fall into ML 2025

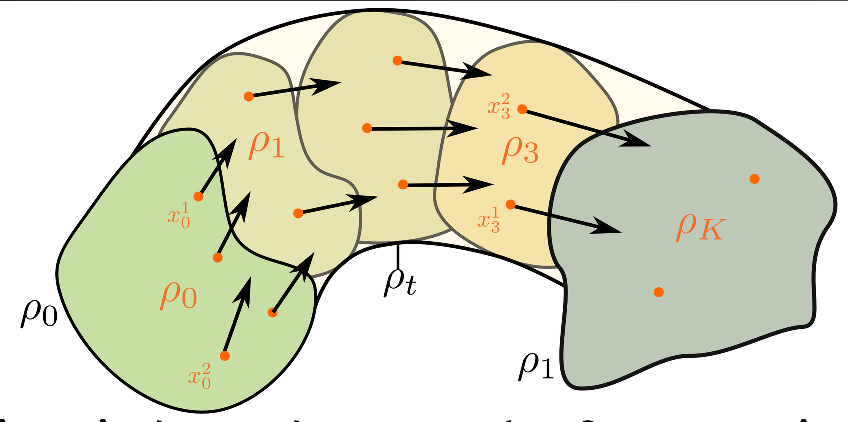
4th conference on machine learning & AI

Learning of Population Dynamics: Inverse Optimization Meets JKO Scheme

Mikhail Persiianov[1] Jiawei Chen[1,2] Petr Mokrov[1] Alexander Tyurin[1,3] Evgeny Burnaev[1,3] Alexander Korotin[1,3]

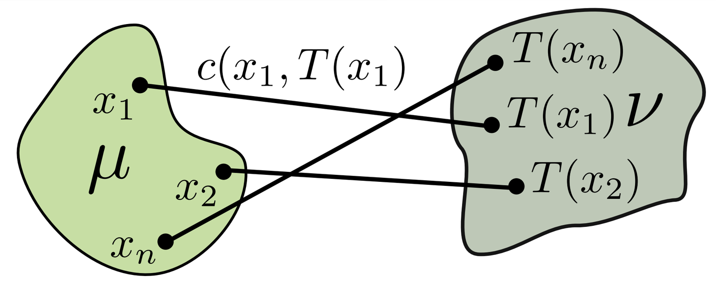
Skoltech[1] MIPT[2] AIRI[3]

Problem Statement



We are given independent samples from marginals $\{\rho_k\}_{k=0}^K$ of the evolving distribution ρ_t at corresponding time points $t_0 < t_1 < \cdots < t_K$. Importantly, each distribution ρ_k may be represented by a different number of samples.

Optimal Transport

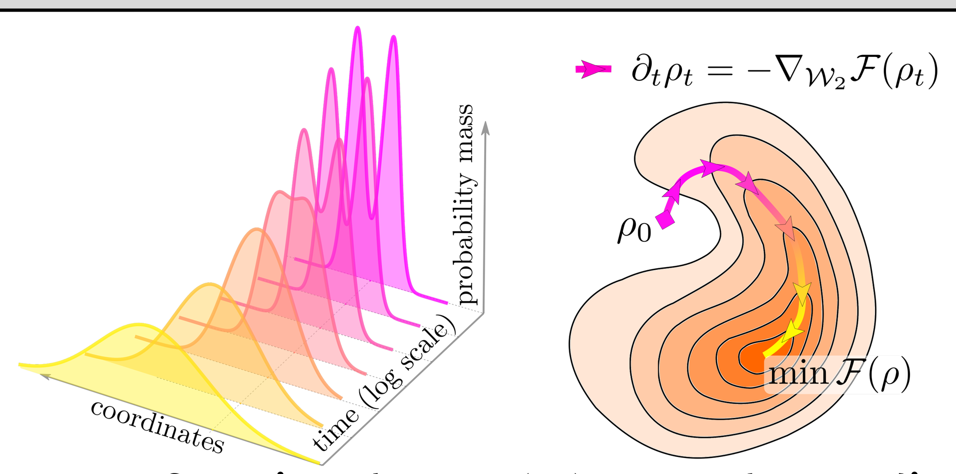


The (squared) *Wasserstein-2 distance* $d_{\mathbb{W}_2}$ between two a.c. probability measures $\mu, \nu \in \mathcal{P}_{ac}(\mathcal{X})$ is defined as:

$$d^2_{\mathbb{W}_2}(\mu, \nu) = \min_{T: T\sharp\mu=\nu} \int_{\mathcal{X}} \|x - T(x)\|_2^2 \, \mathrm{d}\mu(x), \quad (1)$$

where the optimal map T^* is known as the *Monge map*.

Wasserstein Gradient Flows



For an energy functional $\mathcal{F} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, the gradient flow in $\mathbb{W}_2(\mathcal{X})$, called the *Wasserstein gradient flow (WGF)*, is an absolutely continuous curve $\rho_t : \mathbb{R}_+ \to \mathcal{P}(\mathcal{X})$ starting at ρ_0 that follows the steepest descent direction of $\mathcal{J}$:

$$\partial_t \rho_t = -\nabla_{\mathcal{W}_2} \mathcal{J}(\rho_t), \quad \rho_{(t=0)} = \rho_0, \quad (2)$$

where $\nabla_{\mathcal{W}_2} \mathcal{J}(\rho_t)$ denotes the *Wasserstein gradient* in $\mathbb{W}_2$. It can be rewritten in the form of the *continuity equation*, expressing mass conservation under the velocity field v_t:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \; v_t = -\nabla \frac{\delta \mathcal{J}}{\delta \rho}(\rho_t). \quad (3)$$

Examples of PDEs as WGFs

Consider the *free energy* functional:

$$\mathcal{J}_{\mathsf{FE}}(\rho) = \underbrace{\int_{\mathcal{X}} V(x) \mathrm{d}\rho(x)}_{\mathcal{V}(\rho)} + \underbrace{\int_{\mathcal{X}\times\mathcal{X}} W(x-y) \mathrm{d}\rho(x) \mathrm{d}\rho(y)}_{\mathcal{W}(\rho)} + \underbrace{\int_{\mathcal{X}} U(\rho(x)) \mathrm{d}x}_{\mathcal{U}(\rho)}$$

where $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{U}$ correspond to the system's *potential*, *interaction*, and *internal* energies, respectively.
When the energy is $\mathcal{J}_{\mathsf{FP}}(\rho) = \mathcal{V}(\rho) - \beta \mathcal{H}(\rho)$, the resulting PDE is the *Fokker-Planck equation* with diffusion coefficient β:

$$\partial_t \rho_t = \nabla \cdot (\nabla V(x) \, \rho_t) + \beta \nabla^2 \rho_t, \quad (4)$$

which is equivalent to the following *Itô SDE*:

$$\mathrm{d}X_t = -\nabla V(X_t) \, \mathrm{d}t + \sqrt{2\beta} \, \mathrm{d}W_t. \quad (5)$$

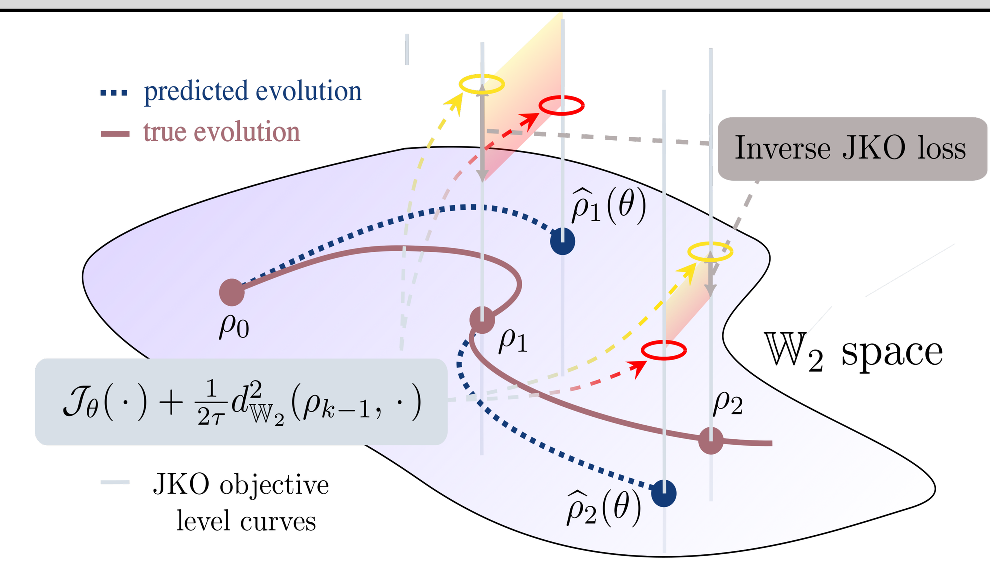
The JKO Scheme

Jordan, Kinderlehrer, and Otto extended the idea of *implicit Euler scheme* to the space of probability measures, introducing a variational time discretization of the Fokker-Planck equation (5), now known as the *JKO scheme*:

$$\rho^\tau_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathcal{X})} \left\{ \mathcal{J}(\rho) + \frac{1}{2\tau} d^2_{\mathbb{W}_2}(\rho, \rho^\tau_k) \right\} = \mathrm{JKO}_{\tau\mathcal{J}}(\rho^\tau_k), \quad \rho^\tau_0 = \rho_0$$

where $\tau > 0$ is the time step. As $\tau \to 0$, the sequence $\rho^\tau_k, k \in \mathbb{N}$ converges to the continuous solution ρ_t of (4), which motivates **our assumption** that the ground truth sequence of measures $\{\rho_k\}_{k=0}^K$ follows $\rho_{k+1} = \mathrm{JKO}_{\Delta t_k \mathcal{J}^*}(\rho_k)$.

iJKOnet Method



Thanks to assumption $\rho_{k+1} = \mathrm{JKO}_{\tau\mathcal{J}^*}(\rho_k)$, we can derive an inequality that becomes an equality if a candidate functional $\mathcal{J}$ matches the ground truth functional $\mathcal{J}^*$:

$$\min_{\rho} \left\{ \mathcal{J}(\rho) + \frac{1}{2\tau} d^2_{\mathbb{W}_2}(\rho_k, \rho) \right\} \le \mathcal{J}(\rho_{k+1}) + \frac{1}{2\tau} d^2_{\mathbb{W}_2}(\rho_k, \rho_{k+1}). \quad (8)$$

Moving the right-hand side to the left yields an expression that is always upper-bounded by zero, regardless of the choice of $\mathcal{J}$. Maximizing the resulting gap, we obtain **loss**:

$$\max_{\mathcal{J}} \min_{T^k} \sum_{k=0}^{K-1} \left[\mathcal{J}(T^k \sharp \rho_k) - \mathcal{J}(\rho_{k+1}) + \frac{1}{2\tau} \int_{\mathcal{X}} \|x - T^k(x)\|_2^2 \rho_k(x) \, \mathrm{d}x \right].$$

Related JKO-based Methods

JKOnet formulates the task of population dynamics recovery as a bi-level optimization problem aimed at minimizing the discrepancy between observed distributions ρ_k and model predictions $\hat{\rho}_k$:

$$\mathcal{L}_{\mathtt{JKOnet}}(\theta, \varphi) = \sum_{k=0}^{K-1} d^2_{\mathbb{W}_2}(\hat{\rho}_k, \rho_k), \quad \text{s.t. } \hat{\rho}_0 = \rho_0, \quad \hat{\rho}_{k+1} = \nabla \psi^*_k \sharp \hat{\rho}_k,$$

$$\psi^*_k \stackrel{\text{def}}{=} \arg\min_{\varphi: \psi_\varphi \in \mathtt{CVX}} \mathcal{J}_\theta(\nabla \psi_\varphi \sharp \hat{\rho}_k) + \frac{1}{2\tau} \int_{\mathcal{X}} \|x - \nabla \psi_\varphi(x)\|^2 \, \mathrm{d}\hat{\rho}_k$$

where CVX denotes the set of continuously differentiable convex functions from $\mathcal{X}$ to $\mathbb{R}$.
JKOnet* addresses the limitations of the original JKOnet by replacing the full JKO optimization problem with its first-order optimality conditions. The method minimizes the following objective, utilizing a *precomputed* optimal transport plans π_k between ρ_k and ρ_{k+1}:

$$\mathcal{L}_{\mathtt{JKOnet^*}}(\theta) = \sum_{k=0}^{K-1} \int_{\mathcal{X}\times\mathcal{X}} \Big\| \nabla V_{\theta_1}(x_{k+1}) + \int_{\mathcal{X}} \nabla U_{\theta_2}(x_{k+1} - y_{k+1}) \mathrm{d}\rho_{k+1}(y_{k+1}) + \theta_3 \frac{\nabla \rho_{k+1}(x_{k+1})}{\rho_{k+1}(x_{k+1})} + \frac{1}{\tau}(x_{k+1} - x_k) \Big\|^2 \mathrm{d}\pi_k(x_k, x_{k+1}).$$

Theoretical Aspects

Theorem (Quality bounds for recovered potential energy). *Let $\varepsilon(V) \doteq \mathcal{L}(V^*, T_{V^*}) - \mathcal{L}(V, T_V)$ be the gap between the optimal and optimized value of inverse JKO loss with internal $\min_T$ problem solved exactly, i.e., $T_V \doteq \min_T \mathcal{L}(V, T)$. Let $\mathcal{X}$ be a convex set; (modified) potentials $V_q := \tau V + \frac{1}{2}\|\cdot\|_2^2 : \mathcal{X} \to \mathbb{R}$ be strictly convex and $\frac{1}{\beta}$-smooth. Then there exists $C = C(\tau, \beta)$ such that following inequality holds:*

$$\int_{\mathcal{X}} \left\| \nabla V^*(y) - \nabla V(y) \right\|^2 \mathrm{d}\rho_1(y) \le C\varepsilon(V).$$

Single-Cell Data

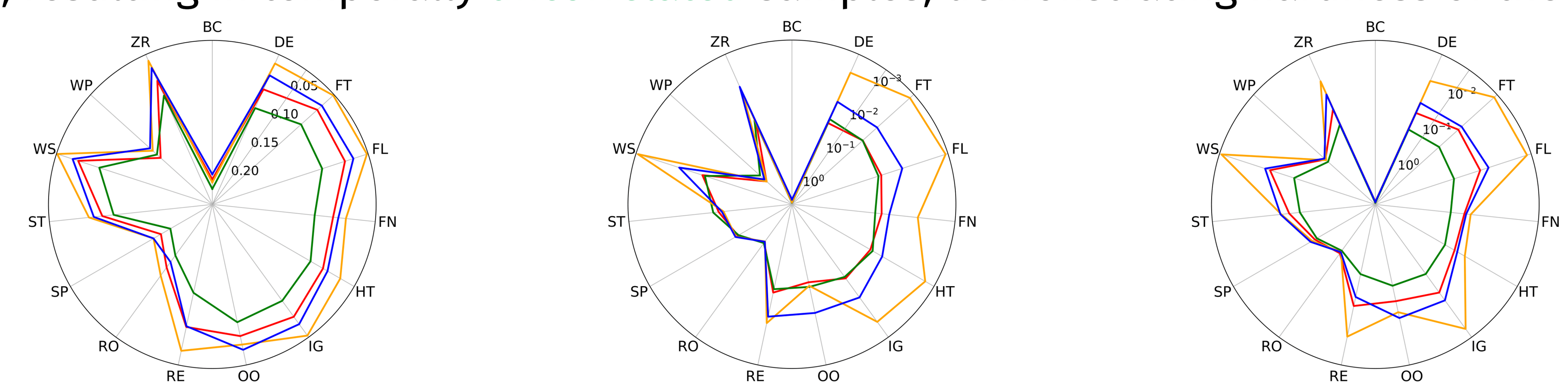
Setup: We perform experiments using a leave-two-out setup. Since the **EB** dataset contains five timesteps, we remove the second (t_2) and fourth (t_4) timesteps, and then evaluate how well our method can reconstruct the data from the remaining t_1 and t_3 timesteps.

Method	t_2	t_4
Vanilla-SB	1.49 ± 0.063	1.55 ± 0.034
DMSB	1.13 ± 0.082	1.45 ± 0.16
TrajectoryNet	2.03 ± 0.04	1.93 ± 0.08
MMSB	1.27 ± 0.028	1.57 ± 0.048
JKOnet$^*_V$	1.145 ± 0.033	2.529 ± 0.014
iJKOnet$_V$ **(Ours)**	$\mathbf{1.082 \pm 0.011}$	$\mathbf{1.147 \pm 0.001}$
JKOnet$^*_{t,V}$	4.414 ± 1.499	2.771 ± 0.197
iJKOnet$_{t,V}$ **(Ours)**	$\mathbf{0.983 \pm 0.037}$	$\mathbf{0.849 \pm 0.021}$

$5D$ experiment. $\mathcal{W}_2$ distance (↓) comparison across t_2 and t_4.

Learning Potential Energy

Goal: Evaluate how our method learns potentials in the **unpaired** setup, i.e., particle trajectories are resampled across time steps, resulting in temporally uncorrelated samples, demonstrating hardness of the corrected setup.



a) EMD b) $Bd^2_{\mathbb{W}_2}$-UVP c) $\mathcal{L}_2$-UVP

Goal: Demonstrate how our method learns potentials in the **paired** setup, i.e., particle trajectories are preserved across time steps, resulting in temporally correlated samples, enabling a direct visual comparison with JKOnet*.