2$). An *oracle-linear* model takes the form

$$\Phi(x) = \phi(\|x\|), \qquad \phi(r) = \beta^\intercal \psi(r), \qquad V(x) = \theta\, q(x),$$

with unknown parameters $\beta \in \mathbb{R}^{K_\Phi}$ and $\theta \in \mathbb{R}$. Gradients entering (6) are $\nabla\Phi(x) = \phi'(\|x\|)\, x/\|x\|$ and $\nabla V(x) = \theta\, \nabla q(x)$. ⌟

Definition III.3.2. Velocity-Matching Loss.
Given fine-scale snapshots $\{X_{t_\ell}\}_{\ell=0}^{L}$ with consistent particle indices and spacing Δt, define empirical velocities $Y_\ell^i := -(X_{t_{\ell+1}}^i - X_{t_\ell}^i)/\Delta t$ and predicted drifts

$$F_\ell^i(\beta, \theta) := \frac{1}{N} \sum_{j \neq i} \nabla\Phi_\beta\big(X_{t_\ell}^i - X_{t_\ell}^j\big) + \nabla V_\theta\big(X_{t_\ell}^i\big).$$

The quadratic velocity loss is

$$\mathcal{J}_{\text{oracle LS}}(\beta,\theta) := \frac{1}{LN}\sum_{\ell=0}^{L-1}\sum_{i=1}^{N}\left\|Y_\ell^i - F_\ell^i(\beta,\theta)\right\|^2. \tag{7}$$

Since $(\beta,\theta)\mapsto F_\ell^i(\beta,\theta)$ is affine for each (ℓ,i), minimizing (7) reduces to a finite-dimensional ridge system

$$(A^\intercal A + \lambda\,\mathrm{Id})\,\hat{\xi} = A^\intercal b, \tag{8}$$

where rows of A are the feature vectors ξ_ℓ^i defined later in Section IV, $\lambda\,\mathrm{Id}$ being the ridge regularization (Tikhonov penalty) to prevent unnecessary complexity due to noise, and $\hat{\xi} = (\hat{\beta},\hat{\theta})$ is the unique minimizer whenever $A^\intercal A + \lambda\,\mathrm{Id}$ is positive definite [HK70].

Then, we are left to derive this A as feature vectors for the specific cases for the least square estimate methods in the next section.

IV Learning Interaction Potentials

In the simulations, we have access to the full state $X_t = (X_t^1,\dots,X_t^N)$ integrated with a small step dt. Indices i are fixed along time, so $\{X_{kdt}^i\}_k$ is a labeled trajectory for each particle. This regime supports least-squares fitting of drifts from finite-difference velocities, as in [Lu+19]; richer nonparametric two-stage pipelines from [WL26].

For evaluation we subsample every m steps with $\Delta t = mdt$ and, at each coarse time, apply an independent uniformly random permutation to particle indices. The observer sees only multisets (or unordered lists) $\{x_{t_\ell,1},\dots,x_{t_\ell,N}\}$; equivalently the empirical measure $\mu_{t_\ell} = \frac{1}{N}\sum_i \delta_{x_{t_\ell,i}}$ without a consistent labeling across ℓ.

IV.1 Labeled Trajectory Loss

For particle systems where trajectories are labeled, we can leverage consecutive position data to infer underlying interaction and external potentials quite directly. Observing positions X_t and $X_{t+\Delta t}$ at successive times allows us to estimate empirical velocities $y_t^i := -(X_{t+\Delta t}^i - X_t^i)/\Delta t$, which approximate the negative drift suggested by discretized Itô dynamics. The learning objective is to determine parameterized potentials $(\hat{\Phi},\hat{V})$ with the following loss function.

Definition IV.1.1. Labeled Trajectory Loss.
Given observed positions X_t and $X_{t+\Delta t}$, define empirical velocities $y_t^i := -(X_{t+\Delta t}^i - X_t^i)/\Delta t$ for each particle. For parameterized potentials $(\hat{\Phi},\hat{V})$, the labeled trajectory loss is

$$\mathcal{L}_{\text{traj}} = \frac{1}{N}\sum_{i=1}^{N}\left\|y_t^i - \left(\frac{1}{N}\sum_{j\neq i}\nabla\hat{\Phi}(X_t^i - X_t^j) + \nabla\hat{V}(X_t^i)\right)\right\|^2.$$

With linear parameterization of the potentials, this becomes a quadratic least-squares problem (possibly with ridge regularization [HK70]) and can be efficiently minimized.

Given the oracle basis setting (7), the closed form solution has the feature vector for i-th entry as

$$\xi^i := \begin{pmatrix} \frac{1}{N}\sum_{j\neq i} \psi'(\|X^i - X^j\|)\frac{X^i - X^j}{\|X^i - X^j\|} \\ \nabla q(X^i) \end{pmatrix} \in \mathbb{R}^{K_\Phi + 1}.,$$

so that $F^i(\xi) = (\xi^i)^\intercal \xi$. Stacking all N particles, the loss becomes $\mathcal{L}_{\text{traj}} = \frac{1}{N}\|A\xi - b\|^2$ where rows of A are $(\xi^i)^\intercal$ and $b_i = y^i$. The closed form solutions are then

$$\hat{\xi} = \begin{pmatrix} \hat{\beta} \\ \hat{\theta} \end{pmatrix} = (A^\intercal A + \lambda I)^{-1} A^\intercal b, \tag{9}$$

where $\lambda \geq 0$ is the ridge penalty ($\lambda = 0$ gives ordinary least squares).

IV.2 Trajectory-Free Self-Test Loss

The above derivation is for the labeled trajectories, which is not entirely the setup for this thesis. The implementation would to utilize the trajectory-free self-test loss in [WL26].

Definition IV.2.1. Trajectory-Free Self-Test Quantities.
Given potentials (Φ, V) and a configuration $X_t = (X_t^1, \ldots, X_t^N)$, define:

- The total energy:
$$E_{\text{tot}}(X_t) = \frac{1}{N}\sum_{i=1}^N V(X_t^i) + \frac{1}{2N^2}\sum_{i,j=1}^N \Phi(X_t^i - X_t^j).$$
- The dissipation functional:
$$J_{\text{diss}}(X_t) = \frac{1}{N}\sum_{i=1}^N \left\| \nabla V(X_t^i) + \frac{1}{N}\sum_{j=1}^N \nabla\Phi(X_t^i - X_t^j) \right\|^2.$$
- The diffusion correction:
$$J_{\text{diff}}(X_t) = \frac{1}{N}\sum_{i=1}^N \Delta V(X_t^i) + \frac{1}{N^2}\sum_{i,j=1}^N \Delta\Phi(X_t^i - X_t^j).$$

⌟

Definition IV.2.2. Trajectory-free self-test loss.
Let $\{X_{t_\ell}\}_{\ell=0}^L$ be a discrete-time trajectory with spacing Δt. The trajectory-free self-test loss associated to (Φ, V) is

$$\mathcal{L}_{\text{TF}}(\Phi, V) = \frac{1}{L}\sum_{\ell=0}^{L-1}\left[\frac{1}{2}J_{\text{diss}}(X_{t_\ell})\,\Delta t - \frac{\sigma^2}{2}J_{\text{diff}}(X_{t_\ell})\,\Delta t + \left(E_{\text{tot}}(X_{t_{\ell+1}}) - E_{\text{tot}}(X_{t_\ell})\right)\right]. \tag{10}$$

⌟

Sketch of Derivation. We derive $\mathcal{L}_{\mathrm{TF}}$ from the weak-form stochastic PDE satisfied by the empirical distribution, following the self-test framework of [WL26].

Step 1: Weak-form PDE. For any test function $f \in C_b^2(\mathbb{R}^d)$, Itô's formula applied to $\frac{1}{N}\sum_{i=1}^N f(X_t^i)$ gives

$$d\left\langle \mu_t^N, f\right\rangle = \left\langle -\nabla\cdot\left[\mu_t^N \nabla\left(\Phi * \mu_t^N + V\right)\right] + \frac{\sigma^2}{2}\Delta\mu_t^N,\, f\right\rangle dt + \sigma\, dm_{X_t}(f),$$

where $m_{X_t}(f) := \frac{1}{N}\sum_i \int_0^t \nabla f(X_s^i)\cdot dW_s^i$ is a martingale with mean zero and variance $\mathrm{Var}(m_{X_t}(f)) = \mathcal{O}(1/N)$ by Itô's isometry. Equivalently, μ_t^N satisfies the weak-form stochastic evolution equation

$$\partial_t \mu_t^N = \nabla\cdot\left[\mu_t^N \nabla\left(\Phi * \mu_t^N + V\right)\right] + \frac{\sigma^2}{2}\Delta\mu_t^N + \sigma \dot{m}_{X_t}.$$

Step 2: Abstract operator form. Define the drift operator

$$R_\varphi[u] := -\nabla\cdot\left[u(\nabla V + \nabla\Phi * u)\right],$$

which maps a probability measure u to the negative divergence of its drift flux, and is *linear* in $\varphi = (V, \Phi)$ despite being nonlinear in u. The weak-form PDE then becomes the linear equation in φ:

$$R_{\varphi^*}[\mu_t^N] - \sigma\dot{m}_{X_t} = g_t^N, \qquad g_t^N := -\partial_t\mu_t^N + \frac{\sigma^2}{2}\Delta\mu_t^N,$$

where g_t^N depends only on μ_t^N and σ, and $\dot{m}_{X_t}$ is unobservable zero-mean noise.

Step 3: Self-test family and quadratic loss. To construct a quadratic loss whose minimizer satisfies the weak-form PDE, we test the equation against the *self-testing family*

$$f_\varphi[u] := V + \Phi * u,$$

which is linear in φ and depends on the data through u. Define the bilinear form, using integration by parts to pass from the divergence form of $R_\varphi[u]$ to the gradient inner product,

$$B_u(\varphi,\psi) := \langle R_\varphi[u],\, f_\psi[u]\rangle \overset{\text{IBP}}{=} \int_{\mathbb{R}^d} u\,\nabla(V + \Phi * u)\cdot\nabla(\tilde{V} + \tilde{\Phi} * u)\, dx,$$

for $\varphi = (V, \Phi)$ and $\psi = (\tilde{V}, \tilde{\Phi})$. This choice satisfies three essential properties:

- **Linearity:** $f_{\varphi+\psi}[u] = f_\varphi[u] + f_\psi[u]$, so $B_u(\varphi,\psi)$ is bilinear in (φ,ψ).
- **Symmetry:** The gradient inner product representation gives $B_u(\varphi,\psi) = B_u(\psi,\varphi)$ directly.
- **Non-negativity:** $B_u(\varphi,\varphi) = \int_{\mathbb{R}^d} u\,|\nabla(V + \Phi * u)|^2\, dx \geq 0$.

These three properties together ensure that the quadratic loss

$$E(\varphi) = \frac{1}{2}B_u(\varphi,\varphi) - \langle g_t^N,\, f_\varphi[u]\rangle$$

has its Fréchet derivative at the true φ^* equal to zero:

$$\lim_{\varepsilon\to 0}\frac{E(\varphi^* + \varepsilon\varphi) - E(\varphi^*)}{\varepsilon} = \langle R_{\varphi^*}[u] - g_t^N,\, f_\varphi[u]\rangle = 0 \quad \text{for all } \varphi,$$

which is precisely the weak-form PDE tested against $\{f_\varphi[u]\}_\varphi$. The $\frac{1}{2}$ in front of $B_u(\varphi,\varphi)$ is the direct analogue of the $\frac{1}{2}x^\intercal Ax$ term in the finite-dimensional quadratic $\frac{1}{2}x^\top Ax - b^\top x$; it is essential for the Fréchet derivative to recover the weak-form PDE exactly, and distinguishes this self-test loss from the energy-dissipation balance approach (which leads to a quartic loss).

Step 4: Assembling the discrete loss. Applying the construction to the empirical distributions $\{\mu_{t_\ell}^{N,m}\}$ with a left-endpoint Riemann-sum approximation for the time integral $\int_{t_\ell}^{t_{\ell+1}} \frac{1}{2}B_{\mu_t^N}(\varphi,\varphi)\,dt$, and averaging over the dataset, we obtain the self-test loss $\mathcal{L}_{\mathrm{TF}}$ in (10). To compute the $\langle g_t^N, f_\varphi[\mu_t^N]\rangle$ term explicitly, note first that by the symmetry of Φ:

$$\langle -\partial_t\mu_t^N,\, V+\Phi*\mu_t^N\rangle = -\frac{d}{dt}\int_{\mathbb{R}^d}\left[V+\frac{1}{2}\Phi*\mu_t^N\right]\mu_t^N\,dx = -\frac{d}{dt}E_{\mathrm{tot}}(X_t),$$

and by integration by parts:

$$\left\langle \frac{\sigma^2}{2}\Delta\mu_t^N,\, V+\Phi*\mu_t^N\right\rangle = \frac{\sigma^2}{2}\int_{\mathbb{R}^d}\mu_t^N\left[\Delta V+\Delta\Phi*\mu_t^N\right]dx = \frac{\sigma^2}{2}J_{\mathrm{diff}}(X_{t_\ell}).$$

For the dissipation term, note that by definition of $\mu_{t_\ell}^N = \frac{1}{N}\sum_i \delta_{X_{t_\ell}^i}$, the non-negativity of $B_u(\varphi,\varphi)$ discretizes exactly as

$$\frac{1}{2}B_{\mu_{t_\ell}^N}(\varphi,\varphi) = \frac{1}{2}\int_{\mathbb{R}^d}\mu_{t_\ell}^N\left|\nabla V+\nabla\Phi*\mu_{t_\ell}^N\right|^2dx = \frac{1}{2}J_{\mathrm{diss}}(X_{t_\ell}).$$

Substituting and applying the Riemann-sum approximation recovers the three terms in (10):

$$\frac{1}{2}J_{\mathrm{diss}}(X_{t_\ell})\,\Delta t - \frac{\sigma^2}{2}J_{\mathrm{diff}}(X_{t_\ell})\,\Delta t + \big(E_{\mathrm{tot}}(X_{t_{\ell+1}}) - E_{\mathrm{tot}}(X_{t_\ell})\big),$$

up to $\mathcal{O}(\Delta t)$ discretization error. At the true (Φ^*, V^*) the residual vanishes in expectation; for misspecified (Φ, V) it measures systematic deviation from the weak-form PDE [WL26].

The loss depends only on individual snapshots, not on particle correspondences across time, and under linear parametrizations of (Φ, V) it is quadratic in the coefficients—reducing estimation to a linear system. When ground-truth labels are available, we additionally report optimal-assignment error and function-space MSE of $\hat{\Phi}, \hat{V}$, developed in Section V.

Remark IV.2.3. Comparison with Other Approaches.
Several alternative strategies have been proposed for learning from unlabeled particle data. These include: (i) alternating label recovery with regression using transport-based pseudo-labels; (ii) minimizing Wasserstein or Sinkhorn distances between simulated and observed ensemble laws; and (iii) maximum-likelihood approaches based on couplings [Lu+19; LL22; Cut13; SST+19]. In contrast, the trajectory-free quadratic loss emphasized here provides a direct algebraic structure when (Φ, V) lie in finite-dimensional linear spaces and circumvents costly inner loops over permutations during training. ⌟

IV.3 Estimation Algorithms for Parametric Setting

This thesis implements learning in the *oracle mean-field basis* of Definition III.3.1: the interaction enters as $\phi(r) = \beta^\intercal \psi(r)$ and the confinement as $V(x) = \theta\, q(x)$ with low-dimensional (β, θ). Two estimation routes are used in the experiments reported later:

- **Labeled velocity matching.** Minimize $\mathcal{J}_{\text{oracle LS}}$ from (7), equivalently the labeled trajectory loss of Section IV when indices are aligned (simulation or pseudo-labels from an alternating loop). This is a single ridge regression in the stacked feature representation.
- **Trajectory-free self-test.** Minimize the quadratic $\mathcal{L}_{\text{TF}}$ in (10) in the same finite-dimensional span; Section IV records the resulting normal equations $(A_D + \lambda I)\hat{\eta} = b_D$ for $\eta = (\beta, \theta)$.

Both objectives are quadratic in (β, θ), so each reduces to a dense ridge solve after assembling A or A_D from radial features ψ and the external template q evaluated on snapshots [HK70].

The closed form derivation for the labeled matching is in (9), while we will derive the trajectory-free version here.

Closed form for A_D and b_D under oracle-linear (Φ, V). With $\phi(r) = \beta^\top \psi(r)$ and $V(x) = \theta\, q(x)$, the predicted drift for particle i is linear in $\eta = (\beta, \theta)$. Define the feature vector for particle i at time t_ℓ as

$$\xi_\ell^i := \begin{pmatrix} \frac{1}{N} \sum_{j \neq i} \psi'(r_{ij})\, u_{ij} \\ \nabla q(X_{t_\ell}^i) \end{pmatrix} \in \mathbb{R}^M,$$

where $r_{ij} = \|X_{t_\ell}^i - X_{t_\ell}^j\|$ and $u_{ij} = (X_{t_\ell}^i - X_{t_\ell}^j)/r_{ij}$ is the unit separation vector, so that the predicted drift satisfies $F_\ell^i(\eta) = (\xi_\ell^i)^\top \eta$, where $\eta = (\beta, \theta)$. The empirical velocities are $y_{t_\ell}^i = -(X_{t_{\ell+1}}^i - X_{t_\ell}^i)/\Delta t$, and their projections onto radial directions give scalar weights

$$w_{i\ell} := \frac{1}{N} \sum_{j \neq i} \langle y_{t_\ell}^i,\, u_{ij} \rangle.$$

Stacking over all particles and times, the self-test loss is quadratic in η and its minimizer satisfies

$$(A_D + \lambda I)\hat{\eta} = b_D,$$

where

$$A_D = \sum_{i,\ell} \xi_\ell^i (\xi_\ell^i)^\top \in \mathbb{R}^{M \times M}, \qquad b_D = \sum_{i,\ell} \xi_\ell^i\, w_{i\ell} \in \mathbb{R}^M.$$

The system admits a unique solution whenever $A_D + \lambda I$ is positive definite [HK70].

Remark IV.3.1. The explicit Euler map $\hat{X}_{t+\Delta t}^i = X_t^i - \hat{b}^i(X_t)\, \Delta t$ with learned $\hat{b}$ feeds the **prediction-informed** costs—this is the bridge from learning to trajectory reconstruction. ⌟

V Trajectory reconstruction via label matching

Here we develop the thesis's **central algorithmic contribution**: reconstruct latent labels between successive unlabeled snapshots by optimal assignment, using costs informed by learned drifts when available. Section II supplies the transport language; Section IV supplies $(\hat{\Phi}, \hat{V})$. The next two subsections reuse least-squares solves from Chapters III and IV inside alternating and trajectory-free matching pipelines; the discrete Kantorovich formulation, Hungarian/Sinkhorn solvers, and prediction-informed costs that follow are the methodological novelty emphasized below.

At each coarse time we observe an unordered cloud $\{x_{t_\ell,1}, \dots, x_{t_\ell,N}\}$. Trajectory reconstruction chooses a permutation $\pi_\ell \in S_N$ between consecutive times, then composes these maps across ℓ to recover piecewise tracks (up to one global relabeling at t_0). We evaluate against simulator ground truth when available. Throughout, matching is posed as discrete OT/assignment with equal masses $1/N$.

Closed-form quadratic minimization before matching Ridge-normal equations $(A^\intercal A + \lambda \,\mathrm{Id})^{-1} A^\intercal b$ for coefficient vectors (β, θ) were already derived in Section III (oracle velocity loss) and assembled explicitly in previous section.

V.1 The Label Assignment Problem

At each coarse time t_ℓ we observe an unordered multiset $\{X^j_{t_\ell}\}_{j=1}^N$ and similarly at $t_{\ell+1}$. A bijection $\pi_\ell \in S_N$ pairs indices across from the previous time step. The assignment problem is to choose π_ℓ that minimizes a sum of edge costs $\sum_{i=1}^N C_{i,\pi_\ell(i)}$ for a nonnegative matrix $C \in \mathbb{R}^{N\times N}_{\geq 0}$ chosen from data and, when available, from a learned model.

Hereby, a natural choice is *minimum Euclidean transport cost*, *i.e.* $C_{ij} = \|x_{t_\ell,i} - x_{t_{\ell+1},j}\|^2$ (or the distance itself). With a learned drift $(\hat{\Phi}, \hat{V})$, we prefer costs that measure how well position j at $t_{\ell+1}$ agrees with a *one-step prediction* started from i at t_ℓ.

Definition V.1.1. Discrete Kantorovich coupling.
Given two snapshots at consecutive times t_ℓ and $t_{\ell+1}$, their empirical measures are

$$\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_{t_\ell,i}}, \qquad \nu = \frac{1}{N}\sum_{j=1}^N \delta_{x_{t_{\ell+1},j}}.$$

A *coupling* between μ and ν is a nonnegative matrix $\gamma = (\gamma_{ij})_{i,j=1}^N$ with row sums $\sum_j \gamma_{ij} = 1/N$ and column sums $\sum_i \gamma_{ij} = 1/N$, meaning that $\gamma \in \Pi(\mu, \nu)$ in the sense of optimal transport [Vil09; PC19]. ⌟

This coupling view is the OT formulation of pairwise relabeling.

Definition V.1.2. Discrete quadratic Kantorovich value.
Given a cost matrix $C_{ij} = \|x_{t_\ell,i} - x_{t_{\ell+1},j}\|^2$, the discrete quadratic Kantorovich (or $\mathcal{W}_2^2$) value between μ

and ν is

$$\mathcal{W}_2^2(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \sum_{i,j=1}^{N} \gamma_{ij}\,\|x_{t_\ell,i} - x_{t_{\ell+1},j}\|^2.$$

Because each row of γ sums to $1/N$, this equals $\frac{1}{N}\min_{\pi\in S_N}\sum_{i=1}^{N} C_{i,\pi(i)}$ whenever an optimizer lives on a permutation (equivalently, multiply the usual assignment objective $\sum_i C_{i,\pi(i)}$ by the per-particle mass N^{-1}). ⌟

This discrete formula is the empirical analogue of (1): there $\pi \in \Pi(\mu,\nu)$ is a *probability* coupling on $\mathbb{R}^d \times \mathbb{R}^d$, whereas here each γ_{ij} is the mass transported from particle i at t_ℓ to particle j at $t_{\ell+1}$, normalised so every row and column sums to $1/N$. The factor N^{-1} outside $\sum_i C_{i,\pi(i)}$ is the same mass-per-particle bookkeeping.
When both empirical measures have equal mass $1/N$, any *permutation coupling* $\gamma_{ij} = \frac{1}{N}\mathbf{1}\{j = \pi(i)\}$ for some $\pi \in S_N$ is an extreme point of $\Pi(\mu,\nu)$. Optimizers for the quadratic cost often lie at such extreme points, yielding the displayed reduction with the explicit N^{-1} prefactor carried outside the discrete sum.

Given the true simulator trajectories $\{X_{t_\ell}^i\}_{i=1}^N$ and a candidate assignment $\pi \in S_N$ pairing particles from t_ℓ to $t_{\ell+1}$, we need a way to evaluate how accurately the proposed permutation recovers the latent particle identities.

Definition V.1.3. Permutation Loss / Misassignment Rate.
The *permutation loss* (or *misassignment rate*) of an assignment $\pi \in S_N$ is defined as

$$\mathrm{Loss}(\pi) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\, X_{t_{\ell+1}}^i \neq x_{t_{\ell+1},\pi(i)} \,\},$$

that is, the proportion of particles whose assigned label at $t_{\ell+1}$ does not correspond to their ground-truth location. ⌟

This will be the **main criterion** for us to evaluate how well the algorithm is doing in terms of matching.

Lower values of $\mathrm{Loss}(\pi)$ indicate more faithful reconstruction of the hidden trajectories. In experiments, this is reported as the **matching error** or **misassignment ratio** relative to ground truth. For random assignment, the expected loss is close to $1 - 1/N$; for a perfect reconstruction, it is 0.

V.2 Linear assignment and the Hungarian algorithm

The **linear assignment problem** seeks the permutation $\pi \in S_N$ that minimizes the total cost $\sum_{i=1}^{N} C_{i,\pi(i)}$ for a given cost matrix C. This problem is central to trajectory reconstruction, as it dictates how to assign particle indices between consecutive snapshots to best reflect plausible physical movement. The cost matrix C can be chosen as the squared Euclidean distance matrix $C_{ij} = \|x_{t_\ell,i} - x_{t_{\ell+1},j}\|^2$.

A classical and efficient method to solve the linear assignment problem is the Kuhn–Munkres (Hungarian) algorithm, which guarantees an exact minimizer in $\mathcal{O}(N^3)$ arithmetic operations [Kuh55; Mun57]. The

output is a **hard** labeling: each past particle index i is matched to exactly one future index $\pi(i)$.

Algorithm V.2.1. Hungarian Method for Linear Assignment.

Require: Cost matrix $C \in \mathbb{R}^{N\times N}$
Ensure: Permutation $\pi \in S_N$ minimizing $\sum_{i=1}^{N} C_{i,\pi(i)}$

1: **Step 1: Row Reduction.** For each row i, subtract $\min_j C_{ij}$ from every entry in row i.
2: **Step 2: Column Reduction.** For each column j, subtract $\min_i C_{ij}$ from every entry in column j.
3: **Step 3: Minimum zero cover.** Among rows and columns, choose a minimum-cardinality family covering every zero entry of the reduced matrix. Equivalently, form the bipartite graph of zero positions and compute a maximum matching together with the dual row/column cover of matching size supplied by Kőnig's theorem; standard presentations of the Hungarian method explain how this certificate drives Step 4 [BDM12; Mun57].
4: **if** the minimum number of covering lines is N **then**
5: **Proceed to Step 5.**
6: **else**
7: **Step 4: Adjust Matrix.**
8: Find the smallest uncovered entry m in C.
9: Subtract m from every uncovered element.
10: Add m to every element covered twice (by both a row and a column).
11: Return to Step 3.
12: **end if**
13: **Step 5: Assignment.**
14: Extract an optimal permutation π from the dual certificate / maximum matching computed alongside Step 3 (equivalently select N independent zeros whose existence is guaranteed when Step 3 admits exactly N covering lines); standard expositions give the greedy completion [BDM12; Mun57].

Remark V.2.2. Runtime Analysis of the Hungarian Algorithm.
The classical Hungarian algorithm for solving the linear assignment problem operates on a cost matrix $C \in \mathbb{R}^{N\times N}$. Its runtime complexity is $\mathcal{O}(N^3)$ in the generic dense case. This cubic scaling arises as follows:

- Each reduction step (row and column minimization) costs $\mathcal{O}(N^2)$.
- The covering and adjustment proceduresidentifying the minimum number of lines covering all zeros, updating uncovered and covered entries, and iteratively searching for augmenting pathsrequire, in the worst case, traversing the entire matrix up to N times.
- Altogether, the standard implementation (e.g., Munkres' algorithm) completes in at most $O(N^3)$ arithmetic operations [Kuh55; Mun57].

In practice, the cost of the Hungarian algorithm is often dominated by data movement and memory access for large N. Specialized sparse or structured cost matrices can yield improved runtimes; classical dense implementations nonetheless match the $\Theta(N^3)$ worst-case scaling [BDM12]. For trajectory reconstruction involving N particles per frame, exact hard assignment therefore scales cubically with N, which can be

significant when N is large. ⌟

By contrast, iterative Sinkhorn–Knopp scaling (Section II.3) can operate with $\mathcal{O}(N^2)$ arithmetic per iteration, trading a modest loss of exactness for substantial computational gains, especially in high-dimensional or repeated-matching settings.

V.3 The Sinkhorn Algorithm

Trajectory reconstruction instantiates the entropic Kantorovich problem (3): empirical marginals supported on consecutive clouds, Gibbs kernel $K_{ij} = \exp(-C_{ij}/\varepsilon)$ built from geometric or prediction-informed costs, followed either by soft couplings or hard extraction [Cut13]. The computational heart of entropic optimal transport is then a classical statement about scaling rows and columns of a positive matrix to prescribed marginals [Sin64; PC19].

Theorem V.3.1. Sinkhorn theorem, doubly stochastic scaling.
Let $K \in \mathbb{R}^{N\times N}_{>0}$ have strictly positive entries. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N_{>0}$ satisfy $\sum_i a_i = \sum_j b_j =: s$. Then there exist vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N_{>0}$, unique up to the gauge freedom $\mathbf{u} \mapsto t\mathbf{u}$, $\mathbf{v} \mapsto t^{-1}\mathbf{v}$ for $t > 0$, such that matrix Γ

$$\Gamma_{ij} := u_i \, K_{ij} \, v_j$$

has row sums $\Gamma\mathbf{1} = \mathbf{a}$ and column sums $\Gamma^\intercal\mathbf{1} = \mathbf{b}$, *i.e.*, Γ is a coupling of $\mathbf{a}$ and $\mathbf{b}$.

Remark V.3.2. The unique minimizer of (3) has the Gibbs form $\gamma^\star_{ij} = u^\star_i K_{ij} v^\star_j$, where $(u^\star, v^\star)$ are the Lagrange multipliers for the marginal constraints. Since $K_{ij} = \exp(-C_{ij}/\varepsilon)$ is strictly positive, Theorem V.3.1 guarantees existence and uniqueness of these scaling factors, so the entropic optimal coupling is precisely the Sinkhorn-scaled matrix with the prescribed marginals [Cut13; PC19]. ⌟

The constructive content of Theorem V.3.1 is realized by an efficient iterative algorithm that alternately scales the rows and columns of the positive matrix K to match prescribed marginals $\mathbf{a}$ and $\mathbf{b}$. This procedure, known as the *Sinkhorn–Knopp algorithm*, produces scaling vectors $(\mathbf{u}, \mathbf{v})$ such that the matrix Γ has the desired row and column sums.

The algorithm is widely used for entropic optimal transport due to its simplicity, parallelizability, and significantly lower arithmetic cost compared to exact assignment solvers, particularly for large N.

Below, we summarize the algorithm in pseudocode.

Algorithm V.3.3. Sinkhorn–Knopp Iteration for Matrix Scaling.

Require: Kernel $K \in \mathbb{R}^{N\times N}_{>0}$, target marginals $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N_{>0}$

1: Initialize $\mathbf{v}^{(0)} \leftarrow \mathbf{1} \in \mathbb{R}^N_{>0}$, $m \leftarrow 0$
2: **repeat**
3: **Row update:** $u_i^{(m+1)} \leftarrow \dfrac{a_i}{\sum_{j=1}^N K_{ij} v_j^{(m)}}$ for all $i = 1, \dots, N$
4: **Column update:** $v_j^{(m+1)} \leftarrow \dfrac{b_j}{\sum_{i=1}^N K_{ij} u_i^{(m+1)}}$ for all $j = 1, \dots, N$
5: $m \leftarrow m + 1$
6: **until** convergence (*e.g.*, all row and column sums within tolerances, or to maximum iteration)
7: **Return** $(\mathbf{u}, \mathbf{v})$

At each iteration, the algorithm enforces one set of marginal constraints (rows, then columns), and the updates can be performed independently across entries. For strictly positive K, convergence to the unique solution is guaranteed, and the final matrix $\Gamma_{ij} = u_i K_{ij} v_j$ satisfies the marginal constraints to within the desired tolerance.

The computational cost per iteration is $\mathcal{O}(N^2)$; typically, only a modest number of iterations is needed for high-precision solutions, making the Sinkhorn–Knopp iteration highly attractive for large-scale problems compared to the cubic scaling of the Hungarian algorithm.

For trajectory reconstruction, we run Sinkhorn on $K_{ij} = \exp(-C_{ij}/\varepsilon)$ to obtain $\gamma^\star$ and read off a hard assignment by $\pi(i) = \arg\max_j \gamma^\star_{ij}$.

Remark V.3.4. Conceptually, we can also use the **Sinkhorn divergence** / entropic transport cost $\sum_{ij}(C_{ij}\gamma^\star_{ij} + \varepsilon\gamma^\star_{ij}\log\gamma^\star_{ij})$ as a permutation-invariant score between predicted and observed clouds without committing to a single π [Cut13; PC19].
The temperature ε trades off faithfulness to low-cost edges (small ε) against numerical stability and spread of mass (larger ε). ⌟

V.4 Alternative Matching Algorithm: Match-LSE-Match Loop

This method alternates between (i) matching with current costs and (ii) closed-form LSE parameter updates from the resulting pseudo-labels. Each outer step treats the permutations $\{\pi_\ell^{(k)}\}$ as giving consistent labels across successive snapshots and fits $(\hat{\Phi}, \hat{V})$ by the same labeled velocity least-squares objective as (7) (finite-difference velocities aligned through $\pi_\ell^{(k)}$), *not* the trajectory-free self-test loss.

Conceptually, we utilize the following loop structure:

- Start from geometric matching between successive frames.
- Fit $(\hat{\Phi}, \hat{V})$ by least squares using the current matched pairs.
- Rebuild prediction-informed costs from the fitted drift and rematch.
- Iterate until convergence or a fixed outer-iteration budget.

Algorithm V.4.1. Match-LSE-match alternating loop.

Require: Permuted snapshots $\{x_{t_\ell}\}_{\ell=0}^{L}$, regularization λ, max iterations $K_{\max}$
Ensure: Final matchings $\{\pi_\ell\}$ and learned coefficients $\hat{\eta} = (\hat{\beta}, \hat{\theta})$
Initialize $\pi_\ell^{(0)} \leftarrow \textsc{Hungarian}(C_\ell^{\text{geom}})$ for all ℓ
for $k = 0, \ldots, K_{\max} - 1$ **do**
  Build labeled velocity least-squares design (7) using $\pi_\ell^{(k)}$ to align $X_{t_{\ell+1}}$ with X_{t_ℓ}
  Solve $(A_D + \lambda I)\hat{\eta}^{(k+1)} = b_D$ (closed-form ridge minimiser)
  for each $\ell = 0, \ldots, L - 1$ **do**
    Compute prediction-informed cost $\tilde{C}_\ell^{(k+1)}$ from $\hat{\eta}^{(k+1)}$
    Update $\pi_\ell^{(k+1)} \leftarrow \textsc{Hungarian}(\tilde{C}_\ell^{(k+1)})$
  end for
end for
return $\{\pi_\ell^{(K_{\max})}\}, \hat{\eta}^{(K_{\max})}$

V.5 Trajectory-free learning then matching

This method decouples learning and matching: first fit the model with the trajectory-free objective, then perform matching once using the learned drift:

- **Stage 1 (trajectory-free fitting):** estimate $(\hat{\Phi}, \hat{V})$ from unlabeled snapshots using the self-test objective and its linear-system solution.
- **Stage 2 (matching):** construct prediction-informed costs from the learned drift and solve assignment (Hungarian or Sinkhorn-hard).

Algorithm V.5.1. Trajectory-free learning then matching.

Require: Unlabeled snapshots $\{x_{t_\ell}\}_{\ell=0}^{L}$, basis choice, regularization λ
Ensure: Matched permutations $\{\pi_\ell\}$ and performance metrics
Assemble trajectory-free linear system from energy, dissipation, and diffusion terms
Solve $(A_D + \lambda I)\hat{\eta} = b_D$ for model coefficients
for each $\ell = 0, \ldots, L - 1$ **do**
  Predict $\hat{X}^i_{t_{\ell+1}} = x_{t_\ell,i} - \hat{b}^i(x_{t_\ell})\Delta t$
  ▷ Implementation-wise, we do propagation with dt for $\Delta t / dt$ times
  Form $\tilde{C}_{\ell,ij} = \|\hat{X}^i_{t_{\ell+1}} - x_{t_{\ell+1},j}\|^2$
  Compute π_ℓ via Hungarian assignment
end for
return $\{\pi_\ell\}, \hat{\eta}$

VI Experiment and Results

This section reports synthetic experiments produced by the consolidated pipeline available in the GitHub repository, aligned with the LIPS-style generation and solver conventions (`lips_unlabeled_data` frame-

work [WL26]). Completed archives summarized below comprise **129** setups. Methods are evaluated on *fine* grids (consecutive simulator states at integration step δt) and *coarse* grids produced by subsampling every m fine steps, so coarse snapshots sit at spacing $\Delta t = m\,\delta t$ whenever $m \geq 1$.

VI.1 Methodology Stack and Setups

The experiments now compare the following reconstruction methodologies:

- **Fine/Coarse Hungarian:** exact assignment from geometric costs.
- **Fine/Coarse Sinkhorn:** hard labels extracted from entropic OT couplings.
- **Sinkhorn iteration tracking:** correctness at the final Sinkhorn iterate.
- **Alternating matching-learning:** iterative relabeling and refitting.
- **Learned-kernel rematching:** MLE, Sinkhorn+MLE, and self-test variants on fine/coarse trajectories.

Here m is the coarse subsampling factor (number of fine steps per coarse step), so $\Delta t = m\,\delta t$.
All algorithm definitions and pseudocode are centralized in Section V (including closed-form LSE updates, match-LSE-match alternation, trajectory-free-then-matching, Hungarian, and Sinkhorn). This section reports only empirical outcomes of those methods.

In the experiments, there are various groups declared with different setups, each with a variety of experiments, characterized as following.

Setup IDs	n	Role
`example_01–example_42`	42	Systematic sweep across Morse, Lennard-Jones, Gaussian well (`model_e`), inverse-square (`model_b`), piecewise (`model_a`), smooth Gaussian pair (`model_gauss`), and Kuramoto periodic dynamics; axes include observation gap $\Delta t/\delta t$, σ, pooled multi-system K, and raised dimension ($d \in \{5, 8, 10\}$ on selected IDs).
`lips_s_000–lips_s_059`	40	Automatic spine over {`model_a`, `model_b`, `model_lj`, `model_morse`, `model_e`} $\times$ $d \in \{2, 5, 10, 20\}$ $\times$ `dt_obs` $\in \{10^{-3}, 10^{-2}, 10^{-1}\}$ with N=10, M=2000, L=100, σ=1.
`lips_ct_*`	15	Condition-number style grid [WL26]: `model_e`, $N \in \{5, 10, 20, 50, 100\}$, $d \in \{2, 5, 10\}$, fine δt=10^{-4}, `dt_obs`=10^{-2}, heavy observation gap.
`lips_m2_*`	5	All interaction laws compared at d=2, `dt_obs`=10^{-2}, `dt_fine`=10^{-3} (same spirit as LIPS model sweep scripts).
`lips_l_01–lips_l_06`	6	Large ensemble sizes ($M \approx$ 5k–7k) and horizons ($L \approx$ 220–300) per model/dimension/`dt_obs` tuple for pooled self-test stress tests (high-d cells omit trajectory maps).
`det_l_01–det_l_08`	8	Deterministic particle flows (σ=0) with macroscopic Δt=$\delta t \geq 1$; archive contains `det_l_01`.
`fixed_01–fixed_20`	20	Regression aliases covering Kuramoto and heterogeneous Kuramoto, deterministic benchmarks, Morse with messy relabelled warm-start, multi-system pooling (K>1), capped Lennard-Jones pools, aggressive observation gaps, and Gaussian-interaction showcases.

Table VI.1. Experimental setup groups for different models.

VI.2 Main Experiment Results

Unless noted otherwise, **correctness** means the mean per-step label recovery in $[0, 1]$ relative to the simulator's hidden permutation (excluding the t=0 identity frame). All pooled rows below aggregate the **129** completed archives under `outputs/`.

Method	Fine mean	Coarse mean	Δ fine vs. Hun.	Strict wins vs. Hun.	Share of 129
Hungarian	0.764	0.745	0	—	—
Sinkhorn	0.757	0.738	−0.007	2	1.6%
Alternating match–LSE	0.762	0.745	−0.002	33	25.6%
Learned self-test rematch	0.762	0.743	−0.002	21	16.3%
Oracle kernel matching	0.783	0.766	+0.019	29	22.5%

Table VI.2. Pooled mean correctness and strict fine-grid improvements vs. Hungarian. "Strict wins" counts setups whose fine-grid mean correctness strictly exceeds Hungarian..

Oracle prediction-informed costs average ≈ 0.019 above Hungarian on fine grids in this corpus (≈ 0.021 on coarse) which implies that over the best learning quality, there is slight improvement over the learning procedure. Below are runs where oracle strictly above the best of {Hungarian, Sinkhorn, alternating, self-test} on fine grids; `det_l_01` is the extreme σ=0 case where geometric costs collapse but oracle costs recover every label (Table VI.3).

Setup ID	Hun. f	Sink. f	Alt. f	Self. f	Ora. f	Hun. c	Alt. c	Self. c	Ora. c
`det_l_01`	0.0000	0.0000	0.0096	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000
`example_21`	0.1874	0.1874	0.1904	0.2015	0.2015	0.1974	0.2105	0.1684	0.1684
`example_28`	0.2343	0.2343	0.2364	0.2480	0.2455	0.2368	0.2579	0.2395	0.2395
`example_37`	0.6717	0.6717	0.6742	0.6829	0.6804	0.6797	0.6471	0.6144	0.5948
`example_05`	0.3936	0.3936	0.3894	0.3824	0.3833	0.3575	0.3865	0.4155	0.4179
`lips_s_041`	0.9778	0.9778	0.9818	0.9818	0.9818	0.9778	0.9818	0.9818	0.9818
`lips_s_002`	0.3051	0.3051	0.2990	0.2768	0.2788	0.3051	0.2990	0.2768	0.2788
`lips_l_02`	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
`lips_l_03`	0.7854	0.7854	0.7872	0.7868	0.7872	0.7854	0.7872	0.7868	0.7872
`lips_l_06`	0.7577	0.7577	0.7544	0.7485	0.7485	0.7577	0.7544	0.7485	0.7485
`fixed_10`	0.9949	0.9949	0.9949	0.9949	0.9949	0.9937	0.9939	0.9937	0.9937

Table VI.3. Representative mean correctness (suffix `f`=fine grid, `c`=coarse).

Radial L^2 vs. oracle (alternating vs. self-test). Each archive stores trapezoid L^2 gaps to the oracle radial parameterization on the r-grids behind.

The criterion we use here is to check how well our learned kernel is compared to the oracle kernel using the L^2 loss, *i.e.*, $\int_{\mathbb{R}_{>0}} |\phi - \phi_{\text{true}}|^2$.

Table VI.4 summarizes how often trajectory-free self-test beats alternating in integrated squared error.

Oracle-relative channel	Self-test $<$ Alt. (# of 129)	Approx. share
$\|\hat{\Phi} - \Phi^*\|_{L^2}^2$	104	$\approx 81\%$
$\|(\hat{\Phi})' - (\Phi^*)'\|_{L^2}^2$	97	$\approx 75\%$
$\|\hat{V} - V^*\|_{L^2}^2$	101	$\approx 78\%$
$\|(\hat{V})' - (V^*)'\|_{L^2}^2$	101	$\approx 78\%$

Table VI.4. Runs where self-test attains strictly smaller oracle-relative L^2 than alternating (n=129).

Table VI.5 records typical *relative* shrinkage when self-test wins.

Median ratio $\mathcal{L}^2_{\text{self}}/\mathcal{L}^2_{\text{alt}}$	Value
Φ channel	≈ 0.18
$(d\Phi/dr)$ channel	≈ 0.18
V channel	≈ 0.09
(dV/dr) channel	≈ 0.11

Table VI.5. Typical relative L^2 improvement of self-test over alternating (medians across setups; means are skewed by outliers)..

Here, we also list the differences of these losses in a few given example setups.

Setup	Alt. $\|\Phi-\Phi^*\|_{L^2}^2$	Self. $\|\Phi-\Phi^*\|_{L^2}^2$	Alt. $\|\Phi'-(\Phi^*)'\|_{L^2}^2$	Self. $\|\Phi'-(\Phi^*)'\|_{L^2}^2$
`det_l_01`	8.56×10^2	3.67	2.33×10^4	8.53×10^1
`example_37`	2.04	2.72×10^{-1}	3.85×10^1	5.77
`fixed_10`	4.82×10^2	6.53	1.02×10^1	5.28×10^1

Table VI.6. Illustrative Φ and $(d\Phi/dr)$ oracle L^2 gaps (rounded): alternating spikes even when assignment scores stay close (`fixed_10`); self-test shrinks Φ error in `example_37`..

A specific example of the learning map comparison is shown below for the kernel, the capping issue is elaborated further in section VI.4.

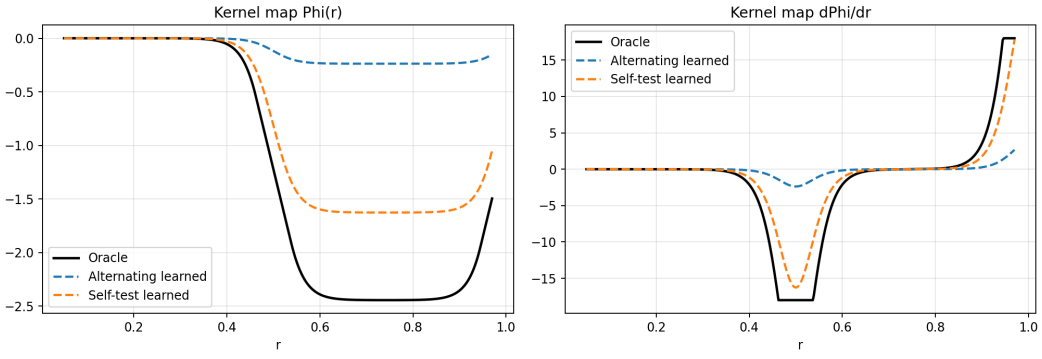

Figure VI.7. Radial kernel maps in `example_37` (`model_a`, d=4, K=8 pooled systems, large $\Delta t/\delta t$): learned self-test tracks oracle Φ more tightly than alternating while improving fine-grid mean correctness over Hungarian (≈ 0.683 vs. ≈ 0.672).

Sinkhorn iterations and cost versus Hungarian. Another perspective we are considering is the computational complexity for the Hungarian algorithm.

The entropic matcher uses $K_{ij} = \exp(-C_{ij}/\varepsilon)$ with $\varepsilon = 0.01 \times \text{median}_{ij}(C_{ij})$, Sinkhorn–Knopp updates until the combined row/column marginal residual in $\|\cdot\|_1$ falls below 10^{-6}, and `max_iter` capped at **40** and **60** in different setups.

Corpus statistic (fine grid unless noted)	Count or value
Setups with per-step *median* Sinkhorn iters in $\{1, 2\}$	54/129
Setups with median at iteration cap	75/129 (63 at cap 40, 12 at cap 60)

Table VI.8. Analysis of Sinkhorn iteration number across all 129 *archives (marginal tolerance* 10^{-6}*,* ε *factor* 0.01*).*

From the above statistics, we see that for the Sinkhorn algorithm, it either converges very quickly ($<< N$) or it would take a long time to converge. As long as we cap `max_iteration` below N, then we can ensure an efficiency in algorithm compared to "Hungarian matching."

Each Sinkhorn round costs two dense matrix–vector products at $\mathcal{O}(N^2)$; the closing Hungarian step costs $\mathcal{O}(N^3)$. When the median iteration count is ≤ 2, scaling plus one polish is often comparable to a single geometric Hungarian call for N; saturating the cap yields $\mathcal{O}(K_{\max}N^2)$ entropic work with $K_{\max} \in \{40, 60\}$ before the final assignment.

VI.3 Interpretation

Prediction-informed costs help most when geometry fails. When particle drift dominates displacement, geometric OT and Hungarian is less powerful. For the deterministic case `det_l_01`: Hungarian and Sinkhorn both score zero while oracle prediction-informed matching recovers every label. This separates intrinsic geometric ambiguity from the headroom available to dynamics-informed costs.

Trajectory-free self-test rematching is regime-specific. Across all 129 setups, self-test correctness tracks Hungarian closely on average, with strict improvements in only 21/129 fine-grid cases. Gains concentrate in the `example_*` sweep where observation gaps are large; in the denser LIPS spine replicas, snapshots already admit stable Euclidean assignment and learned drift adds little.

Alternating refinement is the most consistently competitive learned method. It achieves strict fine-grid wins over Hungarian in 33/129 setups — more than any other learned method — because iterated reassignment couples parameter updates directly to permutation recovery. The tradeoff is sensitivity to initialization and longer computation runtime: when early matchings are poor, the loop can reinforce incorrect labels and worsen performance.

Sinkhorn stays close to Hungarian but does not reliably match it. Fine-grid correctness trails Hungarian by $\approx 7 \times 10^{-3}$ on average, with strict wins in only 2/129 cases. The cause is visible in the iteration telemetry: 54 setups ($\approx$ 42%) converge in one or two Sinkhorn updates, while 75 ($\approx$ 58%) exhaust the scaling cap without tightening marginals. In the latter regime, the terminal Hungarian polish on debiased log-plan costs diverges from purely geometric matching even at small N.

Better kernel recovery does not imply better matching. Self-test learning achieves substantially smaller oracle-relative L^2 error than alternating fits (median component-wise ratios between 0.09 and 0.18), yet the Pearson correlation between kernel L^2 improvement and correctness gain over Hungarian is essentially zero across all 129 setups. Matching quality depends on local gradient accuracy in one-step Euler predictions, not on global functional similarity to the true kernel.

VI.4 Limitations and Future Directions

Throughout these setups, the experiment is limited under a few constraints and room for improvements.

Clustering and Infeasibility Throughout these experiments, there are various cases when the particles very quickly cluster into a few location, and given the noise in impact, the matching itself becomes an infeasible problem, which causes the low matching correctness in various cases, and this can be improved by using better models and prevent very close concentrations in terms of the total time span, in the following setup for `fixed_13`:

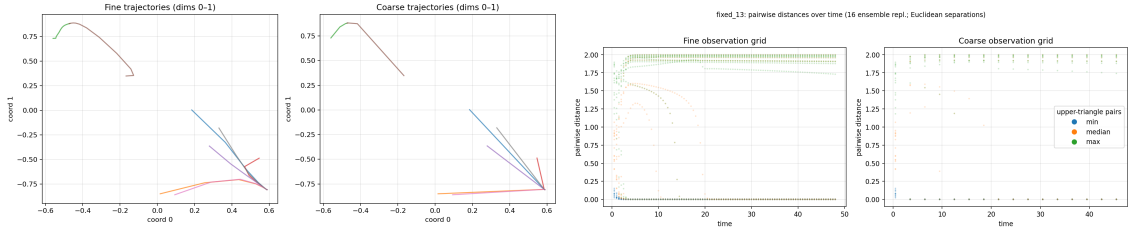

Figure VI.9. Trajectories and Pairwise Distance Distribution in `fixed_13` (`model_gauss`, d=2, L=96): It can be noticed that all the trajectories concentrated.

Enlarged Experiment Size and Models Given the constraint of the computational resources, the "unparametric" learning methods were not explored. Specifically, we have not touched upon kernel learning with a huge number of basis or using neural networks to learn the kernels through the particle system. Meanwhile, a larger K and N in the setup would also enhance the learning quality of the experiments overall.

Singular Kernel and Capping Convention Currently, to handle the singular kernel and prevent explosion of trajectories, we have just added a force cap to the potential and interaction, which would not be ideal for the parametric learning since the actual kernel function with the cap is not contained in the current parameter spaced. It would have been better to handle the singular kernels via other ways and specify the parametric learning settings better for these cases.

For future work, it would be ideal to introduce a variety of new kernels and setup families so that the reconstruction of the trajectories is **feasible** and **meaningful**. On the other hand, it would be better to consider larger model sizes and many real applications would be constrained to larger particle amounts and more variated dynamics.

Moreover, we can use scaling alternating and self-test pipelines to snapshot data without oracle bases; heterogeneous particles; quantitative theory linking energy-dissipation residuals to assignment stability;

applications to single-cell and other real-world unlabeled panels [SST+19; Lav+24; ZMM20].

References

[BDM12] Rainer E. Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems*. Revised reprint. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2012. DOI: 10.1137/1.9781611972238.

[Cut13] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

[FS02] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd. Academic Press, 2002.

[HK70] Arthur E. Hoerl and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. DOI: 10.1137/S0036141096303359.

[Jon24] John Edward Jones. "On the determination of molecular fields. II. From the equation of state of a gas". In: *Proceedings of the Royal Society of London. Series A* 106.738 (1924), pp. 463–477. DOI: 10.1098/rspa.1924.0082.

[KP92] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Vol. 23. Applications of Mathematics. Springer, 1992. DOI: 10.1007/978-3-662-12616-5.

[Kuh55] Harold W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1–2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109.

[Lav+24] Hugo Lavenant et al. "Towards a mathematical theory of trajectory inference". In: *The Annals of Applied Probability* 34.1A (2024), pp. 1–42. DOI: 10.1214/23-AAP1969.

[LL22] Quanjun Lang and Fei Lu. "Learning interaction kernels in mean-field equations of first-order systems of interacting particles". In: *SIAM Journal on Scientific Computing* 44.1 (2022), A260–A285. DOI: 10.1137/20M1377072.

[Lu+19] Fei Lu et al. "Nonparametric inference of interaction laws in systems of agents from trajectory data". In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14424–14433. DOI: 10.1073/pnas.1822012116.

[Mar55] Gisiro Maruyama. "Continuous Markov processes and stochastic equations". In: *Rendiconti del Circolo Matematico di Palermo* 4 (1955), pp. 48–90. DOI: 10.1007/BF02846028.

[McK66] Henry P. McKean. "A class of Markov processes associated with nonlinear parabolic equations". In: *Proceedings of the National Academy of Sciences* 56.6 (1966), pp. 1907–1911. DOI: 10.1073/pnas.56.6.1907.

[Mil+23] Jason Miller et al. "Learning theory for inferring interaction kernels in second-order interacting agent systems". In: *Sampling Theory, Signal Processing, and Data Analysis* 21.1 (2023), p. 21. DOI: 10.1007/s43670-023-00055-9.

[Mun57] James Munkres. "Algorithms for the assignment and transportation problems". In: *Journal of the Society for Industrial and Applied Mathematics* 5.1 (1957), pp. 32–38. DOI: 10.1137/0105003.

[Øks03] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. 6th. Springer, 2003. DOI: 10.1007/978-3-642-14394-6.

[PC19] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Vol. 11. 5–6. Foundations and Trends in Machine Learning, 2019, pp. 355–607. DOI: 10.1561/2200000073.

[Ris96] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. 2nd. Vol. 18. Springer Series in Synergetics. Springer, 1996. DOI: 10.1007/978-3-642-61544-3.

[Sin64] Richard Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *The Annals of Mathematical Statistics* 35.2 (1964), pp. 876–879. DOI: 10.1214/aoms/1177703591.

[SST+19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, et al. "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming". In: *Cell* 176.4 (2019), pp. 928–943. DOI: 10.1016/j.cell.2019.01.006.

[Sze75] Gábor Szeg. *Orthogonal Polynomials*. 4th. Vol. 23. Colloquium Publications. American Mathematical Society, 1975.

[Szn91] Alain-Sol Sznitman. "Topics in propagation of chaos". In: *École d'Été de Probabilités de Saint-Flour XIX—1989*. Vol. 1464. Lecture Notes in Mathematics. Springer, 1991, pp. 165–251. DOI: 10.1007/BFb0085169.

[Tre19] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM, 2019.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer, 2009. DOI: 10.1007/978-3-540-71050-9.

[WL26] Viska Wei and Fei Lu. "Learning interacting particle systems from unlabeled data". In: *arXiv preprint arXiv:2604.02581* (2026).

[WSL25] Xiong Wang, Inbar Seroussi, and Fei Lu. *Optimal minimax rate of learning nonlocal interaction kernels*. 2025. arXiv: 2311.16852 [math.ST]. URL: https://arxiv.org/abs/2311.16852.

[ZMM20] Ming Zhong, Jason Miller, and Mauro Maggioni. "Data-driven discovery of emergent behaviors in collective dynamics". In: *Physica D: Nonlinear Phenomena* 411 (2020), p. 132542. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2020.132542. URL: https://www.sciencedirect.com/science/article/pii/S0167278919308152.