

OPTIMAL TRANSPORTATION BY ORTHOGONAL COUPLING DYNAMICS

Mohsen Sadr, Peyman Mohajerin Esfahani and Hossein Gorji

ABSTRACT. Many numerical and learning algorithms rely on the solution of the Monge-Kantorovich problem and Wasserstein distances, which provide appropriate distributional metrics. While the natural approach is to treat the problem as an infinite-dimensional linear programming, such a methodology limits the computational performance due to the polynomial scaling with respect to the sample size along with intensive memory requirements. We propose a novel alternative framework to address the Monge-Kantorovich problem based on a projection type gradient descent scheme. The dynamics builds on the notion of the conditional expectation, where the connection with the opinion dynamics is leveraged to devise efficient numerical schemes. We demonstrate that the resulting dynamics recovers random maps with favourable computational performance. Along with the theoretical insight, the proposed dynamics paves the way for innovative approaches to construct numerical schemes for computing optimal transport maps as well as Wasserstein distances.

1. INTRODUCTION

1.1. Background

The importance of the optimal transport lies in its widespread application at different fronts of computational studies, along with its unique theoretical properties. It plays an essential role in machine-learning by offering relevant metrics for comparing different distributions [5, 42, 26]. It has led to a significant progress in the theory of partial differential equations by its by-product, Wasserstein gradient flows [2, 37, 59]. At the same time, existing algorithms to compute such distances and corresponding maps remain expensive and complex. Given the versatile role of optimal transport in different branches of machine-learning [35, 1, 5], analysis [66, 37, 58], density functional theory [18, 24], optimization [45], and inference [33], among others, several numerical approaches have been pursued for its efficient computations. Classical methods were developed based on linear programming [14], gradient descent [19], dynamic flow formulation [10, 65], or elliptic solvers [60]. For efficient computation and improved performance-scaling with respect to the dimension, entropy regularization [25, 53] and its stochastic treatment [7] have been pursued. Approximate approaches using moment formulation [47, 57] of Monge-Kantorovich problem have been proposed, as well.

Date: December 14, 2025.

Corresponding author: Hossein Gorji.

Emails: msadr@mit.edu, p.mohajerinesfahani@utoronto.ca, Mohammadhossein.Gorji@empa.ch. Mohsen Sadr: Massachusetts Institute of Technology, Cambridge, USA, and Paul Scherrer Institute, Villigen, Switzerland. Peyman Mohajerin Esfahani: University of Toronto, Canada, and Delft University of Technology, the Netherlands. Hossein Gorji: Laboratory for Computational Engineering, Empa, Dübendorf, Switzerland.

Despite these progresses, to the best of our knowledge, there is no concise Ordinary Differential Equation (ODE), in the sense of particle/agent based dynamics, which leads to the solution of the (un-regularized) Monge-Kantorovich problem. Such a model would be particularly valuable, beyond theoretical and conceptual contributions, for its potential to further enable efficient optimal transport computations. In the opening chapter of Villani’s seminal book [66], a list of coupling dynamics, of ODE or Stochastic Differential Equation (SDE) forms, are reviewed with the intention of creating correlations between two given random variables. While interesting results can be obtained e.g. by the Knoth-Rosenblatt rearrangement [28] or Moser’s coupling [17], such schemes remain limited; e.g. due to coordinate dependency of the Knoth-Rosenblatt rearrangement, or the constraint of small deviation between the marginals for Moser’s coupling.

In principle, finding the optimal transport map between two probability spaces is a global optimization in the sense of linear programming on an infinite dimensional space. We are confronted with the question whether there exist local dynamic rules that can lead to the optimal transportation between two sample spaces? An interesting class of dynamic processes is given by orthogonal dynamics. They naturally arise in projecting Hamiltonian systems onto a sub-space, e.g. by the use of conditional expectation [21, 32]. These processes have an interesting feature that the remainder/unresolved portion remains orthogonal to the current state of the resolved variable.

Recent study of Conforti et al. [23] leverages the projection method in the context of Langevin dynamics. Their proposed stochastic dynamics leads to the coupling which converges to the Sinkhorn regularization of the optimal transport problem [25]. Such a stochastic dynamics, which is closely linked to the Schrödinger Bridge [31, 43, 20], gives rise to the Fokker-Planck equation, evolving the joint distribution. It remains to be addressed what happens to the introduced projected Langevin dynamics when the contribution of Brownian motion vanishes, relevant for the (un-regularized) Monge-Kantorovich problem.

The simplest form of the orthogonal dynamics can be constructed by projection of the gradient descent (with respect to the Monge-Kantorovich cost) onto a sub-space where marginals are preserved. In the absence of entropy regularization, this dynamics becomes a good candidate to create optimal transport between two measure spaces. At the level of distribution, one expects such a dynamics gives rise to the Vlasov equation for the evolution of the joint. Despite the theoretical appeal of the coupling through projection, numerical treatment of the conditional expectation can become prohibitive. In general, conditional expectation is the solution of the Bayesian regression [48]. The Hegselmann-Krause model of opinion dynamics [36, 44] offers non-parametric numerical approach to deal with such coupling dynamics. With these observations in hand and leveraging the recent study [23], we pursue a projected dynamic approach as a concise solution algorithm of the Monge-Kantorovich problem.

1.2. Main Contributions

We propose a numerical scheme for the Monge-Kantorovich problem built on the following steps.

- (1) **Coupling Dynamics.** We introduce the dynamics in Eq. (14) of section 2.1, motivated by projection of the gradient descent on a marginal-preserving tangent space.
- (2) **General Structural Properties.** In section 2.2, assuming that the well-posedness of the dynamics is granted, we justify several structural properties for the general cost setting.
 - (a) *Marginal Preservation.* We prove in Proposition 2.1 that the introduced dynamics preserves the marginals.
 - (b) *Descent in Cost.* The dynamics has a monotone behaviour in decreasing the cost, as shown in Proposition 2.2.
 - (c) *Instability of Sub-optimal Couplings.* We find that the sub-optimal maps are unstable as proved in Proposition 2.3.
 - (d) *McKean-Vlasov System.* We derive the corresponding kinetic description of the introduced evolution in Theorem 2.4, where the kinetics takes the form of the Vlasov equation (20).
- (3) **Key Features in L^2 -Monge-Kantorovich.** Next, we turn our attention to the common case of L^2 -cost and derive two results in section 2.3.
 - (a) *Symmetric Positive Semi-Definite Correlation.* We observe that the dynamics keeps the centered cross-correlation between samples of the two measure spaces in the cone of symmetric positive semi-definite matrices, as proved in Proposition 2.5.
 - (b) *Sharp Descent and Variational Formulation.* We derive the variational form of the dynamics, which according to Theorem 2.6 gives the sharpest descent in the L^2 -Monge-Kantorovich cost among a class of marginal-preserving dynamics.
- (4) **Characterization in Gaussian Setting.** In section 4, we study the Riccati dynamics governing the cross-correlation of Gaussian marginals under L^2 -OCD and prove in Proposition 4.1, that the algorithm converges to the true OT if the marginal covariances commute. We further justify that the convergence is exponentially fast in Theorem 4.3.
- (5) **Numerical Algorithm.** In section 5, after motivating the analogy with opinion dynamics, a non-parametric recipe is proposed for computing the conditional expectation.
- (6) **Numerical Experiments.** We perform several numerical experiments to verify the convergence of the dynamics, as well as its relevance in practical settings, in section 6. We demonstrate that the presented dynamics, equipped with the devised numerical algorithm, successfully recover nonlinear Monge maps. Furthermore, we showcase its performance in distribution learning and color interpolation.

Our focus remains on basic structural properties of the dynamics along with its computational implications. We demonstrate the relevance of the orthogonal coupling dynamics and leave more thorough theoretical and numerical analysis of the dynamics for future studies.

1.3. Notation and Setup

Let $\mathcal{P}_2(\Omega)$ be the set of Borel measures over Ω with a finite second moment, i.e. $\mathcal{P}_2(\Omega) := \{\mu \in \mathcal{P}(\Omega) : \int_{\Omega} \|x\|_2^2 \mu(dx) < \infty\}$, where $\|\cdot\|_2^2$ is the usual L^2 Euclidean norm, and for $f \in L^2(\mu; \mathbb{R}^n)$ let

$$\|f\|_{L^2(\mu)} := \left(\int_{\mathbb{R}^n} \|f(x)\|_2^2 \mu(dx) \right)^{1/2} \quad (1)$$

be the L^2 norm with respect to the measure μ . We write $\mu \otimes \nu$ for the product measure on $\mathbb{R}^n \times \mathbb{R}^n$. For vectors $A, B \in \mathbb{R}^n$, the symbol $A \otimes B$ denotes the rank-one matrix AB^\top . Suppose $\Pi(\mu, \nu)$ is the set of all joint measures in $\mathcal{P}_2(\mathbb{R}^n \times \mathbb{R}^n)$ with marginals $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^n)$. We denote by π^{opt} the solution of the Monge-Kantorovich problem

$$\pi_{\mu, \nu}^{\text{opt}} := \arg \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x, y) \pi(dx, dy) \quad (2)$$

with $c : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ as a non-negative smooth function (see e.g. [13] for details). For the standard setting of $c(x, y) = \frac{1}{2} \|x - y\|_2^2$, the optimal map $\hat{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ exists, if μ and ν are absolutely continuous with respect to the Lebesgue measure [2]. It gives $\hat{T} \# \mu = \nu$ ($\#$ is the push-forward) and accordingly the 2-Wasserstein distance reads

$$d^2(\mu, \nu) = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \pi_{\mu, \nu}^{\text{opt}}(dx, dy) \quad (3)$$

$$= \frac{1}{2} \int_{\mathbb{R}^n} \|\hat{T}(x) - x\|_2^2 \mu(dx) . \quad (4)$$

Let $\mathcal{H} = \mathcal{L}^2(\Omega, \mathcal{A}, P)$ be the Hilbert space comprised of square-integrable \mathcal{A} -measurable functions on Ω . Suppose $X_t, Y_t : \Omega \rightarrow \mathbb{R}^n$ are random processes in \mathcal{H} with the index $t \in [0, +\infty)$ and joint law p_t . We employ the conditional expectation $\mathbb{E}_{p_t}[X_t|Y_t]$ as the orthogonal projection of X_t onto the span of $\sigma(Y_t)$ -measurable functions in \mathcal{H} , denoted by $P_{p_t, Y_t}[X_t]$, where $\sigma(A)$ is the smallest σ -algebra with respect to which the random variable A is measurable (see e.g. [22, 15] for further details). Accordingly, we have

$$P_{p_t, X_t}[Y_t] = \mathbb{E}_{p_t}[Y_t|X_t] \quad \text{and} \quad P_{p_t, Y_t}[X_t] = \mathbb{E}_{p_t}[X_t|Y_t] . \quad (5)$$

We consider the probability flows following the continuity equation of the form

$$\partial_t p_t + \nabla \cdot (v p_t) = 0 , \quad (6)$$

initialized by $p_0 \in \Pi(\mu, \nu)$, where the velocity vector $v = (v_1, v_2)$ is a functional of p_t and belongs to the tangent space [3]

$$\text{Tan}_{p_t} \mathcal{P}_2(\mathbb{R}^{2n}) := \overline{\{\nabla h : h \in C_c^\infty(\mathbb{R}^{2n})\}}^{L^2(p_t; \mathbb{R}^{2n})} . \quad (7)$$

Note that $\text{Tan}_{p_t} \mathcal{P}_2(\mathbb{R}^{2n})$ is a Hilbert space comprised of gradient vector fields. Besides $\text{Tan}_{p_t} \mathcal{P}_2(\mathbb{R}^{2n})$, we specifically consider two tangent spaces $\text{Tan}_{p_t, 1} \Pi(\mu, \nu)$ and $\text{Tan}_{p_t, 2} \Pi(\mu, \nu)$. The former, following [23], reads

$$\text{Tan}_{p_t, 1} \Pi(\mu, \nu) := \left\{ v \in \text{Tan}_{p_t} \mathcal{P}_2(\mathbb{R}^{2n}) : \int_{\mathbb{R}^{2n}} v \cdot \begin{pmatrix} \nabla h_1 \\ \nabla h_2 \end{pmatrix} dp_t = 0, \forall h_{1,2} \in C_c^\infty(\mathbb{R}^n) \right\} , \quad (8)$$

which gives the set of all velocities that preserve the marginals μ and ν . Following notation of [23], the latter tangent space is given by

$$\text{Tan}_{p_t, 2} \Pi(\mu, \nu) := \left\{ v \in \text{Tan}_{p_t} \mathcal{P}_2(\mathbb{R}^{2n}) : \begin{pmatrix} \mathbb{E}_{p_t}[v_1(X, Y)|X] \\ \mathbb{E}_{p_t}[v_2(X, Y)|Y] \end{pmatrix} = 0, \text{ a.s.} \right\} , \quad (9)$$

which is a sub-space of $\text{Tan}_{p_t, 1} \Pi(\mu, \nu)$. We denote it by $\ker(P_{p_t, X}) \times \ker(P_{p_t, Y})$ for simplicity and as shown in Proposition 2.1, it preserves the marginals μ and ν . Since for $v \in \text{Tan}_{p_t, 2} \Pi(\mu, \nu)$,

we have $\mathbb{E}_{p_t}[v_1(X_t, Y_t)|X_t] = \mathbb{E}_{p_t}[v_2(X_t, Y_t)|Y_t] = 0$, the projection operator $S_{p_t} : \text{Tan}_{p_t}\mathcal{P}_2(\mathbb{R}^{2n}) \rightarrow \text{Tan}_{p_t,2}\Pi(\mu, \nu)$ is orthogonal to P_{p_t} , i.e.

$$S_{p_t, X_t}[v_1(X_t, Y_t)] = v_1(X_t, Y_t) - \mathbb{E}_{p_t}[v_1(X_t, Y_t)|X_t] \quad (10)$$

$$\text{and } S_{p_t, Y_t}[v_2(X_t, Y_t)] = v_2(X_t, Y_t) - \mathbb{E}_{p_t}[v_2(X_t, Y_t)|Y_t], \quad (11)$$

for $v \in \text{Tan}_{p_t}\mathcal{P}_2(\mathbb{R}^{2n})$.

2. ORTHOGONAL COUPLING

2.1. Main Dynamics

Starting with independent samples of $X_0 \sim \mu$ and $Y_0 \sim \nu$, we would like to construct a micro-dynamics in the sense of an ODE on X_t and Y_t , which converges to the solution of (2). In other words, we look for a velocity vector v which updates X_t and Y_t in a way that the minimum cost $c(X_t, Y_t)$, in the average sense, is attained as $t \rightarrow \infty$. One can consider the conventional gradient descent for X_t and Y_t with respect to c , suggesting $v_1^{\text{gd}} = -\nabla_x c$ and $v_2^{\text{gd}} = -\nabla_y c$, and updating X_t and Y_t according to v_1^{gd} and v_2^{gd} , respectively. However, it is easy to spot that the distributions of X_t and Y_t would go out of the set of couplings $\Pi(\mu, \nu)$.

An interesting approach to keep the marginals in $\Pi(\mu, \nu)$ is to project the dynamics of the gradient descent onto the tangent space which preserves the marginals. In particular, if we consider the probability flow given by

$$\partial_t p_t + \nabla \cdot (\hat{v} p_t) = 0, \quad (12)$$

and restrict the velocities to those $\hat{v} = (\hat{v}_1, \hat{v}_2) \in \text{Tan}_{p_t}\mathcal{P}_2(\mathbb{R}^{2n})$ which fulfill

$$\int_{\mathbb{R}^{2n}} (\hat{v}_1 \cdot \nabla h_1(x) + \hat{v}_2 \cdot \nabla h_2(y)) p_t(dx, dy) = 0 \quad \forall h_{1,2} \in C^\infty(\mathbb{R}^n), \quad (13)$$

the marginals of p_t remain unchanged. However, the general form of this tangent space, given by $\text{Tan}_{p_t,1}\Pi(\mu, \nu)$ in Eq. (8), does not provide us with a concise and convenient projection (this avenue was explored, to some degree, in [4]). Yet, if instead of $\text{Tan}_{p_t,1}\Pi(\mu, \nu)$, we focus on a sub-manifold, which still preserves the marginals, the problem can become tractable. In particular, the tangent space $\text{Tan}_{p_t,2}\Pi(\mu, \nu) = \ker(P_{p_t,X}) \times \ker(P_{p_t,Y})$ offers a straight-forward projection, while preserving the marginals. Noting that the Riemannian gradient is the orthogonal projection of the Euclidean one onto the tangent space [16], we get the projected version of v^{gd} , as shown in Fig. 1. In fact, projecting v^{gd} onto $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$ gives us the update rule

$$\begin{aligned} \dot{X}_t &= v_1^{\text{ocd}} := -\nabla_x c(X_t, Y_t) + \mathbb{E}_{p_t}[\nabla_x c(X_t, Y_t)|X_t] \\ \text{and } \dot{Y}_t &= v_2^{\text{ocd}} := -\nabla_y c(X_t, Y_t) + \mathbb{E}_{p_t}[\nabla_y c(X_t, Y_t)|Y_t] \end{aligned} \quad (14)$$

which we refer to as the Orthogonal Coupling Dynamics (OCD). The orthogonality comes from the observation that $\mathbb{E}[v_1^{\text{ocd}} \cdot X_t] = \mathbb{E}[v_2^{\text{ocd}} \cdot Y_t] = 0$. In general, there is no guarantee that OCD reaches global minimum of (2), as the projection onto the sub-manifold $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$ can prevent the dynamics from visiting all possible couplings between μ and ν . This can be readily noticed, e.g. if Y_0 is initialized as a sub-optimal map of X_0 . As a result, the dynamics remains frozen at a

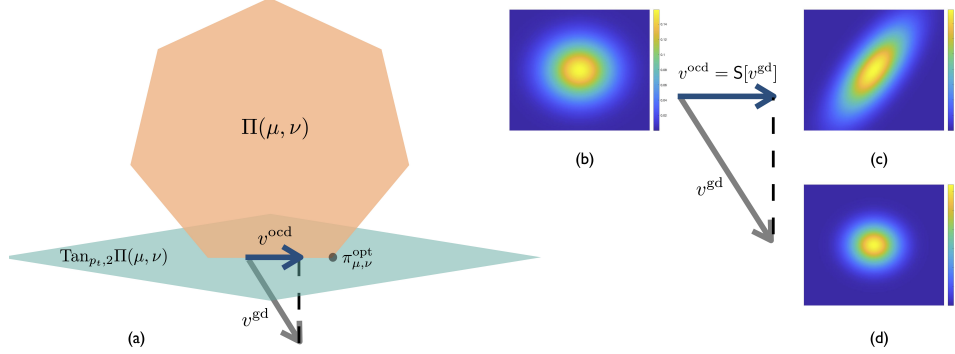


FIGURE 1. Geometric illustration of OCD; (a) The gradient descent, which leaves the polytope $\Pi(\mu, \nu)$, is projected onto the tangent space $\ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$; (b) Initial condition as product Gaussian probability density; (c) Evolved by v^{gcd} , the initial density is updated by projected gradient descent in L^2 -cost; (d) Evolved by v^{gcd} , the initial density is updated by gradient descent in L^2 -cost, leaving $\Pi(\mu, \nu)$.

sub-optimal fixed point. However, as we will discuss, there are several nuances in (14), which warrant investigation of OCD on both theoretical and numerical fronts. In particular, we observe that OCD ensures the cost is non-increasing. This result plays a crucial role in two important properties that we deduce: we observe that sub-optimal couplings turn out to be unstable (see Proposition 2.3), and that OCD gives the sharpest decay in L^2 -cost among processes with velocity in $\text{Tan}_{p_t, 2}\Pi(\mu, \nu)$ (see Theorem 2.6). In practice, the dynamics reaches accurate estimations of the optimization problem (2), as we demonstrate in results section.

From conceptual point of view, by replacing the optimization problem (2) with the OCD system (14), we turn the Monge-Kantorovich problem into a series of regression problems, implied by the conditional expectation. While the original problem, given by Eq. (2), is a constrained optimization over distributions, the OCD entails an unconstrained optimization over functions, in L^2 . This transformation allows us to leverage tools of conditional expectation estimation in order to achieve efficient computation of optimal transport in contrast to the original constrained infinite dimensional linear programming problem.

However, before analyzing OCD, let us mention that alternatively, one can consider the Langevin dynamics

$$\begin{aligned} dX_t &= (-\nabla_x c + \mathbb{E}_{p_t}[\nabla_x c | X_t] + \epsilon \nabla_x \log f_\mu(X_t)) dt + \sqrt{2\epsilon} dW_t \\ \text{and } dY_t &= (-\nabla_y c + \mathbb{E}_{p_t}[\nabla_y c | Y_t] + \epsilon \nabla_y \log f_\nu(Y_t)) dt + \sqrt{2\epsilon} dB_t, \end{aligned} \quad (15)$$

where $\epsilon > 0$ is the regularization parameter; W_t, B_t are independent Wiener processes with values in \mathbb{R}^n ; and $f_{\mu, \nu}$ are densities of the measures μ, ν . Assuming log-concave densities, Conforti et al [23] have shown that the measure generated by the Langevin dynamics (15) converges to the solution of (2), once the latter is regularized by entropy with parameter ϵ . Despite the theoretical grounding of SDE (15), we have three main motivations behind focusing on OCD (14), instead:

- (1) While the SDE system gives us an approximation of the optimal joint density, it would not give us a map between X_t and Y_t , due to the presence of the Brownian motion. Therefore, for applications where the random map is of interest, the SDE may not be employed; at least without further treatments.
- (2) The densities of the measures may not be known in many relevant applications. We are typically left only with samples of μ and ν . Therefore the evaluation of the SDE dynamics requires further regularizations.
- (3) From the numerical aspect, computing the grad-log terms of the density, even if the latter is given, can be significantly costly if n is large in \mathbb{R}^n .

The open question remains under which technical assumptions the case of $\epsilon = 0$, which is OCD (14), would converge to the solution of the Monge-Kantorovich problem. While we do not directly address this problem and despite this knowledge gap, we present new results on both theoretical and numerical aspects of OCD, presenting its basic features along with computational relevance for practical optimal transport problems.

2.2. General Structural Properties

The OCD dynamics (14) leads to a certain favourable theoretical features. These are rooted in the fact that the conditional expectation is a projection operator in L^2 with self-adjoint and contraction characteristics. We leverage these characteristics to derive theoretical properties of the OCD. We provide an overview of the main theoretical results below, where the proofs are given in section 3.

- (1) **Marginal Preservation.** Since the velocities $v_{1,2}^{\text{od}}$, introduced in Eq. (14), are orthogonal with respect to the functions of X_t and Y_t , respectively, a direct computation shows that the distributions of X_t and Y_t remain unchanged. More precisely we have the following result.

Proposition 2.1. *Consider (X_t, Y_t) to be the solution of the ODE system*

$$\dot{X}_t = v_1 \quad \text{and} \quad \dot{Y}_t = v_2 \tag{16}$$

with velocities $v = (v_1, v_2) \in \text{Tan}_{p_t, 2}\Pi(\mu, \nu)$, and the initial condition $(X_0, Y_0) \sim \pi$, where $\pi \in \Pi(\mu, \nu)$ and p_t is their joint law. Under the regularity assumptions stated in Assumption 3.1, $p_t \in \Pi(\mu, \nu)$.

- (2) **Descent in Cost.** As a result of the conditional expectation contraction property, the OCD leads to a monotone decay in the cost. More specifically, we have the following result.

Proposition 2.2. *Consider (X_t, Y_t) to be the solution of OCD (14) with the initial condition $(X_0, Y_0) \sim \pi$, from an arbitrary coupling $\pi \in \Pi(\mu, \nu)$. Under the regularity assumptions stated in Assumption 3.1, we have*

$$\frac{d}{dt} \mathbb{E}[c(X_t, Y_t)] \leq 0. \tag{17}$$

In other words, the OCD (14) monotonically reduces the cost $\mathbb{E}[c(X_t, Y_t)]$ until the latter reaches its stationary value. This property is crucial to show that the sub-optimal fixed points of the dynamics are unstable (as discussed in the following).

- (3) **Instability of Sub-Optimal Couplings.** The monotonic decay in the cost allows us to investigate stability of sub-optimal couplings. In particular, we have that the solution (X_t, Y_t) of OCD with a measure from the sub-optimal coupling set

$$\Pi^*(\mu, \nu) := \left\{ \pi \in \Pi(\mu, \nu) : \int_{\mathbb{R}^{2n}} c(x, y) \pi(dx, dy) > \int_{\mathbb{R}^{2n}} c(x, y) \pi^{\text{opt}}(dx, dy) \right\}$$

is unstable, in the sense described below.

Proposition 2.3. *Consider (X_t, Y_t) to be the solution of the ODE system (14) with the initial condition $(X_0, Y_0) \sim \pi$, from an arbitrary coupling $\pi \in \Pi(\mu, \nu)$. Suppose that the joint measure of (X_t, Y_t) fulfills*

$$p_T \in \Pi^*(\mu, \nu) \quad (18)$$

at some $T > 0$. Under the regularity assumptions stated in Assumption 3.1, the coupling p_T is unstable in the sense that if the law is perturbed by

$$p_T^\epsilon = (1 - \epsilon)p_T + \epsilon\pi_{\mu, \nu}^{\text{opt}}, \quad (19)$$

where $0 < \epsilon < 1$, the solution (X_t, Y_t) would not have the measure p_T for any time later.

- (4) **McKean-Vlasov System.** The OCD (14) can be equivalently interpreted as a probability flow in the sense of Eq. (6). Due to the absence of Brownian motion, this probability flow is in the form of the Vlasov equation. The latter arises in the mean field limit of interacting particles, e.g. see [34, 51]. By exploiting the projection operator S , the corresponding probability density evolution takes the concise form

$$\partial_t \rho_t + \nabla_x \cdot (\rho_t S_{p_t, x} [\nabla_x c]) + \nabla_y \cdot (\rho_t S_{p_t, y} [\nabla_y c]) = 0, \quad (20)$$

where

$$\begin{aligned} S_{p_t, x} [\nabla_x c] &= \nabla_x c(x, y) - \mathbb{E}_{(X_t, Y_t) \sim p_t} [\nabla_x c(X_t, Y_t) | X_t = x] \\ \text{and } S_{p_t, y} [\nabla_y c] &= \nabla_y c(x, y) - \mathbb{E}_{(X_t, Y_t) \sim p_t} [\nabla_y c(X_t, Y_t) | Y_t = y]. \end{aligned} \quad (21)$$

Theorem 2.4. *Consider (X_t, Y_t) to be the solution of the ODE system (14) with the initial condition $(X_0, Y_0) \sim \pi$, from an arbitrary coupling $\pi \in \Pi(\mu, \nu)$. Suppose p_t is the joint law of (X_t, Y_t) with the density ρ_t . Under the regularity assumptions stated in Assumption 3.1, ρ_t is weak solution of the Vlasov-type equation Eq. (20).*

We expect that the Vlasov equation (20) admits a variational formulation in terms of Otto calculus [37, 3, 59, 9], if appropriate set of couplings is considered (those that can be generated by $\text{Tan}_{p_t, 2} \Pi(\mu, \nu)$). However, we leave this interesting question to a separate work. Instead, we will present a variational formulation of OCD in the probability space for the L^2 Monge-Kantorovich cost, in the next section.

2.3. Key Features in L^2 Monge-Kantorovich

So far, we discussed that OCD (14) brings a given joint distribution closer to the solution of the Monge-Kantorovich problem. In other words, the dynamics decreases the cost $\mathbb{E}[c(X_t, Y_t)]$. In this

section, we present a finer result on the decay of the quadratic cost $c(x, y) = \|x - y\|_2^2$, resulted from OCD, i.e. for the ODE system

$$\dot{X}_t = Y_t - \mathbb{E}_{p_t}[Y_t|X_t] \quad \text{and} \quad \dot{Y}_t = X_t - \mathbb{E}_{p_t}[X_t|Y_t] . \quad (22)$$

In this scenario, the optimal transport problem is equivalent to maximizing correlation between X_t and Y_t , see [54, 56]. Focusing on (22), we show the following results.

- (1) **Symmetric Positive Semi-Definite Correlation.** Consider S_+^n to be the set of symmetric positive semi-definite matrices

$$S_+^n := \{A \in \mathbb{R}^{n \times n} | A = A^T, A \succeq 0\} . \quad (23)$$

Let us define the cross-correlation matrix

$$J_t := \mathbb{E}[X'_t \otimes Y'_t] \quad (24)$$

with $X'_t := X_t - \mathbb{E}[X_t]$ and $Y'_t := Y_t - \mathbb{E}[Y_t]$. We have the following result.

Proposition 2.5. *Consider (X_t, Y_t) to be the solution of the ODE system (14) with the initial condition $(X_0, Y_0) \sim \mu \otimes \nu$, given the regularity assumptions stated in Assumption 3.1. Therefore*

$$J_t \in S_+^n . \quad (25)$$

In other words the L^2 -OCD (22) ensures that the cross-correlation created between X_t and Y_t remains in the space of symmetric positive semi-definite matrices.

- (2) **Sharp Descent and Variational Formulation.** The L^2 -OCD (22) gives rise to a sharp decay in the cost $\mathbb{E}[c(X_t, Y_t)]$. More precisely, if we restrict the set of velocities to the tangent space $\ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$, the OCD gives fastest decay in the cost $\mathbb{E}[c(X_t, Y_t)]$.

Theorem 2.6. *Under the regularity assumptions stated in Assumption 3.1, consider (X_t, Y_t) to be the solution of the ODE system (16) and the initial condition $(X_0, Y_0) \in \pi$, where $\pi \in \Pi(\mu, \nu)$. Let*

$$\mathcal{C}(p_t) := \frac{1}{2} \int_{\mathbb{R}^{2n}} \|x - y\|_2^2 p_t(dx, dy) \quad (26)$$

be the L^2 transport cost and

$$\mathcal{L}_{p_t}(v) := \frac{d}{dt} \mathcal{C}(p_t) + \frac{1}{2} \int_{\mathbb{R}^{2n}} \|v\|_2^2 p_t(dx, dy) , \quad (27)$$

its decay rate penalized by the L^2 norm of the velocities. Define optimal velocities

$$v_{p_t}^{opt} := \arg \min_{v \in \ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})} \mathcal{L}_{p_t}(v) . \quad (28)$$

Therefore

$$v_{p_t, 1}^{opt}(X_t, Y_t) = Y_t - \mathbb{E}_{p_t}[Y_t|X_t] \quad (29)$$

$$\text{and } v_{p_t, 2}^{opt}(X_t, Y_t) = X_t - \mathbb{E}_{p_t}[X_t|Y_t] . \quad (30)$$

This result can be employed as variational formulation of L^2 -OCD. The velocities $v_{1,2}^{\text{ocd}}$ admit the variational formulation

$$v_{1,2}^{\text{ocd}} = \arg \min_{v \in \ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})} \left[\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}^{2n}} \|x - y\|_2^2 dp_t + \frac{1}{2} \int_{\mathbb{R}^{2n}} \|v\|_2^2 dp_t \right].$$

Remark 2.7. *In 1D, we expect that the variational formulation given above can be cast in the Jordan-Kinderlehrer-Otto (JKO) type flow [37]. We do not present a proof but motivate this by the following. Assume $n = 1$, $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ are absolutely continuous. Then one can define the JKO form of L^2 -OCD as*

$$p^{k+1} \in \arg \min_{p \in \Pi(\mu, \nu)} \left\{ \frac{1}{2\tau} W_2^2(p, p^k) + \mathcal{C}(p) \right\} \quad (31)$$

with $\mathcal{C}(p)$ given by (26). We expect the limit curve p_t of (31) as $\tau \downarrow 0$ becomes the gradient flow of $\mathcal{C}(p)$ on $\Pi(\mu, \nu)$, with the velocity field v_t given by L^2 -OCD. This can be justified since in 1D, any marginal preserving scheme should have a velocity field $v_t \in \ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$. However we leave the technical proof of this analogy to dedicated studies.

Remark 2.8. *Note that one can extend the steep descent formulation to a general continuously differentiable cost $c \in C^1(\mathbb{R}^n \times \mathbb{R}^n)$. While not pursued here in detail, we give a general argument. For the functional $\mathcal{C}(p) = \int c dp$, the first variation in a marginal-preserving direction $v = (v_1, v_2) \in \ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$ is*

$$D\mathcal{C}_p[v] = \int (\nabla_x c v_1 + \nabla_y c v_2) dp. \quad (32)$$

Consequently, the unique minimizer of the quadratic functional

$$v \mapsto D\mathcal{C}_p[v] + \frac{1}{2} \|v\|_{L^2(p)}^2 \quad \text{over } v = (v_1, v_2) \in \ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$$

is the orthogonal projection of $-(\nabla_x c, \nabla_y c)$ onto $\ker(\mathbf{P}_{p_t, X}) \times \ker(\mathbf{P}_{p_t, Y})$, namely

$$v_1^{\text{ocd}} = -\nabla_x c + \mathbb{E}[\nabla_x c \mid X], \quad v_2^{\text{ocd}} = -\nabla_y c + \mathbb{E}[\nabla_y c \mid Y].$$

3. TECHNICAL RESULTS

3.1. Review of Conditional Expectation Properties

In order to establish basic theoretical properties of the OCD (14), we often make use of the two following properties of the conditional expectation for random variables $A, B, C \in \mathcal{H}$ [22, 15].

- (1) *Contraction property:* As a result of Jensen's inequality, we have

$$\mathbb{E}[\|A\|_2^2] \geq \mathbb{E}[\|\mathbb{E}[A|B]\|_2^2]. \quad (33)$$

- (2) *Self-adjoint property:* Due to the law of total expectation, we have

$$\mathbb{E}[A \cdot \mathbb{E}[B|C]] = \mathbb{E}[B \cdot \mathbb{E}[A|C]]. \quad (34)$$

3.2. Assumptions

We do not attempt to prove well-posedness of (14) in this work, instead we collect below a set of regularity assumptions under which the subsequent derivations are justified. These are standard in the literature on McKean–Vlasov dynamics, see e.g. [63, 11]. In particular, in order to justify the analytical steps, we make explicit the following regularity assumptions.

Assumption 3.1 (Regularity of densities and velocities). *For each $t \geq 0$, the joint law p_t of (X_t, Y_t) admits a Lebesgue density $\rho_t \in L^1(\mathbb{R}^{2n})$ with finite second moment. Moreover:*

- (1) *The map $t \mapsto \rho_t$ is weakly continuous in L^1 .*
- (2) *The velocity field $v_t = (v_{1,t}, v_{2,t})$ satisfies $v_t \in L^2(p_t)$ and $v_t \rho_t \in L^1(\mathbb{R}^{2n})$.*
- (3) *The cost satisfies $c \in C^1(\mathbb{R}^{2n})$ and its gradients $\nabla_x c, \nabla_y c$ have at most linear growth*

$$|\nabla_x c(x, y)| + |\nabla_y c(x, y)| \leq C(1 + |x| + |y|).$$

- (4) *Conditional expectations such as $\mathbb{E}_{p_t}[\nabla_x c(X_t, Y_t) \mid X_t = x]$ exist in $L^2(\mu)$ and define measurable functions of (x, y) .*
- (5) *There exists $L > 0$ such that for all couplings $p, q \in \Pi(\mu, \nu)$,*

$$\|\mathbb{E}_p[\nabla_x c(X, Y) \mid X] - \mathbb{E}_q[\nabla_x c(X, Y) \mid X]\|_{L^2(\mu)} \leq L W_2(p, q),$$

and similarly for the conditional expectation with respect to Y .

Remark 3.2 (Stability of the orthogonal complement operator S_p). *For fixed $p \in \Pi(\mu, \nu)$, the maps*

$$S_{p,x} : L^2(p; \mathbb{R}^n) \rightarrow L^2(p; \mathbb{R}^n), \quad S_{p,y} : L^2(p; \mathbb{R}^n) \rightarrow L^2(p; \mathbb{R}^n),$$

are linear L^2 -contractions, since they are of the form $S_{p,x} = I - P_{p,X}$ and $S_{p,y} = I - P_{p,Y}$, where $P_{p,X}, P_{p,Y}$ are conditional-expectation operators. Moreover, Assumption 3.1(5) implies a stability of S_p with respect to the law p in the directions $f = \nabla_x c$ and $f = \nabla_y c$. Indeed, for any $p, q \in \Pi(\mu, \nu)$,

$$\|S_{p,x}[\nabla_x c] - S_{q,x}[\nabla_x c]\|_{L^2(\mu)} = \|\mathbb{E}_q[\nabla_x c \mid X] - \mathbb{E}_p[\nabla_x c \mid X]\|_{L^2(\mu)} \leq L W_2(p, q),$$

and similarly for $S_{p,y}[\nabla_y c]$.

Remark 3.3 (On the tangent space $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$). *Note that*

$$\text{Tan}_{p_t,2}\Pi(\mu, \nu) = \ker(P_{p_t,X}[v_1]) \times \ker(P_{p_t,Y}[v_2]),$$

where by definition

$$P_{p,X}[v_1] = \mathbb{E}_p[v_1(X, Y) \mid X], \quad P_{p,Y}[v_2] = \mathbb{E}_p[v_2(X, Y) \mid Y].$$

Since conditional expectation is a bounded linear operator, their kernels are closed linear subspaces of $L^2(p; \mathbb{R}^n)$. Thus $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$ is a closed linear subspace of $L^2(p; \mathbb{R}^{2n})$.

3.3. Proofs

Proof of Proposition 2.1 (Marginal preservation). By definition of the tangent space $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$ we have

$$\mathbb{E}_{p_t}[v_1(X_t, Y_t) \mid X_t] = 0, \quad \mathbb{E}_{p_t}[v_2(X_t, Y_t) \mid Y_t] = 0. \quad (35)$$

Let $g \in C_c^\infty(\mathbb{R}^n)$ be a smooth, compactly supported test function. Using Assumption 3.1(1)–(3), the chain rule for the ODE (14) yields

$$\frac{d}{dt}\mathbb{E}[g(X_t)] = \mathbb{E}[\nabla g(X_t) \cdot v_{1,t}(X_t, Y_t)],$$

where the exchange of derivative and expectation is justified because ∇g is bounded and $v_t \in L^2(p_t)$. Using the self-adjointness of conditional expectation (see property 2) we obtain

$$\mathbb{E}[\nabla g(X_t) \cdot v_{1,t}(X_t, Y_t)] = \mathbb{E}[\nabla g(X_t) \cdot \mathbb{E}[v_{1,t}(X_t, Y_t) \mid X_t]].$$

By (35) the right-hand side vanishes, hence $\frac{d}{dt}\mathbb{E}[g(X_t)] = 0$ for all $g \in C_c^\infty(\mathbb{R}^n)$. This implies that the marginal of X_t is constant in time, i.e. $X_t \sim \mu$. The argument for Y_t is identical. \square

In particular, along the OCD dynamics the marginals remain fixed: if $(X_0, Y_0) \sim \pi \in \Pi(\mu, \nu)$, then $X_t \sim \mu$ and $Y_t \sim \nu$ for all $t \geq 0$.

Proof of Proposition 2.2 (Descent in Cost). By definition of the cost functional,

$$\mathcal{C}(t) := \mathbb{E}[c(X_t, Y_t)] = \int_{\mathbb{R}^{2n}} c(x, y) p_t(dx, dy).$$

Using that $c \in C^1$ with at most linear growth of its gradients and that $v_t \in L^2(p_t)$ (Assumption 3.1), we may differentiate under the expectation to obtain

$$\frac{d}{dt}\mathcal{C}(t) = \mathbb{E}[\nabla_x c(X_t, Y_t) \cdot v_{1,t}(X_t, Y_t) + \nabla_y c(X_t, Y_t) \cdot v_{2,t}(X_t, Y_t)]. \quad (36)$$

For OCD, the velocity field is

$$v_{1,t} = \nabla_y c - \mathbb{E}[\nabla_y c \mid X_t], \quad v_{2,t} = \nabla_x c - \mathbb{E}[\nabla_x c \mid Y_t].$$

Substituting these into (36) and rearranging gives

$$\frac{d}{dt}\mathcal{C}(t) = -\mathbb{E}[(\nabla_x c - \mathbb{E}[\nabla_x c \mid Y_t]) \cdot \nabla_x c + (\nabla_y c - \mathbb{E}[\nabla_y c \mid X_t]) \cdot \nabla_y c].$$

Using the self-adjointness of conditional expectation in $L^2(p_t)$ (see property 2), we obtain

$$\mathbb{E}[\nabla_x c \cdot \mathbb{E}[\nabla_x c \mid Y_t]] = \mathbb{E}[\mathbb{E}[\nabla_x c \mid Y_t] \cdot \mathbb{E}[\nabla_x c \mid Y_t]],$$

and similarly for the y -component. Hence

$$\frac{d}{dt}\mathcal{C}(t) = -\left(\|\nabla_x c - \mathbb{E}[\nabla_x c \mid Y_t]\|_{L^2(p_t)}^2 + \|\nabla_y c - \mathbb{E}[\nabla_y c \mid X_t]\|_{L^2(p_t)}^2\right). \quad (37)$$

Since conditional expectation is an L^2 -contraction (see property 1), each term in (37) is nonnegative, so the right-hand side is nonpositive. Therefore

$$\frac{d}{dt}\mathcal{C}(t) \leq 0,$$

with equality if and only if the two OCD velocity components vanish p_t -almost surely. \square

Proof of Proposition 2.3 (Instability of Sub-Optimal Couplings). First note that marginals of p_T^ϵ are μ and ν (due to Proposition 2.1). Next, following the definition of $\pi_{\mu,\nu}^{\text{opt}}$ (see Eq. (2)), we have

$$\int_{\mathbb{R}^{2n}} c(x, y) p_T^\epsilon(dx, dy) < \int_{\mathbb{R}^{2n}} c(x, y) p_T(dx, dy) . \quad (38)$$

However since

$$\frac{d}{dt} \mathbb{E}[c(X, Y)] \leq 0 \quad (39)$$

due to Proposition 2.2, we have

$$\int_{\mathbb{R}^{2n}} c(x, y) p_t(dx, dy) \leq \int_{\mathbb{R}^{2n}} c(x, y) p_T^\epsilon(dx, dy) < \int_{\mathbb{R}^{2n}} c(x, y) p_T(dx, dy) \quad (40)$$

for all $t > T$. Thus given $t > T$, $\exists A \in \mathcal{B}(\mathbb{R}^{2n})$ for which $p_t(A) \neq p_T(A)$. \square

Proof of Theorem 2.4 (McKean–Vlasov system). Let $g \in C_c^\infty(\mathbb{R}^{2n})$ be a smooth, compactly supported test function. Since $c \in C^1$ with at most linear growth of its gradients, and $v_t \in L^2(p_t)$ (Assumption 3.1), the chain rule for the ODE (14) and dominated convergence yield

$$\frac{d}{dt} \mathbb{E}[g(X_t, Y_t)] = \mathbb{E}[\nabla_x g(X_t, Y_t) \cdot v_{1,t}(X_t, Y_t) + \nabla_y g(X_t, Y_t) \cdot v_{2,t}(X_t, Y_t)] . \quad (3.22)$$

Since $\mathbb{E}[g(X_t, Y_t)] = \int g(x, y) \rho_t(x, y) dx dy$, the left-hand side of (3.22) equals

$$\frac{d}{dt} \int_{\mathbb{R}^{2n}} g(x, y) \rho_t(x, y) dx dy .$$

We now rewrite the right-hand side of (3.22). Recall that

$$v_{1,t}(x, y) = \nabla_x c(x, y) - \mathbb{E}_{p_t}[\nabla_x c(X_t, Y_t) \mid X_t = x] = S_{p_t, x}[\nabla_x c](x, y),$$

and analogously $v_{2,t} = S_{p_t, y}[\nabla_y c]$. Since g has compact support and $v_t \rho_t \in L^1(\mathbb{R}^{2n})$ (Assumption 3.1(2)), we may apply integration by parts in the sense of distributions to obtain

$$\int_{\mathbb{R}^{2n}} \nabla_x g \cdot v_{1,t} \rho_t dx dy = - \int_{\mathbb{R}^{2n}} g(x, y) \nabla_x \cdot (\rho_t(x, y) v_{1,t}(x, y)) dx dy,$$

and likewise for the y -term

$$\int_{\mathbb{R}^{2n}} \nabla_y g \cdot v_{2,t} \rho_t dx dy = - \int_{\mathbb{R}^{2n}} g(x, y) \nabla_y \cdot (\rho_t(x, y) v_{2,t}(x, y)) dx dy.$$

Substituting these identities into (3.22), we obtain

$$\frac{d}{dt} \int_{\mathbb{R}^{2n}} g \rho_t dx dy = - \int_{\mathbb{R}^{2n}} g \nabla_x \cdot (\rho_t v_{1,t}) dx dy - \int_{\mathbb{R}^{2n}} g \nabla_y \cdot (\rho_t v_{2,t}) dx dy.$$

Since g is an arbitrary test function in $C_c^\infty(\mathbb{R}^{2n})$, we have the weak formulation

$$\partial_t \rho_t + \nabla_x \cdot (\rho_t v_{1,t}) + \nabla_y \cdot (\rho_t v_{2,t}) = 0,$$

which is the Vlasov equation (20). All divergences above are understood in the distributional sense, i.e. no pointwise derivatives of ρ_t are required. \square

Lemma 3.4 (Preparatory for Proposition 2.5). *Consider $A, B \in \mathcal{H}$ and let*

$$R := \mathbb{E}[A \otimes A - A \otimes \mathbb{E}[A|B]] . \quad (41)$$

Therefore

$$R \in S_+^n . \quad (42)$$

Proof. First note that R is symmetric due to the self-adjoint property of the conditional expectation (see property 2). Choose an arbitrary $v \in \mathbb{R}^n$ and let $L := \sum_{i=1}^n v_i A_i$. We have

$$\begin{aligned} \sum_{i,j=1}^n v_i \mathbb{E}[A_i \mathbb{E}[A_j|B]] v_j &= \mathbb{E}[L \mathbb{E}[L|B]] \leq \frac{1}{2} (\mathbb{E}[\|L\|_2^2] + \mathbb{E}[\|\mathbb{E}[L|B]\|_2^2]) \quad (\text{Cauchy-Schwartz}) \\ &\leq \mathbb{E}[\|L\|_2^2] \quad (\text{contraction property 1}) \\ &= \sum_{i,j=1}^n v_i \mathbb{E}[A_i A_j] v_j . \end{aligned}$$

Therefore $\sum_{i,j} v_i R_{ij} v_j \geq 0$, and thus $R \in S_+^n$. \square

Proof of Proposition 2.5 (Symmetric positive semidefinite correlation). Recall that $J_t = \mathbb{E}[X_t \otimes Y_t]$. Differentiating in time and using the L^2 -OCD dynamics

$$\dot{X}_t = Y_t - \mathbb{E}[Y_t | X_t], \quad \dot{Y}_t = X_t - \mathbb{E}[X_t | Y_t],$$

together with Assumption 3.1 and the chain rule, we obtain

$$\frac{\partial}{\partial t} J_t = \mathbb{E}[\dot{X}_t \otimes Y_t] + \mathbb{E}[X_t \otimes \dot{Y}_t].$$

Substituting the expressions for \dot{X}_t and \dot{Y}_t yields

$$\begin{aligned} \frac{\partial}{\partial t} J_t &= \mathbb{E}[(Y_t - \mathbb{E}[Y_t | X_t]) \otimes Y_t] + \mathbb{E}[X_t \otimes (X_t - \mathbb{E}[X_t | Y_t])] \\ &= \mathbb{E}[Y_t \otimes Y_t - Y_t \otimes \mathbb{E}[Y_t | X_t]] + \mathbb{E}[X_t \otimes X_t - X_t \otimes \mathbb{E}[X_t | Y_t]] \\ &=: R_t^J . \end{aligned} \quad (43)$$

We claim that $R_t^J \in S_+^n$ for all $t \geq 0$. Indeed, the first term in (43) has the form

$$\mathbb{E}[A \otimes A - A \otimes \mathbb{E}[A | B]] \quad \text{with } A = Y_t, \ B = X_t,$$

and is therefore symmetric positive semidefinite by Lemma 3.4. The same lemma applies to the second term in (43), with $A = X_t$ and $B = Y_t$, so both contributions lie in S_+^n and hence $R_t^J \in S_+^n$. The matrix J_t can be written as

$$J_t = J_0 + \int_0^t R_s^J ds.$$

By assumption $J_0 \in S_+^n$, and the integral of symmetric positive semidefinite matrices is again symmetric positive semidefinite. Therefore $J_t \in S_+^n$ for all $t \geq 0$. \square

Lemma 3.5 (Orthogonal projections onto marginal-preserving subspaces). *Let $(X, Y) \sim \pi \in \Pi(\mu, \nu)$ and consider the Hilbert space $L^2(\pi; \mathbb{R}^n)$ with inner product $\langle f, g \rangle = \mathbb{E}_\pi[f \cdot g]$. Define the closed linear subspaces*

$$\mathcal{H}_X^\perp := \{ \phi \in L^2(\pi; \mathbb{R}^n) : \mathbb{E}_\pi[\phi(X, Y) | X] = 0 \},$$

$$\mathcal{H}_Y^\perp := \{ \psi \in L^2(\pi; \mathbb{R}^n) : \mathbb{E}_\pi[\psi(X, Y) \mid Y] = 0 \}.$$

Both are closed in $L^2(\pi)$ since conditional expectation is a bounded linear operator. Consider the minimization problems

$$\tilde{\phi} := \arg \min_{\phi \in \mathcal{H}_X^\perp} \mathbb{E}_\pi [\|Y - \phi(X, Y)\|_2^2],$$

$$\tilde{\psi} := \arg \min_{\psi \in \mathcal{H}_Y^\perp} \mathbb{E}_\pi [\|X - \psi(X, Y)\|_2^2].$$

Then the minimizers exist, are unique, and satisfy

$$\tilde{\phi}(X, Y) = Y - \mathbb{E}_\pi[Y \mid X], \quad \tilde{\psi}(X, Y) = X - \mathbb{E}_\pi[X \mid Y].$$

Proof. The functional $\phi \mapsto \mathbb{E}\|Y - \phi\|^2$ is strictly convex, continuous, and coercive on $L^2(\pi)$, and \mathcal{H}_X^\perp is a closed linear subspace. Hence, there exists a unique minimizer, namely the orthogonal projection of Y onto \mathcal{H}_X^\perp . Let P denote this projection. Since \mathcal{H}_X^\perp is the kernel of the bounded operator $\phi \mapsto \mathbb{E}[\phi \mid X]$, we have

$$P_{p,X}Y = Y - \mathbb{E}_p[Y \mid X].$$

The same argument applies for ψ , yielding $\tilde{\psi} = X - \mathbb{E}[X \mid Y]$. \square

In the next proof we show that the OCD (14) gives rise to sharpest decay in quadratic cost among processes evolving in $\text{Tan}_{p_t,2}\Pi(\mu, \nu)$.

Proof of Theorem 2.6 (Sharp descent and variational formulation). For any admissible velocity field $v = (v_1, v_2) \in \ker(P_{p_t,X}) \times \ker(P_{p_t,Y})$, the instantaneous rate of change of the quadratic cost is

$$\frac{d}{dt}\mathcal{C}(p_t) = \mathbb{E}_{p_t}[Y \cdot v_1 + X \cdot v_2].$$

A direct computation shows that

$$\mathcal{L}_{p_t}(v) = \frac{1}{2} (\mathbb{E}_{p_t} [\|v_1 - Y\|^2 + \|v_2 - X\|^2] - \mathbb{E}_{p_t} [\|Y\|^2 + \|X\|^2]),$$

and the second term does not depend on v . Therefore minimizing $\mathcal{L}_{p_t}(v)$ over the constraint $\ker(P_{p_t,X}) \times \ker(P_{p_t,Y})$ is equivalent to minimizing

$$v \mapsto \mathbb{E}_{p_t} [\|v_1 - Y\|^2 + \|v_2 - X\|^2] \quad \text{over } \ker(P_{p_t,X}) \times \ker(P_{p_t,Y}).$$

Since $\ker(P_{p_t,X}) \times \ker(P_{p_t,Y})$ is a closed linear subspace of $L^2(p_t)$ (see Remark 3.3), and \mathcal{L}_{p_t} is strictly convex, continuous, and coercive, there exists a unique minimizer. By Lemma 3.5, the unique minimizer is

$$v_{1,t}^{\text{opt}} = Y - \mathbb{E}[Y \mid X], \quad v_{2,t}^{\text{opt}} = X - \mathbb{E}[X \mid Y],$$

which is precisely the OCD velocity field. \square

Remark 3.6. Note there is a similarity between the sharp decay property of OCD and Martingale series approximation employed in [55] (see Lemma 9 and Theorem 10 in [55]).

4. CHARACTERIZATION IN GAUSSIAN SETTING

We analyze the OCD dynamics for quadratic cost with Gaussian marginals and show when it coincides with OT. Let μ and ν be centered Gaussian measures with covariance matrices Σ_μ and Σ_ν . The dynamics of correlation matrix $J_t = \mathbb{E}[X_t \otimes Y_t]$, following L^2 -OCD (14), reads

$$\partial_t J_t = \Sigma_\mu + \Sigma_\nu - J_t^T \Sigma_\mu^{-1} J_t - J_t \Sigma_\nu^{-1} J_t^T, \quad (44)$$

due to the identities

$$\mathbb{E}_{p_t}[Y_t|X_t] = J_t^T \Sigma_\mu^{-1} X_t \quad \text{and} \quad \mathbb{E}_{p_t}[X_t|Y_t] = J_t \Sigma_\nu^{-1} Y_t, \quad (45)$$

which results from p_t being Gaussian, see [23]. The matrix-Riccati equation (44) admits many stationary solutions, since any bijection between μ and ν is a stationary solution of (14), as discussed in [23]. However, as we justified in Proposition 2.2, starting with independent samples, i.e. $\mathbb{E}[X_0 Y_0] = 0$, J_t remains symmetric positive semi-definite and its trace increases monotonically. This restricts the stationary state and results in the OT solution if Σ_μ and Σ_ν commute.

Proposition 4.1 (OCD recovers OT for commuting Gaussian covariances). *Let*

$$\mu = \mathcal{N}(0, \Sigma_\mu), \quad \nu = \mathcal{N}(0, \Sigma_\nu),$$

with $\Sigma_\mu, \Sigma_\nu \succ 0$, and let (X_t, Y_t) evolve according to the L^2 -OCD dynamics. If the covariances commute, i.e.

$$[\Sigma_\mu, \Sigma_\nu] = 0,$$

then L^2 -OCD converges to the optimal transport correlation, namely

$$J_t \longrightarrow J_\infty = \Sigma_\mu^{1/2} (\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \Sigma_\mu^{1/2} = J_{\text{OT}}.$$

Proof. When $[\Sigma_\mu, \Sigma_\nu] = 0$, the two matrices are simultaneously diagonalizable:

$$\Sigma_\mu = Q \text{diag}(\sigma_{\mu,i}) Q^\top, \quad \Sigma_\nu = Q \text{diag}(\sigma_{\nu,i}) Q^\top,$$

for some orthogonal Q . Under this change of coordinates, (44) decouples component-wise

$$\dot{j}_i = \sigma_{\mu,i} + \sigma_{\nu,i} - (\sigma_{\mu,i}^{-1} + \sigma_{\nu,i}^{-1}) j_i^2, \quad j_i(0) = 0.$$

By direct computation we find that the equilibrium (given $j_i(0) = 0$) is

$$j_i^* = \sqrt{\sigma_{\mu,i} \sigma_{\nu,i}}.$$

Hence $J_t \rightarrow J_\infty = Q \text{diag}(\sqrt{\sigma_{\mu,i} \sigma_{\nu,i}}) Q^\top$. On the other hand, the Gaussian OT cross-correlation is

$$J_{\text{OT}} = \Sigma_\mu^{1/2} (\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \Sigma_\mu^{1/2},$$

see e.g. [30]. When Σ_μ and Σ_ν commute, we get

$$J_{\text{OT}} = Q \text{diag}(\sqrt{\sigma_{\mu,i} \sigma_{\nu,i}}) Q^\top = J_\infty.$$

□

Remark 4.2. *When the covariances commute (i.e., the Gaussian marginals share eigenvectors), OCD reduces to n independent scalar Riccati flows, each converging to the correct cross-variance $\sqrt{\sigma_{\mu,i} \sigma_{\nu,i}}$. If the covariances do not commute, the symmetry of J_t is preserved, preventing convergence to the generally non-symmetric OT solution.*

Finally we address the speed of convergence to the OT solution in the case of commuting covariance matrices.

Theorem 4.3 (Exponential convergence of OCD to OT in the commuting Gaussian case). *Let $\mu = \mathcal{N}(0, \Sigma_\mu)$ and $\nu = \mathcal{N}(0, \Sigma_\nu)$ with $\Sigma_\mu, \Sigma_\nu \succ 0$ and $[\Sigma_\mu, \Sigma_\nu] = 0$. Consider the L^2 -OCD dynamics initialized with independent samples, and let $J_t = \mathbb{E}[X_t Y_t^\top]$ denote the cross-correlation. Let*

$$J_{OT} = \Sigma_\mu^{1/2} (\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \Sigma_\mu^{1/2}$$

be the Gaussian OT cross-correlation. Then $J_t \rightarrow J_{OT}$ exponentially fast: there exist constants $C > 0$ and $\kappa > 0$ such that

$$\|J_t - J_{OT}\|_2 \leq C e^{-2\kappa t} \quad \text{for all } t \geq 0,$$

where $\|\cdot\|_2$ denotes the spectral norm. Expressing $\Sigma_\mu = Q \text{diag}(\sigma_{\mu,i}) Q^\top$ and $\Sigma_\nu = Q \text{diag}(\sigma_{\nu,i}) Q^\top$ in a common orthonormal basis,

$$\kappa = \min_{1 \leq i \leq n} \left(\sqrt{\frac{\sigma_{\nu,i}}{\sigma_{\mu,i}}} + \sqrt{\frac{\sigma_{\mu,i}}{\sigma_{\nu,i}}} \right) \geq 2.$$

Proof. Since $[\Sigma_\mu, \Sigma_\nu] = 0$, there exists orthogonal Q with $\Sigma_\mu = Q \text{diag}(\sigma_{\mu,i}) Q^\top$ and $\Sigma_\nu = Q \text{diag}(\sigma_{\nu,i}) Q^\top$. The Riccati ODE

$$\dot{J}_t = \Sigma_\mu + \Sigma_\nu - J_t^\top \Sigma_\mu^{-1} J_t - J_t \Sigma_\nu^{-1} J_t^\top$$

preserves symmetry, so writing $J_t = Q \text{diag}(j_i(t)) Q^\top$ gives n decoupled scalar ODEs

$$\dot{j}_i(t) = \sigma_{\mu,i} + \sigma_{\nu,i} - (\sigma_{\mu,i}^{-1} + \sigma_{\nu,i}^{-1}) j_i(t)^2 = \alpha_i ((j_i^*)^2 - j_i(t)^2),$$

where $\alpha_i := \sigma_{\mu,i}^{-1} + \sigma_{\nu,i}^{-1}$ and $j_i^* := \sqrt{\sigma_{\mu,i} \sigma_{\nu,i}}$. The explicit solution is

$$j_i(t) = j_i^* \tanh(\alpha_i j_i^* t).$$

Hence

$$|j_i(t) - j_i^*| = j_i^* \left| \tanh(\alpha_i j_i^* t) - 1 \right| = \frac{2j_i^* e^{-2(\alpha_i j_i^* t)}}{1 + e^{-2(\alpha_i j_i^* t)}} \leq 2j_i^* e^{-2\rho_i t},$$

with $\rho_i := \alpha_i j_i^* = \sqrt{\sigma_{\nu,i}/\sigma_{\mu,i}} + \sqrt{\sigma_{\mu,i}/\sigma_{\nu,i}}$. Therefore, for the spectral norm we have

$$\|J_t - J_{OT}\|_2 = \|Q \text{diag}(j_i(t) - j_i^*) Q^\top\|_2 \leq \max_i |j_i(t) - j_i^*| \leq 2 \max_i j_i^* e^{-2(\min_i \rho_i) t}.$$

Setting $C := 2 \max_i j_i^*$ and $\kappa := \min_i \rho_i$ gives the claim. Finally, $\rho_i \geq 2$ (by arithmetic-geometric mean inequality), hence $\kappa \geq 2$. \square

5. NUMERICAL ALGORITHM THROUGH OPINION DYNAMICS

Opinion dynamics at its core tries to answer how different agents shape their opinions in contact with each-other [39]. While seemingly unrelated to the Monge-Kantorovich problem, in this section, we leverage analogy between opinion dynamics and OCD to provide a novel, yet simple, numerical algorithm for the latter. The main idea is to first, split the data points into a certain number of clusters. Next, the conditional expectation is estimated in each cluster using an appropriate hypothesis class.

5.1. Analogy with Opinion Dynamics

Consider a population of agents indexed by $i \in \{1, \dots, N\}$. Suppose each person i has certain two-dimensional position $P_i(t) \in \mathbb{R}^2$ as their opinion in time t . In the standard form of the bounded interval Hegselmann-Krause consensus dynamics, e.g., see [36, 46], these positions relax towards a global consensus via

$$\dot{P}_i = \sum_{i \neq j} a_{ij} (P_i(t) - P_j(t)) , \quad (46)$$

where $a_{ij} \geq 0$ captures the influence of other agents positions. We can decompose it as $a_{ij} = \phi(r_{ij})\alpha_{ij}$ where $r_{ij} = \|P_i - P_j\|_2$ is the Euclidean distance between the positions of two persons i and j , and α a constant matrix. Based on α , we can intuitively consider two idealized regimes:

- (1) The positions are completely independent from each-other and α becomes the identity matrix.
- (2) The positions are perfectly tied to each-other and α becomes the exchange matrix J (with ones on its anti-diagonal and zeros elsewhere).

The first one degenerates the system into independent positions. More interestingly, let us focus on the latter and denote $P = [P_X, P_Y]^T$, therefore the dynamics takes the form

$$\dot{P}_{X,i} = \sum_{i \neq j} \phi(r_{ij}) (P_{Y,i}(t) - P_{Y,j}(t)) \quad \text{and} \quad \dot{P}_{Y,i} = \sum_{i \neq j} \phi(r_{ij}) (P_{X,i}(t) - P_{X,j}(t)).$$

If we assume $\phi(\cdot)$ is simply a step function i.e. it is one if $r \leq \epsilon$ and zero otherwise (for $\epsilon > 0$), we can further simplify the dynamics to

$$\dot{P}_{X,i} = \frac{1}{N_{\bar{I}^{X_i, \epsilon}}} \sum_{j \in \bar{I}^{X_i, \epsilon}} (P_{Y,i}(t) - P_{Y,j}(t)) \quad \text{and} \quad \dot{P}_{Y,i} = \frac{1}{N_{\bar{I}^{Y_i, \epsilon}}} \sum_{j \in \bar{I}^{Y_i, \epsilon}} (P_{X,i}(t) - P_{X,j}(t)) ,$$

where $\bar{I}^{X_i, \epsilon}$ and $\bar{I}^{Y_i, \epsilon}$ are the set of all agents with $\|P_{X,i} - P_X\|_2^2 \leq \epsilon^2$ and $\|P_{Y,i} - P_Y\|_2^2 \leq \epsilon^2$, respectively, with the corresponding number of elements $N_{\bar{I}^{X_i, \epsilon}}$ and $N_{\bar{I}^{Y_i, \epsilon}}$. In the limit of $N \rightarrow \infty$ and $\epsilon \rightarrow 0$, it is suggestive to express the above dynamics in the statistical form. Treating $P_{X,i}(t)$ and $P_{Y,i}(t)$ as samples of random variables $X_t \in \mathcal{H}$ and $Y_t \in \mathcal{H}$, respectively, the dynamics takes the form

$$\dot{X}_t = Y_t - \mathbb{E}[Y_t|X_t] \quad \text{and} \quad \dot{Y}_t = X_t - \mathbb{E}[X_t|Y_t] , \quad (47)$$

which is OCD (14) for L^2 -cost. From technical point of view, the discrete version is justified due to the approximation of conditional expectation by a discrete kernel [48]. Through the lens of opinion dynamics, the OCD describes a dynamics which is local in a sense that each agent is only influenced by agents in its vicinity, yet as discussed before, it is closely linked to the global optimization problem (2).

5.2. Non-parametric Monte-Carlo Algorithm

The analogy with opinion dynamics hints a numerical recipe for OCD. Conceptually, the parameter ϵ gives rise to formation of clusters in the cloud of (X_t, Y_t) points. We consider two numerical schemes:

- (1) **OCD-piecewise constant.** In each cluster, X_t and Y_t relax towards their locally averaged values, i.e. the conditional expectation is estimated to be constant in the cluster.

- (2) **OCD-piecewise linear.** In each cluster, the conditional expectation is estimated using linear regression, i.e. the joint distribution of each cluster is estimated by a Gaussian.

Starting with N_p , i.i.d., samples $\hat{X}_{i,t_0} \sim \mu$ and $\hat{Y}_{i,t_0} \sim \nu$, we update the state of each sample according to discretized form of OCD. It is important to note that here we are focusing on an approximate solution of the optimization problem (2) over atom-less distributions μ and ν . An alternative interpretation, where the distributions are assumed to be sum of diracs (assignment problem) is not pursued here (see [65]). Therefore throughout what follows, we assume that there exists an underlying continuous map between the two measure spaces. Three hyper-parameters control the discretization error: number of samples N_p , time step size Δt , and cluster cut-off ϵ . Once these values are fixed, relying on Euler's scheme (see Runge-Kutta's scheme in Supplementary Information), each sample is updated according to

$$\hat{X}_i^{n+1} = \hat{X}_i^n + \nabla_x c(\hat{X}_i^n, \hat{Y}_i^n) \Delta t - \hat{\mathcal{K}}_X^n \Delta t \quad (48)$$

$$\text{and } \hat{Y}_i^{n+1} = \hat{Y}_i^n + \nabla_y c(\hat{X}_i^n, \hat{Y}_i^n) \Delta t - \hat{\mathcal{K}}_Y^n \Delta t, \quad (49)$$

where superscript n and $n + 1$ denote the approximations at t^n and $t^n + \Delta t$, respectively. The functionals $\hat{\mathcal{K}}_{X,Y}^n(\cdot)$ estimate conditional expectations with respect to X and Y , respectively, and they are found depending on the following estimators.

- (1) **OCD-piecewise constant.** The estimator is constant in each cluster (equivalent to using square kernel in the estimation of the conditional expectation). The expressions for $\hat{\mathcal{K}}_{X,Y}^n(\cdot)$ are given in Supplementary Information.
- (2) **OCD-piecewise linear.** The estimator follows linear regression (accounts for using Gaussian kernel in the estimation of the conditional expectation). In other words here we look for L^2 projection of $\nabla_{x,y} c$ onto the span of linear functions [38]. The final expressions are given in Supplementary Information.

The overview of the algorithm is given in Supplementary Information. The convergence of the algorithm, in general, is controlled by $N_p, \Delta t, \epsilon$. Assuming that the OCD initialized by samples of μ, ν is well-posed, it is reasonable to consider fixed-order convergence with respect to Δt (e.g. first-order for Euler time integration, see [52]). Also the error due to finite N_p is expected to behave similar to conventional Monte-Carlo schemes [41]. However, the error dependency on the cut-off value ϵ is more delicate. Two asymptotic limits of $\epsilon \rightarrow 0$ and $\epsilon \rightarrow \infty$ can be intuitively understood. The limit $\epsilon \rightarrow 0$ accounts for no interaction between agents, i.e. each point forms its own cluster. Therefore the points remain frozen in time. On the other hand, $\epsilon \rightarrow \infty$ gives rise to perfect interaction among all agents, forming one giant cluster. In this case, the dynamics converges to the average value among all sample points, for the 1st order approximation, and towards linear regression, for the 2nd order case. Obviously, the optimal value of ϵ needs to be tuned in-between, and is expected to depend on N_p . The transition between frozen dynamics in $\epsilon \rightarrow 0$ and homogenization $\epsilon \rightarrow \infty$ suggests a phase transition type behaviour, resulting from ϵ . The transition point can serve as a proxy for the optimal cut-off value, as discussed in Supplementary Information.

Once the pairing between \hat{X}_i and \hat{Y}_i reaches its stationary state (according to a stopping criterion), the sampled points $(\hat{X}_i^s, \hat{Y}_i^s)_{i \in \{1, \dots, N_p\}}$ can be employed to estimate the Monge map $\hat{T}()$,

using a suitable hypothesis class. We train a Neural Network (NN) with a few layers and $\tanh(\cdot)$ as the activation function to approximate the map and denote it by $M_{X \rightarrow Y}$. As the loss function, we simply consider the expectation of L^2 point-wise error between NN's output $M_{X \rightarrow Y}(\hat{X}^*)$ and \hat{Y}^* , i.e. $\text{Loss} = \sum_{i=1}^{N_p} \|M_{X \rightarrow Y}(\hat{X}_i^*) - \hat{Y}_i^*\|_2^2 / N_p$.

5.3. Computational Complexity

In the following, we provide complexity estimates of OCD. Note that the discussion is mainly based on heuristic arguments and rely on practical scaling observations rather than formal asymptotic analysis. For a given time step (iteration), the main computational task of the proposed OCD algorithm is the evaluation of the conditional expectation, whose estimation relies on forming clusters by finding neighbours of each point in a given distribution. For the latter, though non-exclusively, the Ball Tree method [49] scales efficiently i.e. $\mathcal{O}(nN_p \log N_p)$ with number of particles N_p and space dimension n . Therefore, we expect that the OCD equipped with piecewise constant estimator for the conditional expectation to have the time complexity of $\mathcal{O}(nN_p \log N_p)$, since the number of ODEs is $2n$ and number of operations in order to compute the distance between each two particles is $\mathcal{O}(n)$. However, since we deployed the direct solver in the computation of the conditional expectation with piecewise linear estimator, the corresponding OCD approach scales poorly with space dimension n leading to the overall time complexity of $\mathcal{O}(n^3 N_p \log N_p)$.

In terms of memory consumption, since OCD, unlike linear programming, does not require computation of the distance matrix, its memory requirement is linear with the number of particles (histogram-dimension) N_p . In terms of space dimension n , OCD becomes memory intensive at high n since the deployed direct linear solver scales poorly in the high dimensional setting.

In order to estimate the total computational cost of OCD, including number of iterations required to reach stationary state, we conducted numerical tests as shown in Fig. 2, where OCD runs between two normal distributions. The flops and storage requirements confirm our analysis on complexity and memory consumption. However, it is evident that OCD requires significantly more operations compared to the reference Earth Mover's Distance (EMD) method [29], if relatively small number of sample points (e.g. 100) is considered. This is mainly due to the fact that while the number of operations scales efficiently, i.e. with $\mathcal{O}(N_p \log N_p)$, the pre-factor which depends on the number of iterations required to reach stationary condition becomes dominant, if N_p is relatively small. For $N_p > 64'000$, we observe that OCD becomes more efficient than EMD. On the other hand, as expected, OCD with the piecewise linear approximation of the conditional expectation scales poorly with the dimension, i.e. $\mathcal{O}(n^{3.5})$, while the measured timing of OCD with the piecewise constant estimator of the conditional expectation is slightly larger than expected $\mathcal{O}(n)$. We performed this study using the Google's Cloud TPU v6e with 172.9 GB RAM.

6. REPRESENTATIVE RESULTS FOR L^2 -MONGE-KANTOROVICH

In this section, we illustrate the accuracy and cost of the proposed OCD and highlight its capability in estimating the optimal map, along with Wasserstein distance, between two marginals in a few

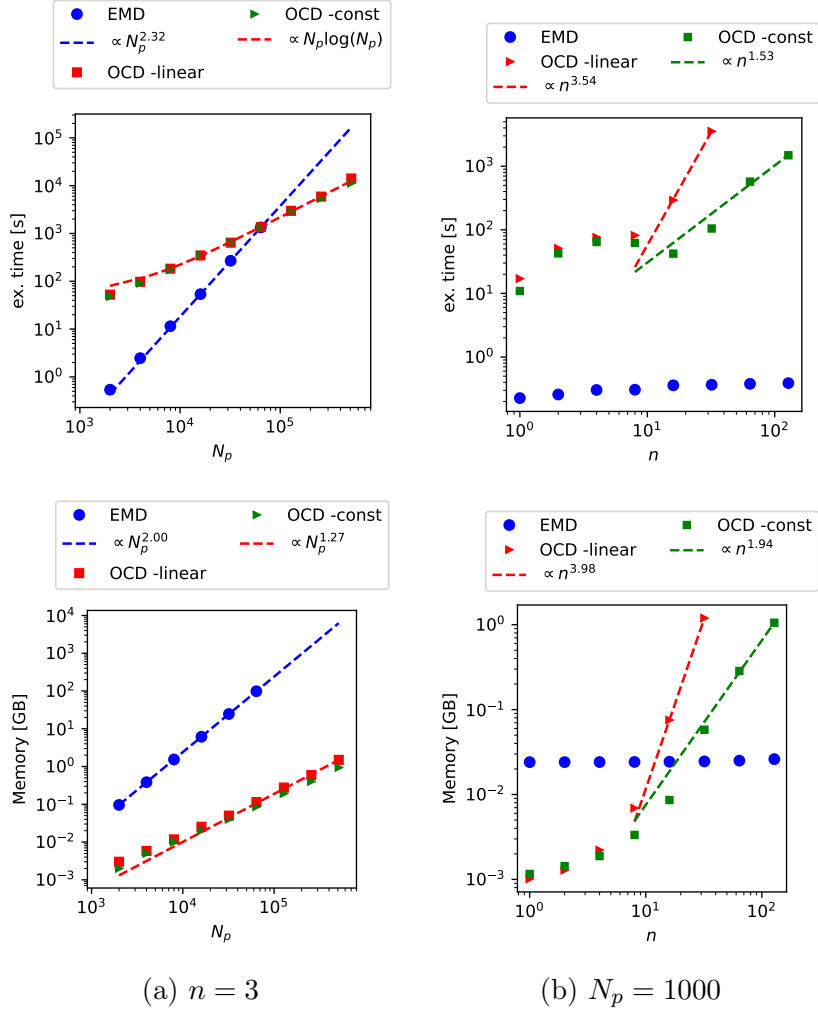


FIGURE 2. Execution time (top) and memory consumption (down) of OCD-piecewise linear algorithm with Runge Kutta 4th order solver with $\Delta t = 0.1$ against Earth Mover's Distance (EMD) method [29] in learning the map between samples of $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$ in (a) $n = 3$ for a range of N_p and (b) $n = 1, \dots, 128$ with $N_p = 1000$. Here we scale ϵ with a rule of thumb $\epsilon = 3nN_p^{-1/4}/4$.

examples. Unless mentioned otherwise, by OCD we are referring to OCD-piecewise linear algorithm equipped with Runge-Kutta 4th order ODE solver. We refer the reader to Supplementary Information for details on the choice of hyper-parameters such as ϵ and Δt , and comparison with benchmarks. All the performance studies are done on a single core and thread of Intel Core i7-8550U CPU with 1.80 GHz frequency.

6.1. Clustering and Prototypical Examples

First, we present and discuss three toy scenarios.

- (1) **Particle Clustering.** Starting from independent samples of two one-dimensional standard normal distributions, we track the particles evolved by L^2 -OCD. The dynamics is illustrated

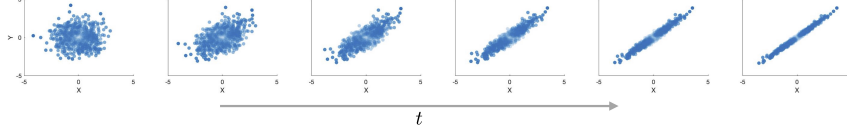


FIGURE 3. Evolution of point particles under the coupling created by OCD.

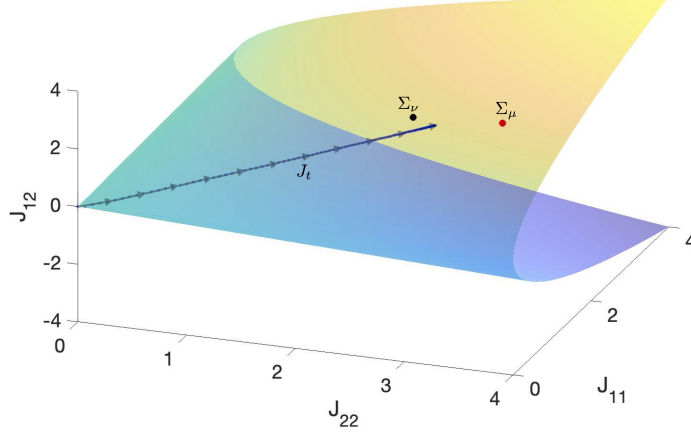
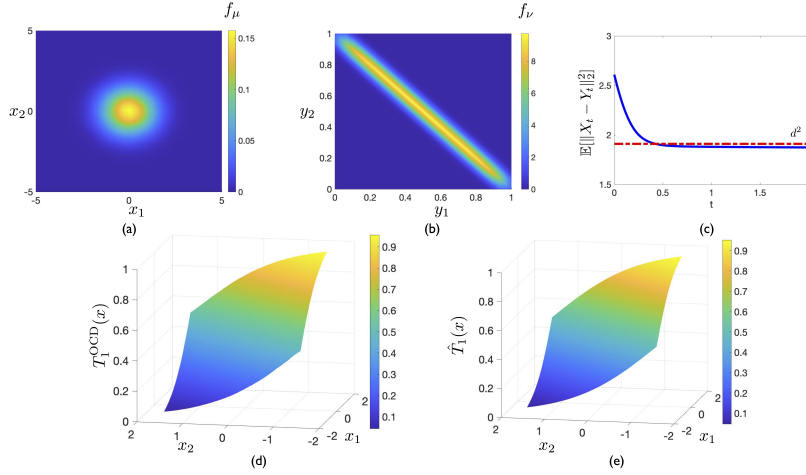
FIGURE 4. Evolution of cross-correlation matrix, J_t , in the cone of symmetric positive semi-definite matrices.

FIGURE 5. Recovery of nonlinear Monge map between (a) normal distribution and (b) its push-forward by softmax map $T_i = \nabla_{x_i} \log(\exp(x_1) + \exp(x_2))$; (c) Starting from independent samples, the pair (X_t, Y_t) is evolved by OCD and $\mathbb{E}\|X_t - Y_t\|_2^2$ converges towards the Wasserstein distance d^2 ; (d) First component of the map generated by OCD; (e) First component of the exact Monge map.

in Fig. 3, where the correlation between the data points monotonically increases in time. The Monge map, given by identity, emerges from particle clusters as time progresses.

- (2) **Symmetric Positive Semi-Definite Cone.** As a by-product of OCD, we justified in Proposition 2.5 that the dynamics ensures the cross-correlation matrix remains in the cone of

symmetric positive semi-definite matrices. Here we verify this finding by numerical simulation, where the initial condition is product measure of two centered Gaussians with covariance matrices Σ_μ and Σ_ν . As shown in Fig. 4, OCD evolves the cross-correlation inside the cone and reaches the maximum on the curve connecting Σ_μ and Σ_ν . Note that there exist interesting results on the geometry of the cross-correlation of Gaussians induced by the L^2 -optimal transport and its connection to geometric average of Σ_μ and Σ_ν , which we leave out of discussion for brevity, but refer the reader to [12].

- (3) **Nonlinear Monge Map.** To demonstrate relevance of OCD in recovering nonlinear multi-dimensional maps, we consider a synthetic setting where the target distribution is chosen such that the Monge map is known analytically. Consider the standard normal distribution over \mathbb{R}^2 , denoted by μ . Let the target distribution be the push-forward by softmax map $\nu = T\#\mu$, with $T = \nabla_x \log(\exp(x_1) + \exp(x_2))$. Since T is gradient of a convex function, the Monge map between μ and ν is simply T . Depicted in Fig. 5, starting from independent 10^6 samples of μ and ν , particles pushed by OCD with $\epsilon = 10^{-3}$ and $\Delta t = 10^{-2}$ successfully recover the Monge map and accurately estimate 2-Wasserstein distance, as t progresses. Note that the underestimation of the 2-Wasserstein distance is due to the fact that the marginals may not be preserved due to different sources of numerical errors, leading to, e.g. smaller variances of each marginal.

6.2. Distribution Learning and Sampling

We deploy the proposed OCD (with L^2 -cost) to learn transport map between the normal source and a target density $\in \{\text{Banana}, \text{Funnel}, \text{Swiss roll}\}$ [8]. We use 10^4 samples of each marginal and carry out OCD to find the samples of the optimal joint density on (\hat{X}^*, \hat{Y}^*) . Afterwards, in order to learn the map via the paired data points, i.e. $M_{X \rightarrow Y} : \hat{X}^* \rightarrow \hat{Y}^*$, we deploy NN with 4 layers, each with 100 neurons, where Adam’s algorithm [27] with a learning rate of 0.002 and 10^4 iterations are utilized to find the NN weights.

For testing, we generate 10^6 normally distributed samples, i.e. $X^{\text{test}} \sim \mathcal{N}(0, I_{n \times n})$ and feed them to the NN to find $M_{X \rightarrow Y}(X^{\text{test}})$. As shown in Fig. 6, OCD is capable of recovering the target density with a reasonable accuracy. However, we note that the estimated densities by OCD are slightly smoothed out compared to the ground-truth. This error could be related to the current approach of estimating the conditional expectation, which resembles Kernel Density Estimation with known smoothing effects. In Supplementary Information, we provide a study of hyper-parameters and comparison with Adaptive Transport Maps (ATM) [8] for each test case. We point out that once the map is learned, computing the new samples is straight-forward, given that the network only performs simple matrix-vector multiplication on the samples of the standard normal distribution. Hence the proposed workflow can efficiently serve as a sampling approach for complex distributions (such as Swiss roll), with ramifications for sampling algorithms based on measure transport [62, 50].

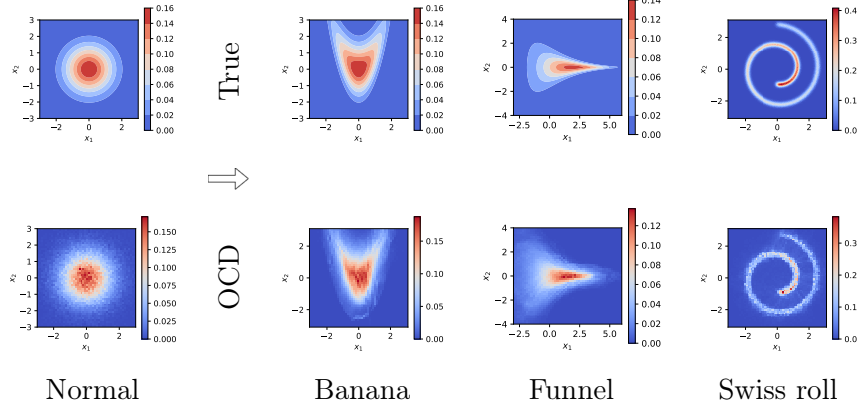


FIGURE 6. Estimating density of Banana, Funnel, and Swiss roll distributions using OCD. Here, first we find optimal pairs among 10^4 samples of normal and target density using OCD. We train NN on the sorted data and generate new samples of the target density using the learned map and 10^6 normally distributed particles. The true marginals are depicted on top and histogram of estimated marginals on bottom.

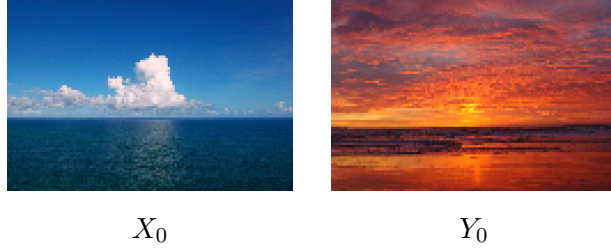


FIGURE 7. Low resolution (84×125 pixels) pictures of ocean in day and sunset time used for training.

6.3. Color Interpolation

As the final example, we consider another application of optimal transport in interpolation between color distribution of different images. Given an image with dimension $(N_x, N_y, 3)$, the pixel colors can be considered as the random variable, leading to $N_x \times N_y$ samples for X with $\dim(X) = 3$, indicating red/green/blue channels. We take a low resolution picture of the ocean taken in day and sunset time for training, as depicted in Fig. 7 (see examples of POT library [29] for details). Given $N_p = 10^5$ training samples of these two pictures, we learn the map $M_{X \rightarrow Y}$ using paired samples found from OCD (with L^2 -cost) and training NN similar to the one described in section 6.2. We employ $\Delta t = 0.1$ and $\epsilon = \epsilon_{\max}(N_{\text{clusters}}/N_p > 0.9)$. We repeat the same procedure for EMD to benchmark, and found that while EMD takes about 2.6 GB of memory, OCD only consumes 115 MB. Once the NN map is learned, we test the map by feeding the NN with the full resolution of day picture with 670×1000 pixels, where we see a good agreement with the EMD solution. We further test the learned OCD map on two new testing picture with 548×1024 and 1200×1920 pixels. As shown in Fig. 8, the trained OCD map is capable of transforming the testing picture into sunset.

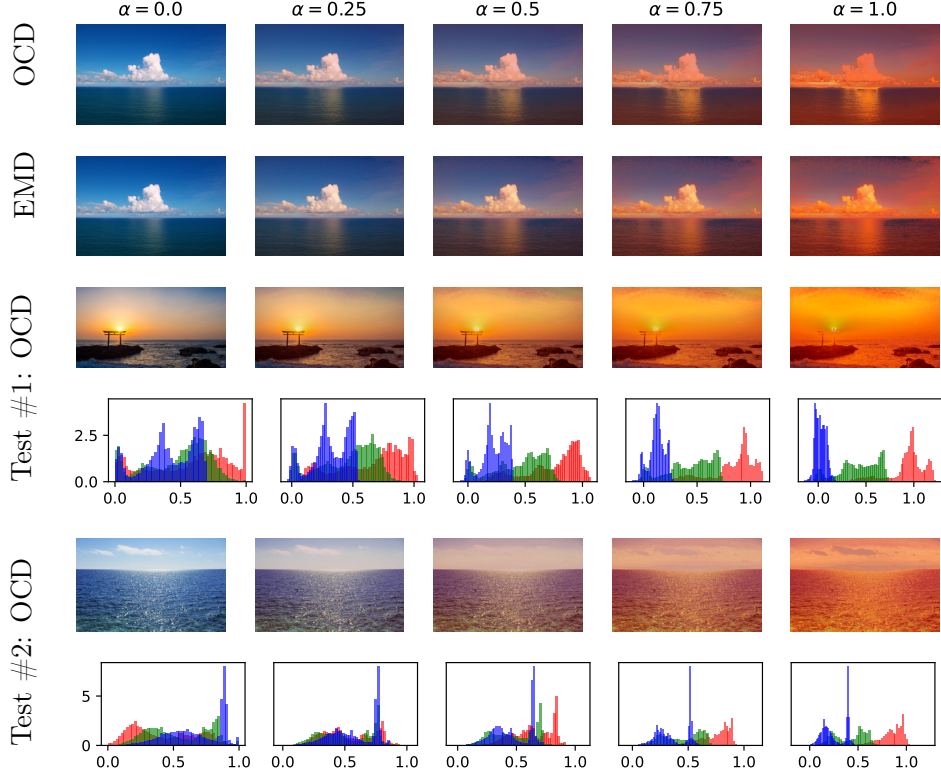


FIGURE 8. Interpolating day and sunset pictures of ocean. The interpolation is done using the learned map $M_{X \rightarrow Y}(\cdot)$ via $(1 - \alpha)X + \alpha M_{X \rightarrow Y}(X)$. The bottom two rows show the transport of colors for two test picture. The histograms show the distribution of red, green, and blue colors for the corresponding image.

6.4. Further numerical studies

In Supporting Information, we have provided more detailed numerical studies for the reader as a reference. This includes error analysis with respect to the OCD's free parameters, and further comparisons against other benchmark methods including the invertible map [8] and Neural optimal transport [40]. Furthermore, we show solution of the OCD with the L^p -cost function as well as OT map between two 10-dimensional multivariate normal distributions.

7. DISCUSSION

7.1. Concluding Remarks

Conceptually, what makes the OCD approach particularly appealing is that it is at cross-road of different topics:

- (1) The original problem of Monge-Kantorovich optimization, constrained on a set of couplings, is reduced to a dynamics whose evolution is controlled by the conditional expectation. The latter is the solution of the regression problem, and thereby, the OCD provides a direct link between Bayesian regression and the optimal transportation.

- (2) Using basic interaction rules, the dynamics gives a particle-type relaxation of independent samples, seeking their optimal coupling. The notion of clustering, which arises from this coupling, resonates with models of opinion dynamics and offers a method for obtaining sparse representations of complex, high-dimensional datasets.
- (3) The OCD framework has potential applications as the foundation for generative models. When combined with supervised learning, the learned maps could be employed to generate new datasets.
- (4) Finally, the dynamics of OCD can be interpreted as a form of projected gradient descent, providing a novel approach through which optimization techniques constrained on a set of distributional coupling can be constructed.

From numerical point of view, the model has favorable scaling behaviour. In the most basic form, it scales nearly linear with number of data points as well as dimension of distributional support, making OCD competitive in comparison to the state-of-the-art random map algorithms as well as direct linear programming approach. This powerful computational efficiency enables us to treat distributional learning problems over more complex and higher dimensional settings.

A crucial structural property of OCD, initialized by independent samples, is that the induced correlations remain symmetric and positive semidefinite. While this constitutes an interesting feature for randomized transport maps, it also implies that OCD, in its current deterministic formulation, does not in general recover the optimal coupling whenever the latter exhibits intrinsic asymmetries in its transport structure. In this sense, the symmetry constraint acts as an implicit bias, different in nature from the entropic bias induced by Schrödinger bridge or Sinkhorn-type methods (e.g. [43, 25, 20]). Whereas entropic algorithms favor couplings close to product measures, the OCD dynamics bias the evolution toward symmetric correlation structures.

The instability of suboptimal stationary solutions suggests a natural direction for relaxing this constraint: introducing controlled randomization into the dynamics, to explore the coupling space more flexibly. We anticipate that stochastic perturbations, e.g. in the form of annealing of the projected Langevin dynamics [23] may allow OCD-type flows to escape symmetric couplings and reach more general stationary couplings. A systematic theoretical characterization of the marginal distributions and structural conditions under which OCD recovers the optimal transport map remains an open research direction.

7.2. Limitations

The current treatment of conditional expectation exhibits diffusive behaviour which can deteriorate marginal distributions with sharp features. More elaborated approaches, e.g. deployment of Gaussian Processes or Neural Nets as hypothesis class of the conditional expectation, may alleviate this issue, at least to some degree, though likely at the cost of increased computational complexity. In addition, moment constraints of the marginal distributions can be leveraged to enforce the marginal densities of OCD to remain close to the initial condition, and hence reducing the numerical dispersion. Another issue that requires further investigation is the optimal choice of ϵ . Although we carried out extensive

numerical tests, see Supplementary Information, we believe computing the conditional expectation efficiently is a separate task by itself. These trade-offs between accuracy and efficiency are interesting areas for future research on OCD. From a theoretical standpoint, our understanding of the underlying dynamics remains incomplete. While we provided some important theoretical results on OCD, questions such as the well-posedness of OCD and its convergence towards optimal solution are still unresolved and require further exploration.

7.3. Broader Context

The presented dynamics, along with devised non-parametric algorithm, gives a means to pair up distinct data-sets, through a dynamics whose increment is a functional of the distribution. This is a prototype of McKean-Vlasov processes, which offer further generalizability compared to the conventional regression, due its dependency on the distribution. We believe this development holds implications which could go beyond the context of optimal transport, specifically given recent attention towards McKean-Vlasov universal approximators emerging in data-driven trainable models [67, 6].

8. ACKNOWLEDGMENTS

MS thanks Andreas Adelmann for his support and constructive feedback. PME was supported by the ERC grant TRUST-949796 and the NSERC Discovery grant RGPIN-2025-06544.

REFERENCES

- [1] D. ALVAREZ-MELIS, T. JAAKKOLA, AND S. JEGELKA, *Structured optimal transport*, in International conference on artificial intelligence and statistics, PMLR, 2018, pp. 1771–1780.
- [2] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.
- [3] L. AMBROSIO AND G. SAVARÉ, *Gradient flows of probability measures*, in Handbook of differential equations: evolutionary equations, vol. 3, Elsevier, 2007, pp. 1–136.
- [4] S. ANGENENT, S. HAKER, AND A. TANNENBAUM, *Minimizing flows for the monge–kantorovich problem*, SIAM journal on mathematical analysis, 35 (2003), pp. 61–97.
- [5] M. ARJOVSKY, S. CHINTALA, AND L. BOTTOU, *Wasserstein generative adversarial networks*, in International conference on machine learning, PMLR, 2017, pp. 214–223.
- [6] L. ATANACKOVIC, X. ZHANG, B. AMOS, M. BLANCHETTE, L. J. LEE, Y. BENGIO, A. TONG, AND K. NEKLYUDOV, *Meta flow matching: Integrating vector fields on the Wasserstein manifold*, arXiv preprint arXiv:2408.14608, (2024).
- [7] M. BALLU, Q. BERTHET, AND F. BACH, *Stochastic optimization for regularized wasserstein estimators*, in International Conference on Machine Learning, PMLR, 2020, pp. 602–612.
- [8] R. BAPTISTA, Y. MARZOUK, AND O. ZAHM, *On the representation and learning of monotone triangular transport maps*, Foundations of Computational Mathematics, (2023), pp. 1–46.
- [9] K. BASHIRI, *Gradient Flows, Metastability and Interacting Particle Systems*, PhD thesis, Universitäts-und Landesbibliothek Bonn, 2020.
- [10] J.-D. BENAMOU AND Y. BRENIER, *A computational fluid mechanics solution to the monge-kantorovich mass transfer problem*, Numerische Mathematik, 84 (2000), pp. 375–393.
- [11] A. BERNOU AND Y. LIU, *Path-dependent mckean-vlasov equation: strong well-posedness, propagation of chaos and convergence of an interpolated euler scheme*, (2023).
- [12] R. BHATIA, T. JAIN, AND Y. LIM, *On the bures–wasserstein distance between positive definite matrices*, Expositiones Mathematicae, 37 (2019), pp. 165–191.

- [13] V. I. BOGACHEV AND A. V. KOLESNIKOV, *The monge-kantorovich problem: achievements, connections, and perspectives*, Russian Mathematical Surveys, 67 (2012), p. 785.
- [14] N. BONNEEL, M. VAN DE PANNE, S. PARIS, AND W. HEIDRICH, *Displacement interpolation using lagrangian mass transport*, in Proceedings of the 2011 SIGGRAPH Asia conference, 2011, pp. 1–12.
- [15] A. A. BOROVKOV, *Probability theory*, CRC Press, 1999.
- [16] N. BOUMAL, *An introduction to optimization on smooth manifolds*, Cambridge University Press, 2023.
- [17] Y. BRENIER, *Extended monge-kantorovich theory*, Lecture Notes in Mathematics, (2003), pp. 91–122.
- [18] G. BUTTAZZO, L. DE PASCALE, AND P. GORI-GIORGI, *Optimal-transport formulation of electronic density-functional theory*, Physical Review A—Atomic, Molecular, and Optical Physics, 85 (2012), p. 062502.
- [19] R. CHARTRAND, B. WOHLBERG, K. VIXIE, AND E. BOLLT, *A gradient descent solution to the monge-kantorovich problem*, Applied Mathematical Sciences, 3 (2009), pp. 1071–1080.
- [20] Y. CHEN, T. T. GEORGIU, AND M. PAVON, *Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge*, Siam Review, 63 (2021), pp. 249–313.
- [21] A. J. CHORIN, O. H. HALD, AND R. KUPFERMAN, *Optimal prediction and the mori-zwanzig representation of irreversible processes*, Proceedings of the National Academy of Sciences, 97 (2000), pp. 2968–2973.
- [22] K. L. CHUNG, *A course in probability theory*, Elsevier, 2000.
- [23] G. CONFORTI, D. LACKER, AND S. PAL, *Projected langevin dynamics and a gradient flow for entropic optimal transport*, arXiv preprint arXiv:2309.08598, (2023).
- [24] C. COTAR, G. FRIESECKE, AND C. KLÜPPELBERG, *Density functional theory and optimal transportation with coulomb cost*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 548–599.
- [25] M. CUTURI, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 26 (2013).
- [26] R. DAKHMOUCHE, I. LUNATI, AND H. GORJI, *Robust symbolic regression for network trajectory inference*, in ICLR 2024 Workshop on Machine Learning for Genomics Explorations.
- [27] P. K. DIEDERIK, *Adam: A method for stochastic optimization*, (2014).
- [28] T. A. EL MOSELHY AND Y. M. MARZOUK, *Bayesian inference with optimal maps*, Journal of Computational Physics, 231 (2012), pp. 7815–7850.
- [29] R. FLAMARY, N. COURTY, A. GRAMFORT, M. Z. ALAYA, A. BOISBUNON, S. CHAMBON, L. CHAPEL, A. CORENFLOS, K. FATRAS, N. FOURNIER, ET AL., *Pot: Python optimal transport*, Journal of Machine Learning Research, 22 (2021), pp. 1–8.
- [30] M. GELBRICH, *On a formula for the l_2 wasserstein metric between measures on euclidean and hilbert spaces*, Mathematische Nachrichten, 147 (1990), pp. 185–203.
- [31] I. GENTIL, C. LÉONARD, AND L. RIPANI, *Dynamical aspects of the generalized schrödinger problem via otto calculus—a heuristic point of view*, Revista matemática iberoamericana, 36 (2020), pp. 1071–1112.
- [32] D. GIVON, R. KUPFERMAN, AND O. H. HALD, *Existence proof for orthogonal dynamics and the mori-zwanzig formalism*, Israel Journal of Mathematics, 145 (2005), pp. 221–241.
- [33] Z. GOLDFELD, K. KATO, G. RIOUX, AND R. SADHU, *Statistical inference with regularized optimal transport*, Information and Inference: A Journal of the IMA, 13 (2024), p. iaad056.
- [34] F. GOLSE, *The mean-field limit for the dynamics of large particle systems*, Journées équations aux dérivées partielles, (2003), pp. 1–47.
- [35] P. GORDALIZA, E. DEL BARRIO, G. FABRICE, AND J.-M. LOUBES, *Obtaining fairness using optimal transport theory*, in International conference on machine learning, PMLR, 2019, pp. 2357–2365.
- [36] R. HEGSELMANN, U. KRAUSE, ET AL., *Truth and cognitive division of labor: First steps towards a computer aided social epistemology*, Journal of Artificial Societies and Social Simulation, 9 (2006), p. 10.
- [37] R. JORDAN, D. KINDERLEHRER, AND F. OTTO, *The variational formulation of the Fokker–Planck equation*, SIAM journal on mathematical analysis, 29 (1998), pp. 1–17.
- [38] I. KLEBANOV, B. SPRUNGK, AND T. J. SULLIVAN, *The linear conditional expectation in hilbert space*, Bernoulli, 27 (2021), pp. 2267–2299.

- [39] M. A. S. KOLARIJANI, A. V. PROSKURNIKOV, AND P. MOHAJERIN ESFAHANI, *Macroscopic noisy bounded confidence models with distributed radical opinions*, IEEE Transactions on Automatic Control, 66 (2020), pp. 1174–1189.
- [40] A. KOROTIN, D. SELIKHANOVYCH, AND E. BURNAEV, *Neural optimal transport*, in The Eleventh International Conference on Learning Representations, 2023.
- [41] D. P. KROESE, T. TAIMRE, AND Z. I. BOTEV, *Handbook of monte carlo methods*, John Wiley & Sons, 2013.
- [42] D. KUHN, P. MOHAJERIN ESFAHANI, V. A. NGUYEN, AND S. SHAFIEEZADEH-ABADEH, *Wasserstein distributionally robust optimization: Theory and applications in machine learning*, in Operations research & management science in the age of analytics, Informs, 2019, pp. 130–166.
- [43] C. LÉONARD, *A survey of the schrödinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215, (2013).
- [44] L. LI, A. SCAGLIONE, A. SWAMI, AND Q. ZHAO, *Consensus, polarization and clustering of opinions in social networks*, IEEE Journal on Selected Areas in Communications, 31 (2013), pp. 1072–1083.
- [45] P. MOHAJERIN ESFAHANI AND D. KUHN, *Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.
- [46] S. MOTSCH AND E. TADMOR, *Heterophilous dynamics enhances consensus*, SIAM review, 56 (2014), pp. 577–621.
- [47] O. MULA AND A. NOUY, *Moment-sos methods for optimal transport problems*, Numerische Mathematik, (2024), pp. 1–38.
- [48] E. A. NADARAYA, *On estimating regression*, Theory of Probability & Its Applications, 9 (1964), pp. 141–142.
- [49] S. M. OMOHUNDRO, *Five balltree construction algorithms*, tech. rep., 1989.
- [50] G. PAPAMAKARIOS, E. NALISNICK, D. J. REZENDE, S. MOHAMED, AND B. LAKSHMINARAYANAN, *Normalizing flows for probabilistic modeling and inference*, Journal of Machine Learning Research, 22 (2021), pp. 1–64.
- [51] T. PAUL AND E. TRÉLAT, *From microscopic to macroscopic scale equations: mean field, hydrodynamic and graph limits*, arXiv preprint arXiv:2209.08832, (2022).
- [52] L. PETZOLD, *Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations*, SIAM journal on scientific and statistical computing, 4 (1983), pp. 136–148.
- [53] G. PEYRÉ, M. CUTURI, ET AL., *Computational optimal transport: With applications to data science*, Foundations and Trends in Machine Learning, 11 (2019), pp. 355–607.
- [54] S. T. RACHEV AND L. RÜSCHENDORF, *Mass Transportation Problems: Volume 1: Theory*, Springer Science & Business Media, 2006.
- [55] L. RÜSCHENDORF, *The wasserstein distance and approximation theorems*, Probability Theory and Related Fields, 70 (1985), pp. 117–129.
- [56] L. RÜSCHENDORF AND S. T. RACHEV, *A characterization of random variables with minimum l_2 -distance*, Journal of multivariate analysis, 32 (1990), pp. 48–54.
- [57] M. SADR, N. G. HADJICONSTANTINOY, AND M. H. GORJI, *Wasserstein-penalized entropy closure: A use case for stochastic particle methods*, Journal of Computational Physics, 511 (2024), p. 113066.
- [58] F. SANTAMBROGIO, *Optimal transport for applied mathematicians*, Birkhäuser, NY, 55 (2015), p. 94.
- [59] ———, *Euclidean, metric, and Wasserstein gradient flows: an overview*, Bulletin of Mathematical Sciences, 7 (2017), pp. 87–154.
- [60] L.-P. SAUMIER, M. AGUEH, AND B. KHOUIDER, *An efficient numerical algorithm for the l_2 optimal transport problem with periodic densities*, The IMA Journal of Applied Mathematics, 80 (2015), pp. 135–157.
- [61] E. SCHUBERT, J. SANDER, M. ESTER, H. P. KRIEGLER, AND X. XU, *Dbscan revisited, revisited: why and how you should (still) use dbscan*, ACM Transactions on Database Systems (TODS), 42 (2017), pp. 1–21.
- [62] J. SUN, J. BERNER, L. RICHTER, M. ZEINHOFFER, J. MÜLLER, K. AZIZZADENESHELI, AND A. ANANDKUMAR, *Dynamical measure transport and neural pde solvers for sampling*, arXiv preprint arXiv:2407.07873, (2024).
- [63] A.-S. SZNITMAN, *Topics in propagation of chaos. école d’été de probabilités de saint-flour xix—1989*, 165–251, Lecture Notes in Math, 1464 (1991).
- [64] C. THRAMPOLIDIS, S. OYMAK, AND B. HASSIBI, *Regularized linear regression: A precise analysis of the estimation error*, in Conference on Learning Theory, PMLR, 2015, pp. 1683–1709.

- [65] G. TRIGILA AND E. G. TABAK, *Data-driven optimal transport*, Communications on Pure and Applied Mathematics, 69 (2016), pp. 613–648.
- [66] C. VILLANI, *Optimal transport: old and new*, vol. 338, Springer, 2009.
- [67] H. YANG, A. HASAN, Y. NG, AND V. TAROKH, *Neural mckean-vlasov processes: Distributional dependence in diffusion processes*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 262–270.

APPENDIX A. NUMERICAL SCHEME AND SOLUTION ALGORITHM

The algorithmic steps of OCD consist of estimating the conditional expectation, along with an appropriate time integration scheme. Explicit formulas are given for both in the following.

A.1. Estimation of Conditional Expectation

- (1) Piecewise constant estimation follows simple averaging per cluster, which gives

$$\hat{\mathcal{K}}_X(\hat{X}_i^n | \hat{X}^n, \hat{Y}^n) = \frac{1}{N_{\bar{I}^{\hat{X}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{X}_i^n, \epsilon}} \nabla_x c(\hat{X}_i^n, \hat{Y}_j^n) \quad (50)$$

$$\text{and } \hat{\mathcal{K}}_Y(\hat{Y}_i^n | \hat{X}^n, \hat{Y}^n) = \frac{1}{N_{\bar{I}^{\hat{Y}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{Y}_i^n, \epsilon}} \nabla_y c(\hat{X}_j^n, \hat{Y}_i^n). \quad (51)$$

- (2) For the piecewise linear estimation, first, let us define the average values per cluster

$$\hat{m}_X^n = \frac{1}{N_{\bar{I}^{\hat{X}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{X}_i^n, \epsilon}} \nabla_x c(\hat{X}_j^n, \hat{Y}_j^n) \quad \text{and} \quad \hat{m}_Y^n = \frac{1}{N_{\bar{I}^{\hat{Y}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{Y}_i^n, \epsilon}} \nabla_y c(\hat{X}_j^n, \hat{Y}_j^n).$$

Accordingly, we get the correlation matrices

$$\begin{aligned} \hat{\Sigma}_{XX}^n &= \frac{1}{N_{\bar{I}^{\hat{X}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{X}_i^n, \epsilon}} \left(\nabla_x c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_X^n \right) \otimes \left(\nabla_x c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_X^n \right), \\ \hat{\Sigma}_{YY}^n &= \frac{1}{N_{\bar{I}^{\hat{Y}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{Y}_i^n, \epsilon}} \left(\nabla_y c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_Y^n \right) \otimes \left(\nabla_y c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_Y^n \right), \\ \hat{\Sigma}_{XY}^n &= \frac{1}{N_{\bar{I}^{\hat{X}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{X}_i^n, \epsilon}} \left(\nabla_x c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_X^n \right) \otimes \left(\nabla_y c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_Y^n \right) \\ \text{and } \hat{\Sigma}_{YX}^n &= \frac{1}{N_{\bar{I}^{\hat{Y}_i^n, \epsilon}}} \sum_{j \in \bar{I}^{\hat{Y}_i^n, \epsilon}} \left(\nabla_y c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_Y^n \right) \otimes \left(\nabla_x c(\hat{X}_j^n, \hat{Y}_j^n) - \hat{m}_X^n \right), \end{aligned}$$

leading to

$$\begin{aligned} \hat{\mathcal{K}}_X(\hat{X}_i^n | \hat{X}^n, \hat{Y}^n) &= \hat{m}_Y^n + (\hat{\Sigma}_{XY}^n)^T (\hat{\Sigma}_{XX}^n)^{-1} (\hat{X}_i^n - \hat{m}_X^n) \\ \text{and } \hat{\mathcal{K}}_Y(\hat{Y}_i^n | \hat{X}^n, \hat{Y}^n) &= \hat{m}_X^n + (\hat{\Sigma}_{YX}^n)^T (\hat{\Sigma}_{YY}^n)^{-1} (\hat{Y}_i^n - \hat{m}_Y^n). \end{aligned}$$

If the number of particles in a cluster drop below two, the correlation matrix becomes singular. In order to ensure that the matrices $\hat{\Sigma}_{XX, YX}^n$ remain invertible, they are regularized by $\hat{\epsilon} > 0$, as the following

$$\hat{\mathcal{K}}_X^{\hat{\epsilon}}(\hat{X}_i^n | \hat{X}^n, \hat{Y}^n) = \hat{m}_Y^n + (\hat{\Sigma}_{XY}^n)^T (\hat{\Sigma}_{XX}^n + \hat{\epsilon} I_X)^{-1} (\hat{X}_i^n - \hat{m}_X^n) \quad (52)$$

$$\text{and } \hat{\mathcal{K}}_Y^{\hat{\epsilon}}(\hat{Y}_i^n | \hat{X}^n, \hat{Y}^n) = \hat{m}_X^n + (\hat{\Sigma}_{XX}^n)^T (\hat{\Sigma}_{YY}^n + \hat{\epsilon} I_Y)^{-1} (\hat{Y}_i^n - \hat{m}_Y^n), \quad (53)$$

where $I_{X,Y}$ are identity matrices with the same size as $\hat{\Sigma}_{XX,YY}^n$, respectively [64].

A.2. Time Discretization

For the fourth-stage Runge-Kutta, we evaluate the slopes

$$k_{1,i}^X = \nabla_x c(\hat{X}_i^n, \hat{Y}_i^n) - \hat{\mathcal{K}}_X(\hat{X}_i^n | \hat{X}^n, \hat{Y}^n) \quad (54)$$

$$\text{and } k_{1,i}^Y = \nabla_y c(\hat{X}_i^n, \hat{Y}_i^n) - \hat{\mathcal{K}}_Y(\hat{Y}_i^n | \hat{X}^n, \hat{Y}^n) \quad (55)$$

at the beginning,

$$k_{2,i}^X = \nabla_x c(\hat{X}_i^{k_1}, \hat{Y}_i^{k_1}) - \hat{\mathcal{K}}_X(\hat{X}_i^{k_1} | \hat{X}^{k_1}, \hat{Y}^{k_1}) \quad (56)$$

$$\text{and } k_{2,i}^Y = \nabla_y c(\hat{X}_i^{k_1}, \hat{Y}_i^{k_1}) - \hat{\mathcal{K}}_Y(\hat{Y}_i^{k_1} | \hat{X}^{k_1}, \hat{Y}^{k_1}) \quad (57)$$

with $\hat{X}_i^{k_1} = \hat{X}_i^n + k_{1,i}^X \Delta t / 2$ and $\hat{Y}_i^{k_1} = \hat{Y}_i^n + k_{1,i}^Y \Delta t / 2$, at the first mid-point,

$$k_{3,i}^X = \nabla_x c(\hat{X}_i^{k_2}, \hat{Y}_i^{k_2}) - \hat{\mathcal{K}}_X(\hat{X}_i^{k_2} | \hat{X}^{k_2}, \hat{Y}^{k_2}) \quad (58)$$

$$\text{and } k_{3,i}^Y = \nabla_y c(\hat{X}_i^{k_2}, \hat{Y}_i^{k_2}) - \hat{\mathcal{K}}_Y(\hat{Y}_i^{k_2} | \hat{X}^{k_2}, \hat{Y}^{k_2}) \quad (59)$$

with $\hat{X}_i^{k_2} = \hat{X}_i^{k_1} + k_{2,i}^X \Delta t / 2$ and $\hat{Y}_i^{k_2} = \hat{Y}_i^{k_1} + k_{2,i}^Y \Delta t / 2$, at the second mid-point, and

$$k_{4,i}^X = \nabla_x c(\hat{X}_i^{k_3}, \hat{Y}_i^{k_3}) - \hat{\mathcal{K}}_X(\hat{X}_i^{k_3} | \hat{X}^{k_3}, \hat{Y}^{k_3}) \quad (60)$$

$$\text{and } k_{4,i}^Y = \nabla_y c(\hat{X}_i^{k_3}, \hat{Y}_i^{k_3}) - \hat{\mathcal{K}}_Y(\hat{Y}_i^{k_3} | \hat{X}^{k_3}, \hat{Y}^{k_3}) \quad (61)$$

with $\hat{X}_i^{k_3} = \hat{X}_i^{k_2} + k_{3,i}^X \Delta t$ and $\hat{Y}_i^{k_3} = \hat{Y}_i^{k_2} + k_{3,i}^Y \Delta t$, at the end-point, where for notation brevity we omit the dependency on the regularizer $\hat{\epsilon}$. The update for full Δt then is given by weighted average of all four slopes

$$\hat{X}_i^{n+1} = \hat{X}_i^n + \frac{1}{6} (k_1^X + 2k_2^X + 2k_3^X + k_4^X) \quad (62)$$

$$\text{and } \hat{Y}_i^{n+1} = \hat{Y}_i^n + \frac{1}{6} (k_1^Y + 2k_2^Y + 2k_3^Y + k_4^Y) . \quad (63)$$

See algorithm 1 for summary of L^2 -OCD computational step.

APPENDIX B. OPTIMAL CUTOFF

We perform numerical experiments to investigate the optimal value of ϵ by considering the transport map between two normal distributions $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$. In particular, given N_p particles of these two marginals, we run the L^2 -OCD algorithm until convergence, for a range of ϵ in one and two dimensional probability space, and analyse our finding. In all experiments here, we consider $N_p \in \{100, 200, 400, 800, 1'600\}$ and $\Delta t = 10^{-3}$, where maximum of 20'000 steps are employed. Furthermore, both piecewise linear and piecewise constant estimation of conditional expectations are used to characterize ϵ .

While the Monge map is a simple affine transformation in this setting, we use a numerical solution of the linear programming, noted here by EMD [14], to understand the quality of OCD

Algorithm 1 Main steps of the OCD algorithm for L^2 -cost, using piecewise linear estimation of the conditional expectation and fourth-order Runge-Kutta.

```

Input  $N_p$  samples of  $(\hat{X}^0, \hat{Y}^0)$ 
Set  $\beta = 0.9$ ,  $\epsilon = \max(\epsilon |N_{\text{clusters}}/N_p > \beta)$  and  $\hat{\epsilon} = 0$ 
Set convergence threshold  $\gamma = 0.01$ 
Set  $t_f, n_t$  and time step size  $\Delta t = t_f/n_t$  (default value  $\Delta t = 0.1$ )
while  $\mathbb{E}[\|\hat{X}^n - \hat{Y}^n\|_2^2] > \gamma$  do
     $\hat{X}^{k_0} \leftarrow \hat{X}^n$ 
     $\hat{Y}^{k_0} \leftarrow \hat{Y}^n$ 
    for  $j = 0 : 2$  do
        indX  $\leftarrow$  BallTree( $\hat{X}^{k_j}, \epsilon$ )
        indY  $\leftarrow$  BallTree( $\hat{Y}^{k_j}, \epsilon$ )
        for  $i = 1 : N_p$  do
             $\bar{I}_{\hat{X}_i^{k_j}, \epsilon} \leftarrow \text{indX}[i]$ 
             $\bar{I}_{\hat{Y}_i^{k_j}, \epsilon} \leftarrow \text{indY}[i]$ 
            Estimate  $\mathbb{E}[\nabla_x c(\hat{X}^{k_j}, \hat{Y}^{k_j}) | \hat{X}_i^{k_j}]$  using Eq. (52)
            Estimate  $\mathbb{E}[\nabla_y c(\hat{X}^{k_j}, \hat{Y}^{k_j}) | \hat{Y}_i^{k_j}]$  using Eq. (53)
            Compute  $k_{j+1,i}^X$  using Eqs. (54), (56), (58)
            Compute  $k_{j+1,i}^Y$  using Eqs. (55), (57), (59)
             $\hat{X}_i^{k_{j+1}} = \hat{X}_i^n + k_{j+1,i}^X \Delta t / w[j]$ 
             $\hat{Y}_i^{k_{j+1}} = \hat{Y}_i^n + k_{j+1,i}^Y \Delta t / w[j]$ 
        end for
        Compute  $k_{4,i}^X$  using Eq. (60)
        Compute  $k_{4,i}^Y$  using Eq. (61)
    end for
     $\hat{X}^{n+1} \leftarrow \hat{X}^n + (k_1^X + 2k_2^X + 2k_3^X + k_4^X)/6$ 
     $\hat{Y}^{n+1} \leftarrow \hat{Y}^n + (k_1^Y + 2k_2^Y + 2k_3^Y + k_4^Y)/6$ 
    Increment  $n$ 
end while

```

results with respect to the EMD benchmark. To find the optimal value denoted by ϵ^* , from several experiments of OCD with a range of ϵ , we compare the solution of OCD with EMD by computing the Wasserstein distance of their joint densities. Therefore, we construct the discrete joint density by concatenation $(X, Y) \sim \pi$ and compute $d^2(\pi^{\text{EMD}}, \pi^{\text{OCD}})$. Furthermore, we study the number of clusters created for each OCD run, using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method [61].

From Figs. 9-10, it is clear that for small ϵ , i.e. $\epsilon \rightarrow 0$, the number of clusters does not differ much from the number of particles N_p . In this case, the initial particles remain frozen and do not see each other, hence the OCD algorithm returns the initial joint density (which is density of the product measure). However, for large values of ϵ , the number of clusters decreases to a

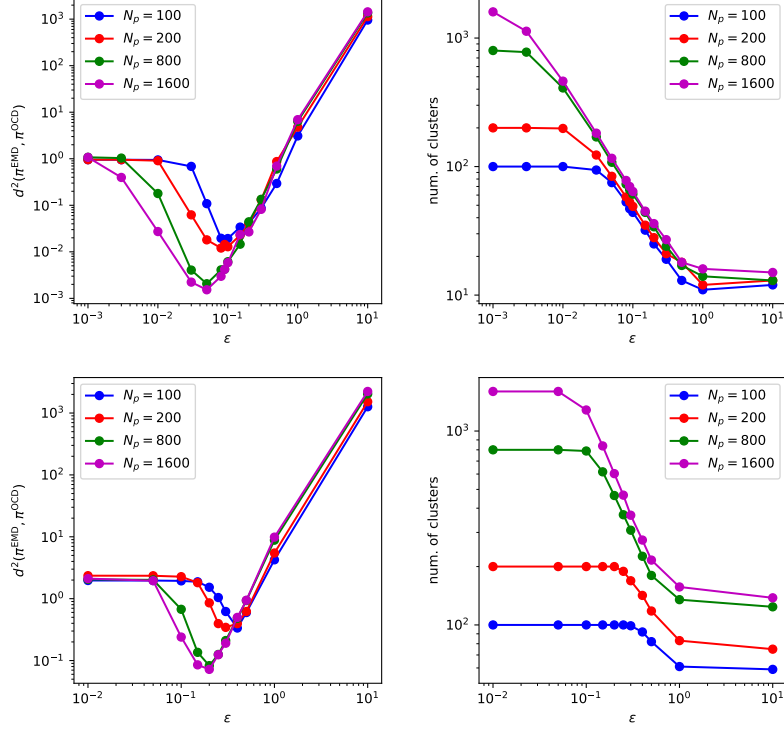


FIGURE 9. Error in finding the optimal map between marginals $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$ as well as the number of clusters in 1-dimensional space, i.e. $\dim(X) = 1$ (top) and 2-dimensional space, i.e. $\dim(X) = 2$ (bottom). The piecewise constant approximation is used to estimate the conditional expectation.

fixed value depending on N_p , dimension of X and Y , while the Wasserstein distance of OCD from the solution of EMD increases exponentially (for piecewise constant approximation of conditional expectation). Note that we observe a decay in error for large ϵ , if linear approximation of conditional expectation is employed. This can be explained by the fact that linear regression is indeed optimal if conditional expectation with respect to Gaussian measure is considered. Since here both marginals are Gaussians, we observe an improved performance of the piecewise linear approximation, for very large ϵ . Nevertheless, in all considered scenarios, the optimal parameter ϵ^* lies somewhere between two extremes of $\epsilon \rightarrow 0$ and $\epsilon \rightarrow \infty$.

In fig. 11, we plot the value of optimal parameter ϵ^* against N_p . Interestingly, in both cases of $\dim(X) = \dim(Y) = 1$ and 2, we observe that the optimal value scales almost $\epsilon^* \sim N_p^{-1/5}$ up to statistical noise. Since our proposed approach in computing the conditional expectation (piecewise constant) can be seen as a kernel density estimator with a step function as the kernel, the scaling of the optimal bandwidth $N_p^{-1/5}$ is expected.

Another insight can be gained by considering number of clusters as a guiding proxy for ϵ^* . We consider two proxies that could relate ϵ^* to the number of clusters:

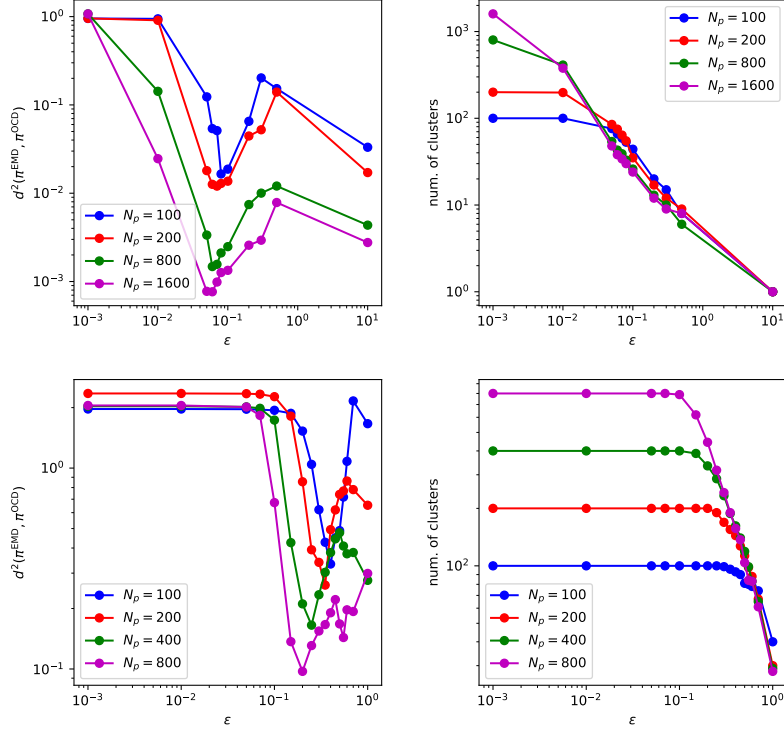


FIGURE 10. Error in finding the optimal map between marginals $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$, as well as the number of clusters in 1-dimensional space, i.e. $\dim(X) = 1$ (top) and 2-dimensional space, i.e. $\dim(X) = 2$ (bottom). The piecewise linear approximation is considered for the estimation of conditional expectation.

- (1) ϵ_{crit} : the critical point of number of clusters versus ϵ .
- (2) ϵ_{max} : the maximum ϵ with $N_{\text{clusters}}/N_p > \beta$, with β is a user-defined tolerance.

Numerical results presented in Figs. 12-13 motivate that both proxies lead to reasonable approximations of ϵ^* . While there is no clear consensus on all the considered numerical tests, we note that the optimal ϵ^* is typically close to one of the mentioned proxies. Therefore, the user can preform a grid search around either ϵ_{crit} or ϵ_{max} . While presented proxies provide some guidance, we emphasize that, in general, the optimal ϵ^* should be treated as a hyper-parameter, and to be set by the user with try and error.

APPENDIX C. GALLERY OF DENSITY ESTIMATION

To further assess the quality, robustness, and computational cost of OCD, with respect to the state-of-the-art, we compare the learned map between Gaussian and a target distribution using OCD (with L^2 -cost) and ATM. As target distributions, we consider Swiss roll, banana, and funnel, as shown in Figs. 14-16. In all these experiments, we varied the hyper-parameters of OCD and ATM, and checked how the learned map is affected. We employed 10^4 samples of the target density in the learning phase, and generated 10^6 samples of the fitted density to produce the histograms. We solve the OCD equations using RK4 solver and use piecewise linear approximation for the conditional

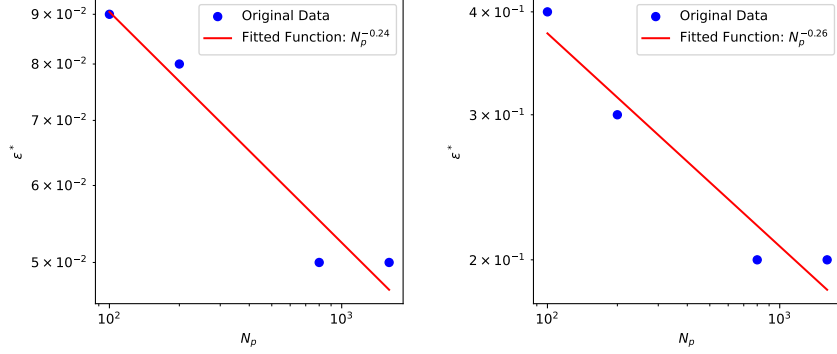


FIGURE 11. Optimal cutoff denoted by ϵ^* as a function of N_p , given two marginals $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$ over 1-dimensional space, i.e. $\dim(X) = 1$ (left) and 2-dimensional space, i.e. $\dim(X) = 2$ (right). The piecewise constant approximation is employed to estimate the conditional expectation.

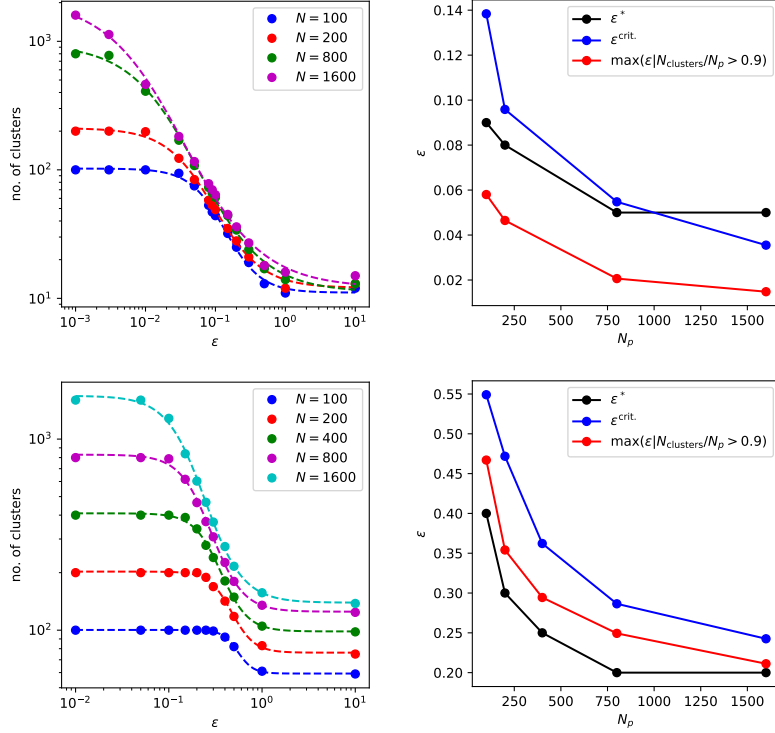


FIGURE 12. Comparing the critical point of the number of cluster versus ϵ in 1-dimensional space (top) and two-dimensional space (bottom), where $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$. The piecewise constant approximation is employed for the estimation of conditional expectation.

expectation, see algorithm 1. In case of ATM, we used an additional of 500 samples for validation and performed 5 folded search. In table 1, we report the computational cost of OCD and ATM for the Swiss roll example. For reference, the computational cost of EDM is also reported.

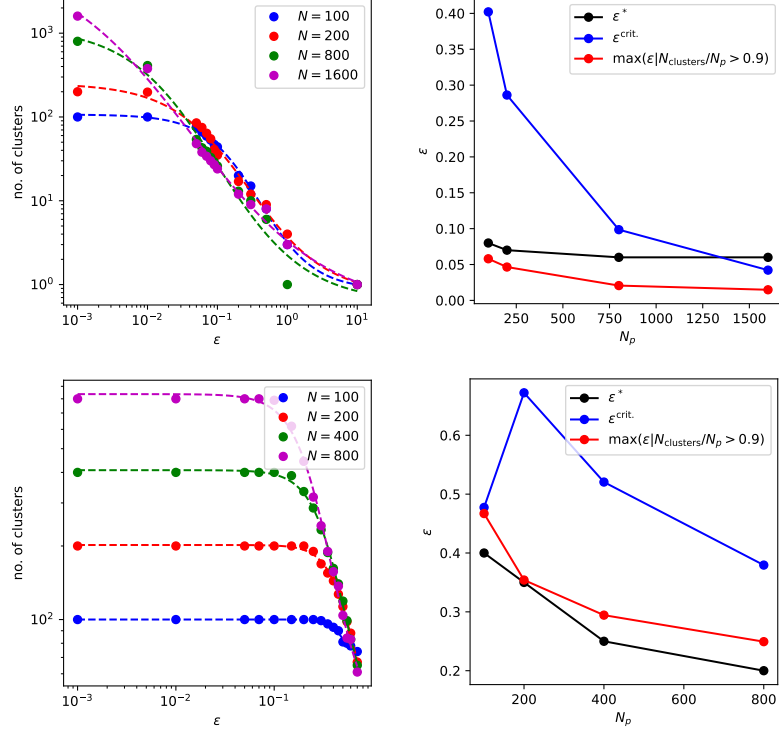


FIGURE 13. Comparing the critical point of the number of cluster versus ϵ in 1-dimensional space (top) and 2-dimensional space (bottom), where $\mu = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(1, I)$. The piecewise linear approximation is considered for the estimation of conditional expectation.

Method	Hyper parameters	Ex. time [s]	Memory [MB]
EMD		22.5	2400.75
OCD	$\epsilon = 0.025, \Delta t = 0.1$	91.6	30.9
	$\epsilon = 0.03, \Delta t = 0.1$	84.4	37.1
	$\epsilon = 0.035, \Delta t = 0.1$	84.5	37.2
	$\epsilon = 0.03, \Delta t = 0.01$	743.2	36.2
	$\epsilon = 0.03, \Delta t = 0.001$	8421.2	40.4
ATM	10 terms	31.29	—*
ATM	20 terms	207.3	—*
ATM	30 terms	3542.1	—*

TABLE 1. Execution time of EMD, OCD, and ATM, along with memory consumption of EMD and OCD in estimating Swiss roll distribution.

* The deployed code of ATM (<https://github.com/baptistar/ATM>) is implemented in Matlab, and we were not able to record the memory consumption accurately on our local computer which uses Linux as the operating system.

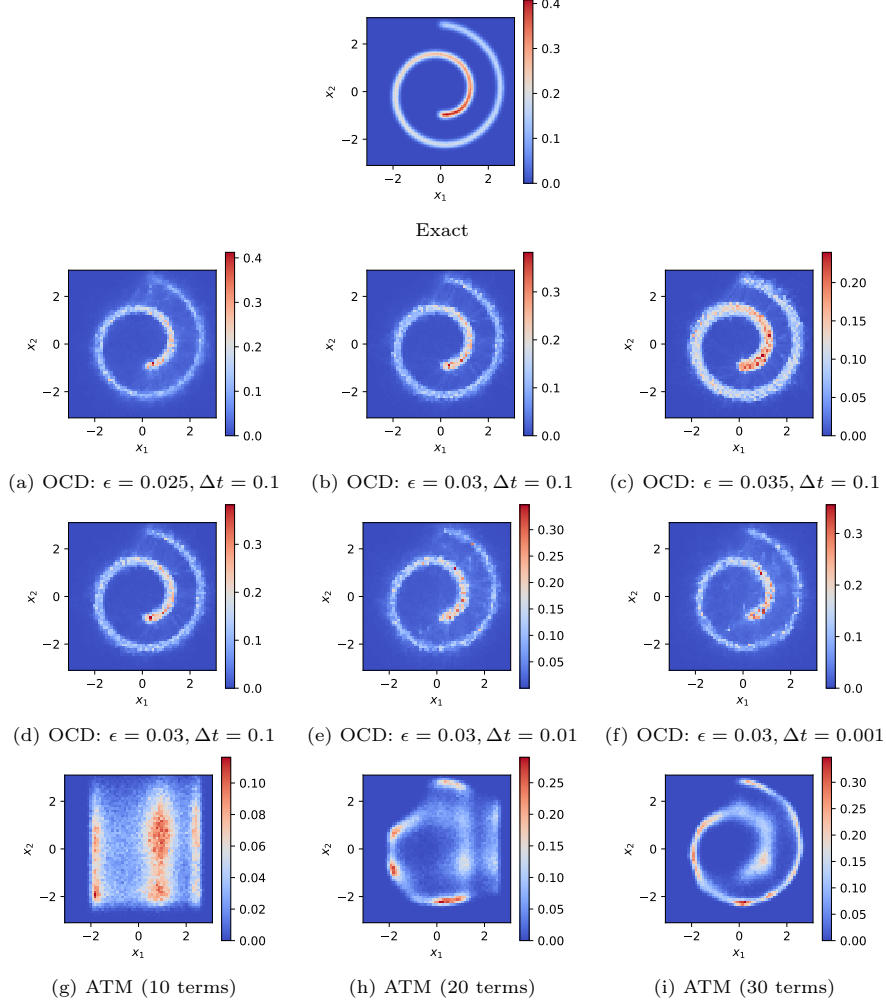


FIGURE 14. Estimating distribution of the Swiss roll as the target density from Gaussian source using L^2 -OCD (a-f) and ATM (g-i).

APPENDIX D. BENCHMARKING OCD AGAINST NEURAL OPTIMAL TRANSPORT

In this section, we further test OCD against another particle method called Neural Optimal Transport [40]. In this method, backpropagation is used to compute the map between marginals. We consider two normal distributions $X, Y \sim N(0_n, I_n)$ in n -dimensional setting. We compare the solution of the discrete samples from each method against EMD and investigate performance in terms of execution time and error. As shown in Fig. 17, for a fixed dimension of $n = 2$, using 200 epochs for NOT and 100 iterations for OCD, the proposed OCD method outperforms NOT for a wide of number of particles N_p , while achieving more accurate solution. However, we expect NOT to reach the performance of OCD at very large $N_p \rightarrow \infty$. In Fig. 18, we further compare NOT to OCD in terms of dimension $n = \dim(X)$. Clearly, the OCD outperforms NOT in low dimensional problems. We expect a turnover in performance in high dimensions.

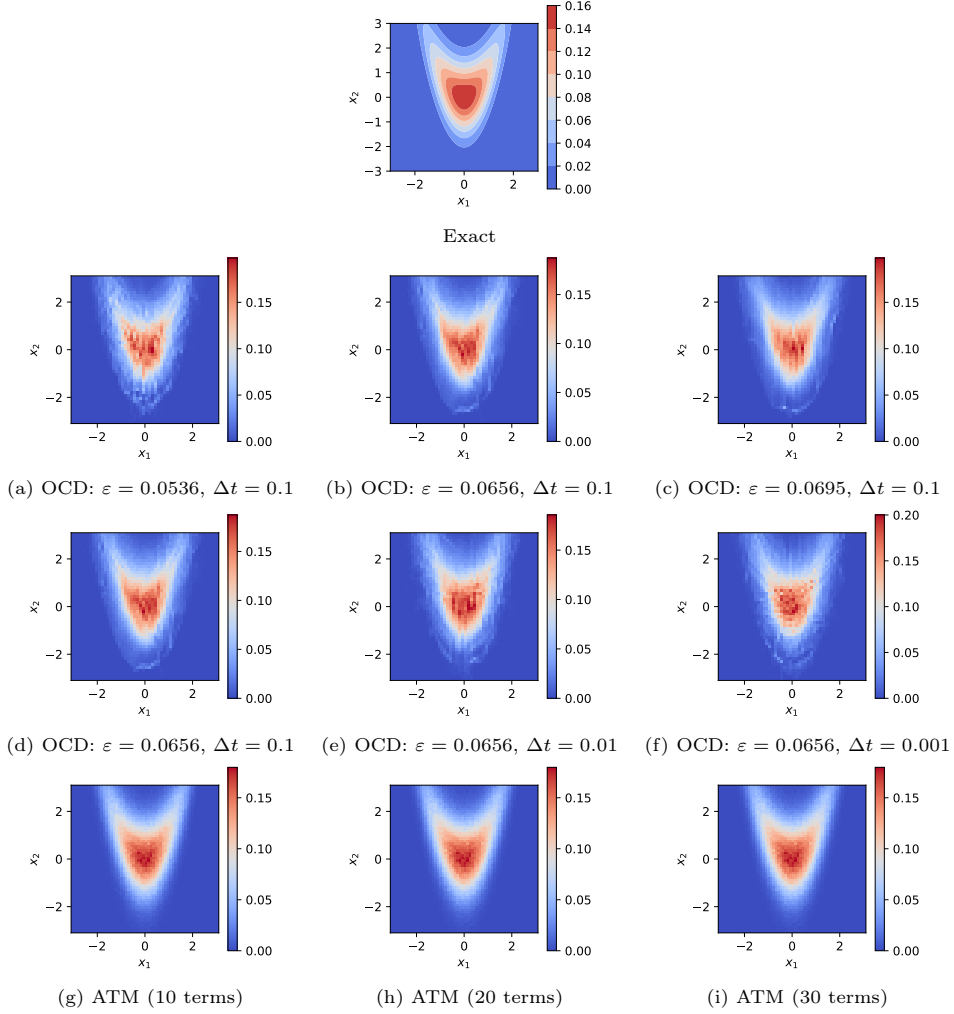


FIGURE 15. Estimating distribution of the banana as the target density from Gaussian source, using L^2 -OCD (a-f) and ATM (g-i).

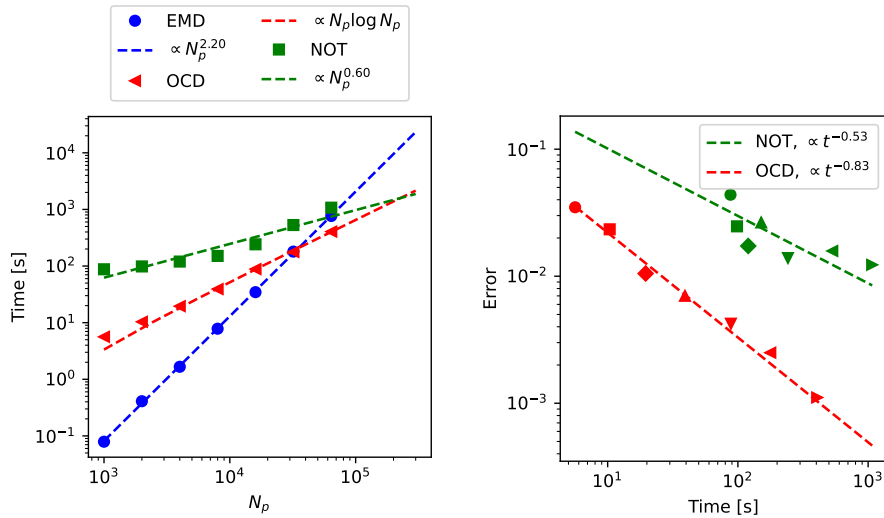


FIGURE 17. Scaling of execution time against number of particles N_p (left) and a comparison of NOT against OCD in terms of L_1 error in the estimation of Wasserstein distance versus computational time (right).

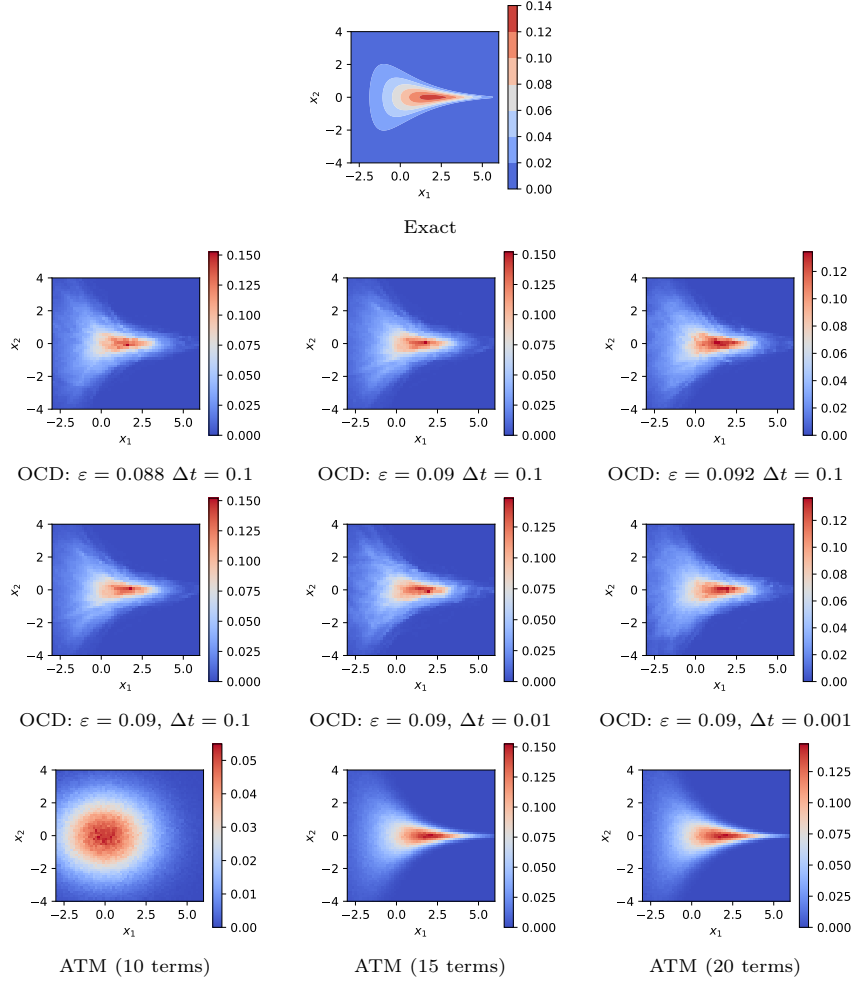


FIGURE 16. Estimating distribution of the funnel as the target density from Gaussian source, using L^2 -OCD (a-f) and ATM (g-i).

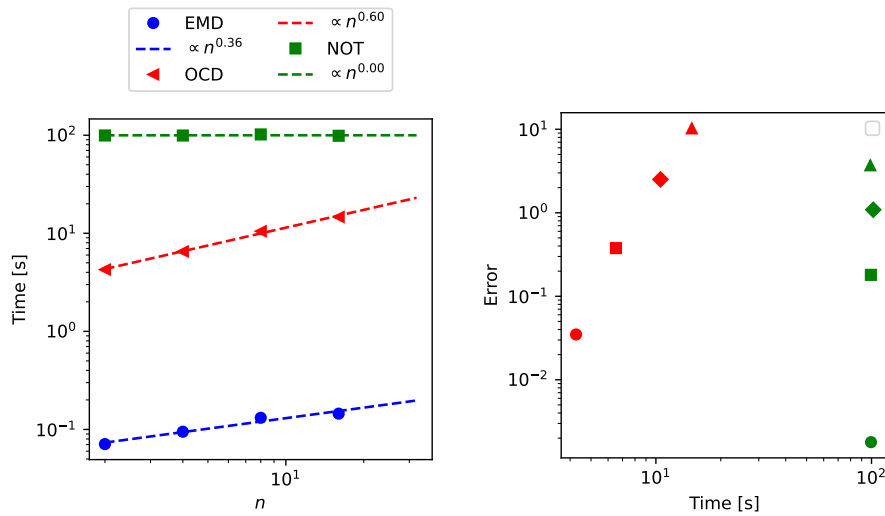


FIGURE 18. Scaling of execution time against dimension $n = \dim(X)$ (left) and a comparison of NOT against OCD in terms of L_1 error in the estimation of Wasserstein distance versus computational time for the corresponding dimensions (right).

APPENDIX E. OCD WITH L^p -COST FUNCTION

Here, we demonstrate the application of the developed L^p -OCD method for convex cost functions other than quadratic one considered in the main text. As an example, let us consider L^p function $c_p(x, y) = \|x - y\|_p^p$ as the cost function for the optimal map problem with value $p = 4$.

As the first example, we consider $n = 1, 2$ and generate 10^4 samples from $X \sim \mathcal{N}(0, I)$ and $Y \sim \mathcal{N}(0, 0.5I)$. As shown in Fig. 19, both of these cases converge to the sub-optimal solution monotonically in time.

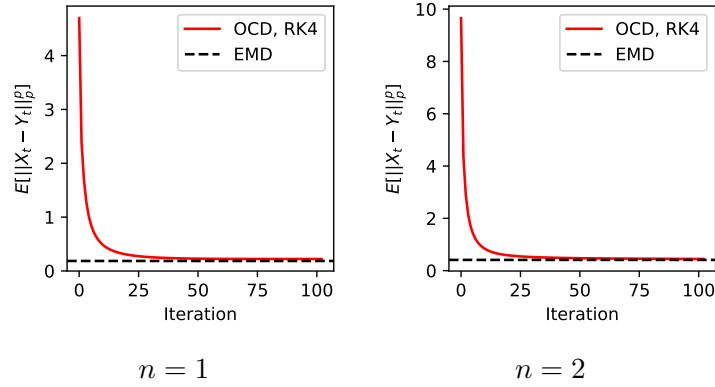


FIGURE 19. Estimate of L^p -Wasserstein distance in the OCD dynamics with $p = 4$ till steady-state between two normally distributed random variables X and Y with $\Delta t = 0.01$.

Second, we consider the optimal map problem between the normal distribution $X \sim N(0, I)$ and its push-forward softmax map $Y = T(X)$, where $T_i = \nabla_{x_i} \log(\exp(x_1) + \exp(x_2))$ in \mathbb{R}^2 . Starting from 10^5 independent random samples of both marginals, we evolve the pair (X_t, Y_t) by RK4-OCD with the linear approximation to the conditional expectation, $\Delta t = 3 \times 10^{-3}$ and $\epsilon = 2 \times 10^{-3}$. As shown in Fig. 20, $\mathbb{E}\|X_t - Y_t\|_p^p$ converges towards the L^p -Wasserstein distance.

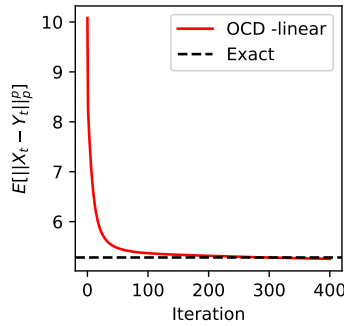


FIGURE 20. Estimate of L^p -Wasserstein distance in the OCD dynamics with $p = 4$ till steady-state between normally distribution X and softmax map Y with $\epsilon = 2 \times 10^{-3}$ and $\Delta t = 3 \times 10^{-3}$.

APPENDIX F. 10-DIMENSIONAL MULTIVARIATE NORMAL DISTRIBUTIONS

As an example in the high-dimensional setting, here we test the proposed OCD in estimating the optimal map between two multivariate normal distributions $\mu = \mathcal{N}(0, \Sigma_\mu)$ and $\nu = \mathcal{N}(0, \Sigma_\nu)$ in \mathbb{R}^n with $n = 10$ where

$$\Sigma_\mu = I + L_\mu^T L_\mu \quad \Sigma_\nu = I + L_\nu^T L_\nu \quad (64)$$

$$(L_\mu)_{i,j} = \begin{cases} \frac{10}{(i+j)^2} & \forall j < i \\ 0 & \text{otherwise} \end{cases} \quad (65)$$

$$(L_\nu)_{i,j} = \begin{cases} \cos(2\pi(i+j)/n) & \forall j < i \\ 0 & \text{otherwise} \end{cases} \quad (66)$$

We generate 10^4 independent samples for μ and ν and estimate the optimal map between them using OCD with $\epsilon = 1.88$ and $\Delta t = 0.05$ against the solution obtained by EMD. As shown in Fig. 21, OCD provides an accurate estimate for the L^2 Wasserstein distance.

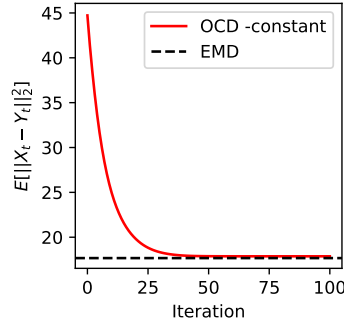


FIGURE 21. Estimate of L^2 -Wasserstein distance in the OCD dynamics with piecewise constant estimator of the conditional expectation till steady-state between multi-variate 10-dimensional normal distributions μ and ν with covariance matrices (64) against the solution obtained with EMD.