ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

OPTIMAL TANSPORT AND APPLICATIONS IN GENERATIVE MODELING

Σταματία - Στυλιανή Δ. Μουζάκη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών στο
πλαίσιο του Προπτυχιακού Προγράμματος
Σπουδών

Αθήνα

Ιούνιος, 2026

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Yannacopoulos, for his guidance, support, and valuable comments throughout the preparation of this thesis.

I am also deeply grateful to my family for their constant encouragement and support during my studies.

Περίληψη

Η Θεωρία Βέλτιστης Μεταφοράς (Optimal Transport) αποτελεί ένα ισχυρό μαθηματικό πλαίσιο για τη σύγκριση κατανομών πιθανότητας, λαμβάνοντας υπόψη τη γεωμετρία του υποκείμενου χώρου. Τα τελευταία χρόνια, η σύνδεσή της με τη μηχανική μάθηση, τα παραγωγικά μοντέλα και τα δυναμικά συστήματα έχει οδηγήσει στην ανάπτυξη μίας ευρείας κατηγορίας μεθόδων παραγωγικής μοντελοποίησης βασισμένων στη μεταφορά κατανομών.

Στην παρούσα διπλωματική εργασία μελετώνται οι μαθηματικές θεμελιώσεις της Βέλτιστης Μεταφοράς και η σύνδεσή τους με σύγχρονες αρχιτεκτονικές generative modeling. Αρχικά παρουσιάζονται οι διατυπώσεις των Monge και Kantorovich, η γεωμετρική δομή των χώρων Wasserstein, το θεώρημα του Brenier, καθώς και η εξίσωση Monge–Ampère. Ιδιαίτερη έμφαση δίνεται στη γεωμετρική ερμηνεία των gradient flows στον χώρο πιθανοτικών μέτρων και στη σχέση τους με δυναμικά συστήματα μεταφοράς.

Στη συνέχεια εξετάζονται σύγχρονες μέθοδοι generative modeling βασισμένες στη γεωμετρία Wasserstein, όπως τα Normalizing Flows, τα Continuous Normalizing Flows, οι Wasserstein-based objectives, οι Sliced-Wasserstein μέθοδοι και το Wasserstein Flow Matching. Παράλληλα, μελετάται η προσέγγιση απεικονίσεων βέλτιστης μεταφοράς μέσω ελεγχόμενων Neural Ordinary Differential Equations και αρχιτεκτονικών βασισμένων σε Lie compositions, αναδεικνύοντας τη γεωμετρική και αναλυτική τους δομή.

Στόχος της εργασίας είναι η ενοποιημένη παρουσίαση της σχέσης μεταξύ Βέλτιστης Μεταφοράς, γεωμετρικής ανάλυσης και συνεχών παραγωγικών μοντέλων. Το προτεινόμενο πλαίσιο αναδεικνύει τον τρόπο με τον οποίο οι δυναμικές μεταφοράς μπορούν να χρησιμοποιηθούν για την κατασκευή ευέλικτων και γεωμετρικά συνεπών generative models σε υψηλές διαστάσεις.

Abstract

Optimal Transport has emerged as a powerful mathematical framework for comparing probability distributions while incorporating the underlying geometry of the sample space. In recent years, its interaction with machine learning, generative modeling, and dynamical systems has led to the development of a broad class of transport-based generative methods, including Wasserstein flows, continuous normalizing flows, and flow matching techniques.

This thesis studies the mathematical foundations and computational aspects of Optimal Transport and investigates their connection with modern generative modeling architectures. We begin by presenting the Monge and Kantorovich formulations of the transport problem, followed by the geometric structure of Wasserstein spaces, Brenier's theorem, and the Monge–Ampère equation. Particular emphasis is placed on the role of Wasserstein geometry in defining gradient flows over probability measures and in constructing transport-driven dynamical systems.

Building on this theoretical framework, we examine several transport-based generative methodologies, including Normalizing Flows, Continuous Normalizing Flows, Wasserstein-based objectives, Sliced-Wasserstein methods, and Wasserstein Flow Matching. The thesis further explores the approximation of optimal transport maps through controlled Neural Ordinary Differential Equations and Lie-composition-based architectures, highlighting their geometric interpretation and approximation capabilities.

The overall objective of this work is to provide a unified mathematical perspective connecting Optimal Transport, geometric analysis, and modern continuous-time generative modeling. Beyond the theoretical development, the presented framework illustrates how transport-based dynamics can be used to construct flexible and geometrically consistent generative models in high-dimensional settings.

Contents

1 Theoretical Foundations of Optimal Transport **1**

1.1 Introduction 1

1.2 The Monge-Kantorovich Problem 1

1.2.1 Monge's Formulation: 1

1.2.2 Kantorovich Relaxation: 3

1.3 Wasserstein Distances 6

1.3.1 Definition and Properties: 6

1.4 Comparison with Classical Divergences: 10

1.5 Geometry of Optimal Transport Maps 12

1.5.1 Brenier's Theorem: 12

1.5.2 The Monge-Ampère Equation: 14

1.6 Entropic Regularization 17

1.6.1 The Sinkhorn Algorithm 17

1.7 Gradient Flows in Wasserstein Space 18

1.7.1 Gradient Flows of Functionals 19

1.7.2 The JKO Scheme 20

1.7.3 Nonparametric Sliced-Wasserstein Flows (SWF) 20

2 Generative Modeling and Optimal Transport **23**

2.1 Introduction 23

2.2 Normalizing Flows 24

2.2.1 Mathematical Foundation and the Change of Variables Formula 25

2.2.2 Flow Architectures and Invertibility Strategies 25

2.2.3 Continuous Normalizing Flows (CNF) 26

2.3 Generative Modeling via Wasserstein-2 Loss 26

2.3.1 The W2 Loss Objective 27

2.3.2 Distribution-Dependent ODEs and Gradient Flows 27

2.3.3 The Role of Kantorovich Potentials in Training . . . 27
2.3.4 Implementation within the Autoencoder Framework . . . 28
2.4 Sliced-Wasserstein Autoencoders (SWAE) . . . 28
2.4.1 SWAE Architecture and Objective Function . . . 29
2.4.2 Comparison with VAE and WAE . . . 30
2.5 Sliced-Wasserstein Flows (SWF) . . . 30
2.5.1 Functional Optimization and Gradient Flows . . . 30
2.5.2 Connection to Stochastic Differential Equations (SDEs) . . . 31
2.6 Wasserstein Generative Learning of Conditional Distributions . . . 31
2.6.1 The Conditional Generator Mechanism . . . 31
2.7 Wasserstein Flow Matching (WFM) . . . 31
2.7.1 Lifting Flow Matching to Families of Distributions . . . 32

3 Approximation of Optimal Transport Maps via Linear-Control Neural ODEs 33
3.1 Introduction . . . 33
3.2 Linear-Control Neural ODEs as Dynamical Systems . . . 34
3.2.1 The Dynamical Formulation of the Problem . . . 34
3.2.2 Existence and Uniqueness of Solutions . . . 37
3.2.3 The Flow of the Dynamical System . . . 39
3.3 Approximation Properties and Lie Algebras . . . 41
3.3.1 Lie Brackets and Geometric Interpretation . . . 42
3.3.2 Lie Algebra of the System . . . 43
3.4 Strong Approximating Property . . . 45
3.4.1 Universal Approximation Theorem . . . 47
3.5 Connection with Optimal Transport Maps . . . 49
3.6 The OT Map as a Flow of a Dynamical System . . . 50
3.6.1 Construction of the Interpolation . . . 51
3.6.2 Construction of a Time-Varying Vector Field . . . 51
3.7 Approximation of the OT Map via Linear-Control Neural ODEs . . . 54
3.8 Conclusion . . . 56

4 Normalizing Flows via Linear-Control Neural ODEs 57
4.1 Introduction . . . 57
4.2 Approximation of Optimal Transport Maps . . . 58
4.2.1 Practical Scenario . . . 58

4.3 Variational Learning Framework 58
4.3.1 Discrete Optimal Transport Plan 58
4.3.2 Objective Functional 58
4.3.3 Data Fidelity Term 59
4.3.4 Regularization Term 59
4.4 Optimization Procedure 59
4.5 Algorithmic Interpretation 61
4.5.1 Geometric Interpretation 61
4.5.2 Generative Modeling Interpretation 61
4.6 Advantages of the Proposed Framework 62
4.6.1 Numerical Implementation and Experimental Evaluation 62
4.7 Conclusion 66

5 Discussion and Conclusion **69**

6 Appendix **71**

List of Figures

4.1 Experiment 1: Source distribution, transported samples, and control trajectories. 64
4.2 Experiment 2: Improved capacity leads to more structured transport. 65
4.3 Experiment 3: Best performing configuration with improved stability. 65

List of Tables

Chapter 1

Theoretical Foundations of Optimal Transport

1.1 Introduction

In this chapter, we present some fundamental general results in the theory of optimal transport that will be essential for the development of the material of this thesis. More specifically, we introduce the Monge and Kantorovich formulations of the transport problem, Wasserstein distances, the geometry of optimal transport maps, and gradient flows in Wasserstein space. Particular emphasis is placed on the quadratic-cost setting, including Brenier's theorem, the Monge–Ampère equation, and entropic regularization techniques such as the Sinkhorn algorithm. The exposition is based primarily on the treatments of Santambrogio (2015) and Figalli–Glaudo (2021).

1.2 The Monge-Kantorovich Problem

1.2.1 Monge's Formulation:

The history of optimal transport begins in 1781 with Gaspard Monge, who posed a practical problem: if sand has to be transported from one place to another (for example, for the construction of fortifications), what is the cheapest way to transport it?

In order to give a rigorous mathematical formulation of the problem, it is necessary to define the transportation cost. In his original formulation, the ambient space was $\mathbb{R}^d$ (with $d = 2$ or 3) and the transportation cost of one unit of mass from point x to y was given by the Euclidean distance:

$$c(x, y) = |x - y|.$$

Let $f, g \geq 0$ be two mass distributions on $\mathbb{R}^d$ such that

$$\int f(x)\,dx = \int g(y)\,dy = 1.$$

Monge's problem is to find a map

$$T : \mathbb{R}^d \to \mathbb{R}^d$$

which transports the distribution f to the distribution g, meaning that it satisfies:

$$\int_A g(y)\,dy = \int_{T^{-1}(A)} f(x)\,dx \quad \text{for every Borel set } A,$$

and at the same time it minimizes the total cost of transport:

$$M(T) = \int_{\mathbb{R}^d} |T(x) - x|^2 f(x)\,dx.$$

Intuitively, consider a set of particles distributed according to the density f. These particles are to be transported so as to produce a new distribution g, in a way that minimizes the average transportation distance.

Viewed through the lens of modern measure theory, if μ is a measure on X and $T : X \to Y$ is a measurable mapping, then the push-forward measure is defined as:

$$(T_{\#}\mu)(A) = \mu(T^{-1}(A)).$$

Thus, the problem can be formulated as:

$$\min \left\{ \int c(x, T(x))\,d\mu(x) \ : \ T_{\#}\mu = \nu \right\}.$$

In the Euclidean case, if the densities are smooth and the map is 1-1, then the mass conservation condition is equivalent to the non-linear equation:

$$g(T(x)) \det(DT(x)) = f(x).$$

This equation is highly nonlinear and creates significant analytical difficulties, particularly regarding the existence and stability of minimizers. To be more precise, the weak convergence of maps does not ensure that the mass transport condition is preserved.

Monge had analyzed some geometrical features of the solutions, like the fact that the "transport rays" do not intersect and that the transport occurs in a monotonic way along them.

Nevertheless, he did not manage to solve the problem of finding a solution.

The main advances occurred after the 20th century with Kantorovich, who introduced a new perspective: instead of using maps, the transport can be described by a transport measure γ on the space $X \times Y$.

1.2.2 Kantorovich Relaxation:

About 150 years after Monge, in the 1940s, Leonid Kantorovich reconsidered the problem of optimal transport from a different perspective. In order to understand his viewpoint, let us consider the following example:

Let us assume that we have N bakeries located at $\{x_i\}_{i=1}^N$ and M coffee shops located at $\{y_j\}_{j=1}^M$. The bakery i produces a quantity $\alpha_i \geq 0$ of bread and the coffee shop j requires a quantity $\beta_j \geq 0$. We assume that the total supply and total demand are normalized so that

$$\sum_i \alpha_i = \sum_j \beta_j = 1.$$

In Monge's formulation, the transportation is deterministic: the mass located at the point x_i is sent to a unique destination $T(x_i)$. Nevertheless, this is not sufficient for the previous example, since a bakery may provide bread to several coffee shops and a coffee shop may buy bread from several bakeries.

For this reason, Kantorovich introduced a new perspective using the concept of **couplings** $\gamma(x, y)$, which describes the amount of mass transported from x to y. In the discrete example, this corresponds to a matrix (γ_{ij}), where:

(a) $\gamma_{ij} \geq 0$, i.e. the amount of bread transported from the bakery x_i to the coffee shop y_j is nonnegative.

(b) $\sum_{j=1}^M \gamma_{ij} = \alpha_i$ for every i, i.e. the total quantity sent from each bakery equals its production.

(c) $\sum_{i=1}^N \gamma_{ij} = \beta_j$ for every j, i.e. the total quantity received by each coffee shop equals its demand.

(d) The transportation cost $\sum_{i,j} \gamma_{ij} c(x_i, y_j)$ is minimized.

Important aspects of Kantorovich's approach are:

- The coupling γ allows **mass splitting**, something that Monge's formulation with the single valued map T does not allow.

- Constraint (a) is convex, constraints (b) and (c) are linear, and the objective function (d) is also linear. Hence, Kantorovich's problem is transformed into a **linear programming problem with convex/linear constraints**.

- The **Kantorovich Duality** is fundamental for computations, since it allows the characterization and verification of optimal solutions through dual potentials.

In this way, Kantorovich's formulation not only makes the problem feasible in cases where Monge's map fails, but also connects it with linear programming, allowing the application of optimization methods and computational algorithms for optimal transport.

Transport Plans in the General Setting. Let (X, μ) and (Y, ν) be probability spaces. A **transport plan** (or coupling) between μ and ν is a probability measure γ on the product space $X \times Y$ such that

$$(\pi_X)_{\#}\gamma = \mu, \qquad (\pi_Y)_{\#}\gamma = \nu,$$

where $\pi_X(x, y) = x$ and $\pi_Y(x, y) = y$ denote the canonical projections.

The set of all such transport plans is denoted by

$$\Pi(\mu, \nu).$$

The Kantorovich problem in the general setting is then formulated as

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\gamma(x, y).$$

Kantorovich Duality. A cornerstone of optimal transport theory is the **Kantorovich Duality Theorem**, which asserts that, under suitable conditions on the cost function c, the primal minimization problem admits the dual formulation

$$\sup_{\varphi,\psi} \left\{ \int_X \varphi\, d\mu + \int_Y \psi\, d\nu \ : \ \varphi(x) + \psi(y) \leq c(x, y) \right\}.$$

The functions φ and ψ are called *Kantorovich potentials*, and this dual representation is fundamental both for theoretical analysis and computational methods in optimal transport.

Assumptions for Kantorovich Duality. To state the theorem rigorously, we assume that:

- The cost function $c : X \times Y \to \mathbb{R}$ is **lower semicontinuous** and **bounded below**.

- The spaces X and Y are **Polish spaces** (complete separable metric spaces).
- The measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ are probability measures.

Under these assumptions, the Kantorovich optimal transport problem

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\gamma(x,y)$$

admits the dual formulation

$$\sup_{\phi\in L^1(\mu),\, \psi\in L^1(\nu)} \left\{ \int_X \phi\, d\mu + \int_Y \psi\, d\nu \;:\; \phi(x)+\psi(y) \le c(x,y),\ \forall x\in X, y\in Y \right\},$$

which makes the duality statement precise and mathematically rigorous.

Weak Kantorovich Duality. Before presenting the full duality theorem, it is instructive to establish the *weak duality inequality*, which shows that every admissible pair of potentials provides a lower bound on the optimal transport cost.

Theorem 1.2.1. *Weak Duality: Let μ and ν be probability measures on X and let $c : X \times Y \to \mathbb{R}$ be a measurable cost function. Then*

$$\sup_{\varphi,\psi} \left\{ \int_X \varphi\, d\mu + \int_Y \psi\, d\nu \;:\; \varphi(x)+\psi(y) \le c(x,y) \right\} \le \inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\gamma(x,y).$$

We now proceed to the proof of the weak duality inequality, showing that every admissible pair (ϕ, ψ) provides a lower bound for the optimal transport cost.

Let (φ, ψ) be any pair of measurable functions satisfying the constraint

$$\varphi(x) + \psi(y) \le c(x,y) \quad \text{for all } (x,y) \in X \times Y.$$

Let $\gamma \in \Pi(\mu,\nu)$ be any transport plan between μ and ν. Integrating the above inequality with respect to γ gives

$$\int_{X\times Y} \big(\varphi(x)+\psi(y)\big)\, d\gamma(x,y) \le \int_{X\times Y} c(x,y)\, d\gamma(x,y).$$

Using the marginal conditions of γ, we obtain

$$\int_{X\times Y} \varphi(x)\, d\gamma(x,y) = \int_X \varphi(x)\, d\mu(x),$$

and

$$\int_{X\times Y} \psi(y)\, d\gamma(x,y) = \int_Y \psi(y)\, d\nu(y).$$

Therefore,

$$\int_X \varphi\, d\mu + \int_Y \psi\, d\nu \le \int_{X\times Y} c(x,y)\, d\gamma(x,y).$$

Since this inequality holds for every transport plan $\gamma \in \Pi(\mu,\nu)$, we obtain

$$\int_X \varphi\, d\mu + \int_Y \psi\, d\nu \le \inf_{\gamma\in\Pi(\mu,\nu)} \int c(x,y)\, d\gamma(x,y).$$

Finally, taking the supremum over all admissible pairs (φ,ψ) yields

$$\sup_{\varphi,\psi} \left(\int_X \varphi\, d\mu + \int_Y \psi\, d\nu \right) \le \inf_{\gamma\in\Pi(\mu,\nu)} \int c(x,y)\, d\gamma(x,y).$$

□

1.3 Wasserstein Distances

1.3.1 Definition and Properties:

We now introduce a family of distances on the space of probability measures, induced by optimal transport, known as the *p-Wasserstein distances*.

Let (X,d) be a locally compact and separable metric space and let $1 \le p < \infty$. We first define the space of probability measures with finite p-moment:

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \ : \ \int_X d(x,x_0)^p\, d\mu(x) < \infty \text{ for some (and hence any) } x_0 \in X \right\}.$$

The definition is independent of the choice of the base point x_0 due to the triangle inequality.

Definition 1.3.1. p-Wasserstein Distance: Given $\mu,\nu \in \mathcal{P}_p(X)$, their p-Wasserstein distance is defined by

$$W_p(\mu,\nu) := \left(\inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times X} d(x,y)^p\, d\gamma(x,y) \right)^{\frac{1}{p}},$$

where $\Pi(\mu,\nu)$ denotes the set of transport plans (couplings) between μ and ν, i.e.,

$$\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(X \times X) : (\pi_1)_{\#}\gamma = \mu,\ (\pi_2)_{\#}\gamma = \nu\}.$$

Important Special Cases.

- The distance W_1 is called the *1-Wasserstein distance* or *Earth Mover's Distance*. It is particularly important due to its dual formulation:

$$W_1(\mu,\nu) = \sup_{\|\varphi\|_{\mathrm{Lip}} \le 1} \left\{ \int_X \varphi \, d\mu - \int_X \varphi \, d\nu \right\}.$$

- The distance W_2 plays a central role in analysis and geometry. In Euclidean spaces, it is closely connected to quadratic cost optimal transport, gradient flows, and Riemannian-type structures on the space of measures.

Metric Properties. The following fundamental theorem justifies the terminology "distance" for the Wasserstein functional.

Theorem 1.3.1. *Let (X,d) be a separable metric space and let $1 \le p < \infty$. Then the function*

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X\times X} d(x,y)^p \, d\gamma(x,y) \right)^{1/p}$$

defines a metric on $\mathcal{P}_p(X)$.

We now verify that W_p satisfies the four defining properties of a metric.

Proof.

1. Non-negativity.

Since $d(x,y)^p \ge 0$ and γ is a probability measure,

$$\int d(x,y)^p \, d\gamma(x,y) \ge 0.$$

Taking the infimum and the p-th root yields

$$W_p(\mu,\nu) \ge 0.$$

2. Identity of indiscernibles.

Assume first that $\mu = \nu$. Then the transport plan

$$\gamma = (\mathrm{Id}, \mathrm{Id})_{\#}\mu$$

is concentrated on the diagonal $\{x = y\}$ and satisfies

$$\int d(x, y)^p \, d\gamma = 0.$$

Hence $W_p(\mu, \mu) = 0$.

Conversely, assume $W_p(\mu, \nu) = 0$. Then there exists a sequence $\gamma_n \in \Pi(\mu, \nu)$ such that

$$\int d(x, y)^p \, d\gamma_n(x, y) \to 0.$$

Since the cost vanishes, the mass of γ_n concentrates near the diagonal. By tightness and weak compactness of probability measures on Polish spaces, we may extract a weakly convergent subsequence $\gamma_n \rightharpoonup \gamma$.

By lower semicontinuity of the integral functional,

$$\int d(x, y)^p \, d\gamma = 0.$$

Thus $d(x, y) = 0$ γ-almost everywhere, meaning that γ is supported on the diagonal. Therefore,

$$\mu = (\pi_1)_{\#}\gamma = (\pi_2)_{\#}\gamma = \nu.$$

3. Symmetry.

Let $\gamma \in \Pi(\mu, \nu)$. Define the map $S(x, y) = (y, x)$ and set

$$\tilde{\gamma} := S_{\#}\gamma.$$

Then $\tilde{\gamma} \in \Pi(\nu, \mu)$ and

$$\int d(x, y)^p \, d\tilde{\gamma}(x, y) = \int d(y, x)^p \, d\gamma(x, y) = \int d(x, y)^p \, d\gamma(x, y),$$

since d is symmetric.

Taking infimum over all γ gives

$$W_p(\mu, \nu) = W_p(\nu, \mu).$$

4. Triangle inequality.

Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(X)$.

Let $\gamma_{12} \in \Pi(\mu_1, \mu_2)$ and $\gamma_{23} \in \Pi(\mu_2, \mu_3)$ be optimal (or ε-optimal) transport plans.

Gluing Lemma. There exists a probability measure Γ on $X \times X \times X$ such that:

- the (x_1, x_2)-marginal equals γ_{12},

- the (x_2, x_3)-marginal equals γ_{23}.

Define γ_{13} as the (x_1, x_3)-marginal of Γ. Then $\gamma_{13} \in \Pi(\mu_1, \mu_3)$.

By the triangle inequality for d,

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3).$$

Raising to the power p and integrating with respect to Γ,

$$\int d(x_1, x_3)^p \, d\Gamma \leq \int (d(x_1, x_2) + d(x_2, x_3))^p \, d\Gamma.$$

Applying Minkowski's inequality in $L^p(\Gamma)$ yields

$$\left(\int d(x_1, x_3)^p \, d\Gamma \right)^{1/p} \leq \left(\int d(x_1, x_2)^p \, d\Gamma \right)^{1/p} + \left(\int d(x_2, x_3)^p \, d\Gamma \right)^{1/p}.$$

Since the marginals of Γ recover γ_{12} and γ_{23},

$$W_p(\mu_1, \mu_3) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3).$$

Thus W_p satisfies the triangle inequality.

Having verified all four properties, we conclude that W_p is a metric on $\mathcal{P}_p(X)$. □

We close this section with the following comments concerning the metric properties of the Wasserstein distance.

- *Non-negativity and identity of indiscernibles:* Clearly $W_p(\mu, \nu) \geq 0$. If $W_p(\mu, \nu) = 0$, then there exists a transport plan concentrated on the diagonal $\{x = y\}$, which implies $\mu = \nu$.

- *Symmetry:* If γ is a transport plan between μ and ν, then the push-forward under the map $(x, y) \mapsto (y, x)$ is a transport plan between ν and μ. Since $d(x, y) = d(y, x)$, it follows that

$$W_p(\mu, \nu) = W_p(\nu, \mu).$$

- *Triangle inequality:* Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(X)$ and let γ_{12} and γ_{23} be optimal transport plans between (μ_1, μ_2) and (μ_2, μ_3) respectively.

  Using the gluing lemma (or disintegration theorem), one constructs a measure on $X \times X \times X$ whose projections recover γ_{12} and γ_{23}. Integrating out the middle variable produces a transport plan between μ_1 and μ_3.

  Applying the triangle inequality in L^p yields

$$W_p(\mu_1, \mu_3) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3).$$

Topology Induced by W_p. Since W_p is a metric, it induces a topology on $\mathcal{P}_p(X)$. An important result is that convergence in W_p is equivalent to weak-* convergence together with convergence of the p-th moments:

$$W_p(\mu_n, \mu) \to 0 \quad \Longleftrightarrow \quad \begin{cases} \mu_n \overset{*}{\rightharpoonup} \mu, \\ \displaystyle\int_X d(x_0, x)^p \, d\mu_n(x) \to \int_X d(x_0, x)^p \, d\mu(x). \end{cases}$$

In particular, if X is compact, then convergence in W_p is equivalent to weak-* convergence.

Thus, the Wasserstein distances provide not only a metric structure on spaces of probability measures, but also a powerful analytical framework linking optimal transport, functional analysis, and geometric measure theory.

1.4 Comparison with Classical Divergences:

In many applications, probability distributions are compared using various discrepancies or divergences, such as the Kullback{Leibler divergence (KL) or the total variation (TV) distance. The p-Wasserstein distances W_p defined above provide an alternative that captures the underlying geometry of the space on which the measures live, and in several respects behaves differently (and often more smoothly) than classical divergences.

Classical Divergences.

- The *Kullback{Leibler (KL) divergence* between two probability measures μ and ν (when μ is absolutely continuous with respect to ν) is defined by

$$D_{\mathrm{KL}}(\mu \,\|\, \nu) = \int_X \log\left(\frac{d\mu}{d\nu}\right) d\mu,$$

and satisfies $D_{\mathrm{KL}}(\mu\|\nu) \geq 0$ with equality iff $\mu = \nu$. However, D_{KL} is not symmetric, it does not satisfy the triangle inequality, and it is not a proper metric. In particular, if μ and ν have disjoint supports, the KL divergence is infinite or undefined.

- The *total variation (TV) distance* between μ and ν is given by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A} |\mu(A) - \nu(A)|,$$

which in discrete cases reduces to half the L^1 norm of the density difference. The TV distance is a metric, symmetric and satisfies the triangle inequality, but it remains insensitive to geometry: it only measures the largest pointwise discrepancy between probabilities, ignoring distances between points in the underlying space. Moreover, by Pinsker's inequality we have

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mu\|\nu)},$$

showing a connection between TV and KL divergences. :contentReference[oaicite:1]index=1

Wasserstein vs KL and TV. The Wasserstein distance W_p is a genuine metric on $\mathcal{P}_p(X)$, symmetric and satisfying the triangle inequality. Unlike KL divergence, it remains finite even for measures with disjoint support, as long as p-moment conditions hold. This is because W_p measures the *cost of transporting mass* from one distribution to another, using the ground metric d; it therefore encodes geometric information that classical divergences ignore. :contentReference[oaicite:2]index=2

To illustrate this difference, consider two Dirac measures in $\mathbb{R}$,

$$\mu_n = \delta_0, \qquad \nu_n = \delta_{1/n},$$

where δ_x is the probability measure concentrated at x. For these measures:

- The KL divergence $D_{\mathrm{KL}}(\mu_n\|\nu_n)$ is infinite for all n, because the support of ν_n does not contain the support of μ_n, so the Radon{Nikodym derivative does not exist.

- Similarly, $D_{\mathrm{KL}}(\nu_n\|\mu_n)$ is also infinite for the same reason | one of the measures is singular with respect to the other.

- The total variation distance between μ_n and ν_n is maximal ($\|\mu_n - \nu_n\|_{\mathrm{TV}} = 1$ for all n), since they do not overlap.

- The p-Wasserstein distance, on the other hand, is simply the distance between the support

points,

$$W_p(\mu_n, \nu_n) = |0 - 1/n| = 1/n,$$

which tends to 0 linearly as $n \to \infty$. This shows that W_p sees the two distributions as *close* in a geometric sense, even though classical divergences view them as infinitely far apart. The behaviour of W_p gives a notion of *smoothness* as distributions drift in space.

Implications for Optimization. Because W_p is sensitive to the underlying metric d, and finite even when supports separate, it often provides a smoother loss landscape in optimization problems compared to KL or TV. In particular, if one is trying to adjust a model by gradient descent to minimize W_p, small changes in the support of the predicted distribution lead to small changes in W_p, whereas the KL divergence may remain infinite or exhibit discontinuous behaviour. This geometric sensitivity gives the Wasserstein distance advantages in applications like generative modelling and variational inference, where distributions evolve continuously in parameter space and supports may be disjoint.

1.5 Geometry of Optimal Transport Maps

1.5.1 Brenier's Theorem:

The deepest and most fundamental mathematical result in the geometry of optimal transport maps is the so-called *Brenier Theorem*.

This theorem provides a complete characterization of the form of the optimal transport map in the case of quadratic cost.

Theorem 1.5.1. *Brenier: Let μ and ν be two probability measures on $\mathbb{R}^d$ with finite second moments, and assume that μ is absolutely continuous with respect to the Lebesgue measure.*

Then, for the cost

$$c(x, y) = \frac{1}{2}|x - y|^2,$$

there exists a unique optimal transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ (unique up to sets of μ-measure zero) that pushes μ to ν. Moreover, T is almost everywhere differentiable, as it is the gradient of a convex function, i.e. there exists a convex potential $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that

$$T(x) = \nabla\varphi(x).$$

Theorem 1.5.2. *Alexandrov: Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then φ is twice differentiable*

$\mathcal{L}^d$-almost everywhere, i.e. for a.e. $x \in \mathbb{R}^d$ there exists a symmetric matrix $D^2\varphi(x)$ such that

$$\varphi(x+h) = \varphi(x) + \nabla\varphi(x) \cdot h + \frac{1}{2} h^\top D^2\varphi(x)\, h + o(|h|^2), \quad h \to 0.$$

In particular, $\nabla\varphi$ is defined $\mathcal{L}^d$-almost everywhere.

Remark 1. The uniqueness of T is not unconditional. It relies crucially on the assumption that μ is absolutely continuous with respect to the Lebesgue measure ($\mu \ll \mathcal{L}^d$). More precisely:

- **Uniqueness μ-a.e.:** The map $T = \nabla\varphi$ is unique μ-almost everywhere. Two optimal transport maps may differ only on a set of μ-measure zero.
- **Role of absolute continuity:** By Theorem 1.5.2, $\nabla\varphi$ is defined $\mathcal{L}^d$-a.e., and hence μ-a.e. whenever $\mu \ll \mathcal{L}^d$. If two convex functions φ_1 and φ_2 both yield optimal transport maps, then $\nabla\varphi_1 = \nabla\varphi_2$ holds μ-almost everywhere.
- **Failure without absolute continuity:** If μ is not absolutely continuous (e.g. a discrete measure), uniqueness may fail entirely, as multiple optimal transport maps can exist.

The 1.5.1 is significant not only as it ensures existence, but also because it gives an explicit expression for the transport map, directly linked to convex analysis and the Legendre transformation.

Connection with the convex analysis. Brenier's theory is based on the ideal combination of optimal transport and convex analysis.

The convex analysis focuses on functions φ satisfying

$$\varphi(tx + (1-t)y) \leq t\,\varphi(x) + (1-t)\,\varphi(y)$$

for every $x, y \in \mathbb{R}^d$ and $t \in [0,1]$ and has highly robust properties, such as gradient's monotonicity and the existence of subgradients.

The results of that theorem, indicate that:

- With the quadratic cost, the optimal transport maps are *cyclically monotone*.
- Every cyclically monotone map corresponds to a subgradient of a convex function φ, which can be chosen to be convex and lower semi-continuous.
- Thus, there is a convex φ in such a way that the optimal map is $T = \nabla\varphi$.

This connection with convex analysis is the modern foundation of optimal transport theory.

Connection with the Legendre transform. The Legendre transform (or convex conjugate) is a fundamental tool in convex analysis: For a convex function φ, its conjugate is

$$\varphi^*(y) = \sup_{x \in \mathbb{R}^d} \big\{ \langle x, y \rangle - \varphi(x) \big\}.$$

In the case of Brenier's theorem, the Kantorovich potentials that appear in the dual formulation of the problem are closely related to the Legendre transform of φ. Indeed, the dual candidates can be expressed as transforms of the form

$$\psi(y) = \varphi^*(y) - \frac{1}{2}|y|^2,$$

where φ^* is the Legendre transformation of φ. This function ensures that the complementarity conditions are satisfied and that φ and its conjugate φ^* are in dual relation according to the structure of the quadratic cost.

This theory shows that optimal transport with quadratic cost is intimately related το the fundamental tools of the convex analysis and the Legendre{Fenchel transformations.

Brenier's Theorem represents the most important mathematical result in the field of optimal transport for the quadratic cost:

- It expresses the optimal transport map as the gradient of a convex function.
- It connects optimal transport with convex analysis and geometry.
- It introduces a natural role for the Legendre transform in understanding the duality.

This geometric and analytical structure is fundamental for understanding the metric and PDE aspects of the theory.

1.5.2 The Monge-Ampére Equation:

Let two probability measures

$$\mu = f(x)\,dx \quad \text{on } \Omega \subset \mathbb{R}^d, \qquad \nu = g(y)\,dy \quad \text{on } \Omega' \subset \mathbb{R}^d,$$

where the densities f, g, are nonnegative and integrable, with

$$\int_\Omega f(x)\,dx = \int_{\Omega'} g(y)\,dy = 1.$$

In the case of quadratic cost,

$$c(x, y) = \frac{1}{2}|x - y|^2,$$

by 1.5.1, the optimal transport map has the form:

$$T(x) = \nabla\varphi(x),$$

where φ is a convex function.

The fundamental constraint of the transport problem is mass conservation:

$$T_{\#}\mu = \nu.$$

this means that for every measurable set $B \subset \Omega'$ holds

$$\nu(B) = \mu(T^{-1}(B)).$$

If T is C^1 and 1-1 (which is guaranteed by the strict convexity of φ), then we can apply the change of variables formula. For every $B \subset \Omega'$, we have

$$\int_B g(y)\, dy = \int_{T^{-1}(B)} f(x)\, dx.$$

By making the change of variable $y = T(x)$, we have

$$\int_B g(y)\, dy = \int_B f(T^{-1}(y))\, \frac{1}{\det(DT(T^{-1}(y)))}\, dy.$$

Therefore, the densities must satisfy the following condition pointwise (almost everywhere):

$$g(T(x)) \det(DT(x)) = f(x).$$

This is the Jacobian equation, which precisely expresses mass conservation.

Monge–Ampère Equation. In the special case

$$T(x) = \nabla\varphi(x),$$

we have

$$DT(x) = D^2\varphi(x),$$

That is, the Hessian matrix of φ. Therefore, the above relation becomes

$$g(\nabla\varphi(x))\det(D^2\varphi(x)) = f(x),$$

or equivalent

$$\boxed{\det(D^2\varphi(x)) = \frac{f(x)}{g(\nabla\varphi(x))}}$$

which is the *Monge-Ampère* equation.

It is a fully nonlinear elliptic partial differential equation, because:

- the nonlinearity is already present in the term $\det(D^2\varphi)$,
- but also in the right-hand side via composition

The ellipticity follows from the fact that, for a convex φ the matrix $D^2\varphi$ is positive semidefinite and the gradient exhibits monotonicity:

$$0 \leq A \leq B \;\Rightarrow\; \det A \leq \det B.$$

Geometric interpretation. The Monge–Ampere equation describes the fact that the curvature of the potential function φ defines the transport of probability density.

Indeed:

- The Hessian $D^2\varphi(x)$ counts the local curvature of φ.
- The $\det(D^2\varphi(x))$ is the local volume change factor (Jacobian).

So:

- If $\det(D^2\varphi(x)) > 1$, then the transformation locally *expands* volumes around.
- If $\det(D^2\varphi(x)) < 1$, then the transformation *contracts* volumes.

The equation

$$\det(D^2\varphi(x)) = \frac{f(x)}{g(\nabla\varphi(x))}$$

indicates that the local curvature of φ is precisely adjusted so that the mass leaving x with density $f(x)$ is redistributed to the point $\nabla\varphi(x)$ with density g.

In other words:

The curvature of φ encodes the geometric deformation required to transform the distribution f into the distribution g.

The Monge–Ampère equation follows directly from the requirement of mass conservation via the change of variables formula. In the quadratic cost case, the optimal map is the convex function gradient and the mass conservation translates into a fully nonlinear elliptic PDE.

Thus, the optimal transport problem acquires a profound geometric character: the form of the potential function φ fully determines how space is deformed to transport probability from one distribution to another.

1.6 Entropic Regularization

According to Wang (2020), direct numerical resolution of fully non-linear partial differential equations, such as the Monge-Ampère equation, becomes computationally prohibitive in high-dimensional spaces due to the curse of dimensionality. To enable the application of optimal transport to large-scale data (such as high-dimensional point clouds), it is necessary to relax the problem through entropic regularization, which transforms it into a strictly convex and computationally efficient optimization task.

As noted by Cuturi (2013) and Santambrogio (2015), the standard Optimal Transport problem is often computationally demanding, especially when dealing with large-scale data or high-dimensional point clouds. To mitigate this, **entropic regularization** is introduced by adding an entropy penalty term to the original Kantorovich objective:

$$W_{2,\epsilon}^2(\mu,\nu) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^2 d\gamma(x,y) + \epsilon H(\gamma|\mu\otimes\nu)$$

where $H(\gamma|\mu\otimes\nu)$ is the relative entropy (Kullback-Leibler divergence) of the transport plan γ with respect to the product measure $\mu\otimes\nu$. The regularization parameter $\epsilon > 0$ controls the trade-off between the transport cost and the smoothness of the plan. As $\epsilon \to 0$, the solution converges to the standard Wasserstein distance, while for larger ϵ, the transport plan becomes more diffused ("blurry"), which effectively prevents over-fitting to individual noise particles.

1.6.1 The Sinkhorn Algorithm

According to Cuturi (2013), the primary advantage of entropic regularization is that it transforms the constrained linear programming problem into a strictly convex optimization problem that can be solved efficiently using the **Sinkhorn algorithm**.

As described in Wang (2020), the Sinkhorn algorithm (also known as the Iterative Proportional Fitting Procedure) exploits the fact that the optimal regularized transport plan has a specific scaling form:

$$\gamma_{ij} = u_i K_{ij} v_j$$

where $K_{ij} = \exp(-|x_i - y_j|^2/\epsilon)$ is the Gibbs kernel. The algorithm proceeds by iteratively updating the scaling vectors u and v:

1. $u^{(t+1)} = \frac{\mu}{Kv^{(t)}}$
2. $v^{(t+1)} = \frac{\nu}{K^T u^{(t+1)}}$

As shown by Cuturi (2013), this iterative scaling process converges rapidly and is highly parallelizable, making it suitable for GPU implementation. In the context of generative models like **Wasserstein Flow Matching (WFM)**, the Sinkhorn algorithm is essential for estimating the transport paths between complex point clouds in high-dimensional spaces where exact OT computations would be prohibitive, as discussed in Haviv et al. (2024).

Algorithm 1 Sinkhorn Algorithm for Entropic Optimal Transport

Input: Cost matrix $M \in \mathbb{R}^{n \times m}$ (where $M_{ij} = \|x_i - y_j\|^2$), marginals μ, ν, regularization ϵ, iterations T.
Initialize: Gibbs kernel $K = \exp(-M/\epsilon)$, $v^{(0)} = \mathbf{1}_m$.
for $t = 0$ to $T - 1$ **do**
    $u^{(t+1)} = \mu \,/\, (Kv^{(t)})$ ▷ Element-wise division
    $v^{(t+1)} = \nu \,/\, (K^T u^{(t+1)})$
end for
Output: Transport plan $\gamma = \mathrm{diag}(u) K \mathrm{diag}(v)$ and cost $W_{2,\epsilon}^2 = \langle \gamma, M \rangle$.

1.7 Gradient Flows in Wasserstein Space

In the previous sections, we introduced the optimal transport problem and the Wasserstein distance. One of the most important developments in optimal transport theory is the discovery that the Wasserstein space admits a rich geometric structure. In particular, it is possible to define gradient flows of functionals on the space of probability measures.

This point of view, developed by Jordan- Kinderlehrer-Otto (P.W.Y. Lee 2015), allows us to interpret several classical partial differential equations as gradient flows in Wasserstein space.

The Wasserstein Space Let $\Omega \subset \mathbb{R}^d$ and denote by $\mathcal{P}_2(\Omega)$ the space of probability measures with finite second moment:

$$\int_\Omega |x|^2 d\mu(x) < \infty.$$

Equipped with the 2-Wasserstein distance, the pair $(\mathcal{P}_2(\Omega), W_2)$ forms a metric space, often called the *Wasserstein space*. This space can be seen as a geometric manifold of probability measures where the Wasserstein distance plays the role of a Riemannian metric.

1.7.1 Gradient Flows of Functionals

Let $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$ be a functional defined on the Wasserstein space. The gradient flow associated with F is formally defined as:

$$\partial_t \mu_t = -\nabla_{W_2} F(\mu_t),$$

where $\nabla_{W_2} F(\mu)$ denotes the **Wasserstein gradient** of F at μ, defined as the element of the tangent space $T_\mu \mathcal{P}_2(\Omega)$ satisfying:

$$\lim_{\epsilon \to 0} \frac{F(\mu^\epsilon) - F(\mu)}{\epsilon} = \langle \nabla_{W_2} F(\mu), v \rangle_{L^2(\mu)},$$

for all perturbations $\mu^\epsilon = (\mathrm{id} + \epsilon v)_\# \mu$, where v is a vector field and $(\mathrm{id} + \epsilon v)_\# \mu$ denotes the pushforward of μ along the map $x \mapsto x + \epsilon v(x)$. More precisely, the Wasserstein gradient is related to the first variation $\frac{\delta F}{\delta \mu}$ of F via:

$$\nabla_{W_2} F(\mu) = -\nabla \frac{\delta F}{\delta \mu},$$

where $\frac{\delta F}{\delta \mu}$ is the functional derivative of F with respect to μ, as discussed in Santambrogio (2015).

This equation describes the steepest descent evolution of the functional F with respect to the Wasserstein metric. Although the Wasserstein space is not a classical finite-dimensional manifold, it is still possible to define this notion rigorously using tools from metric geometry, as developed by Jordan, Kinderlehrer, and Otto (1998) and formalized in Lee (2012).

Continuity Equation Formulation If the measure μ_t admits a density ρ_t with respect to the Lebesgue measure, then the gradient flow can be written as a continuity equation:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0,$$

where v_t is a velocity field associated with the functional derivative of F. More precisely:

$$v_t = -\nabla \frac{\delta F}{\delta \rho}(\rho_t).$$

To understand this formulation, we recall that the **continuity equation** expresses conservation of mass: the density ρ_t is transported by the velocity field v_t, meaning that mass is neither created nor destroyed. The velocity field v_t is not chosen arbitrarily | it is determined by the functional F through its first variation $\frac{\delta F}{\delta \rho}$, which measures how F changes under infinitesimal perturbations of ρ_t. Intuitively, the density ρ_t moves in the direction that decreases F most steeply, at each point $x \in \Omega$ following the negative gradient of $\frac{\delta F}{\delta \rho}$.

Example 1.7.1. A classical instance is the **Fokker–Planck equation**, which arises as the gradient flow of the free energy functional:

$$F(\rho) = \int_\Omega \rho \log \rho \, dx + \int_\Omega V \rho \, dx,$$

where $V : \Omega \to \mathbb{R}$ is a potential. Its first variation is $\frac{\delta F}{\delta \rho} = \log \rho + 1 + V$, giving the velocity field $v_t = -\nabla(\log \rho_t + V)$, and the continuity equation becomes:

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla(\log \rho_t + V)) = \Delta \rho_t + \nabla \cdot (\rho_t \nabla V),$$

which is precisely the Fokker–Planck equation, as shown in Jordan, Kinderlehrer and Otto (1998).

This interpretation connects optimal transport theory with the study of partial differential equations.

1.7.2 The JKO Scheme

In this context, a gradient flow describes the trajectory of a distribution μ_t that evolves toward the minimum of a cost functional $F(\mu)$.

A fundamental result in this area is the time-discretization method introduced by Jordan, Kinderlehrer, and Otto. Given a time step $\tau > 0$, the sequence of measures $\{\mu_k\}$ is defined recursively by:

$$\mu_{k+1} = \arg \min_{\mu \in \mathcal{P}_2(\Omega)} \left(F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu_k) \right).$$

Each step of this scheme balances minimizing the functional F and remaining close to the previous measure μ_k. The second term acts as a penalty preventing abrupt changes.

1.7.3 Nonparametric Sliced-Wasserstein Flows (SWF)

As detailed in Liutkus et al. (2019), the gradient flow framework can be extended to the *Sliced-Wasserstein* distance to develop nonparametric generative models. Unlike parametric

models that update network weights, **Sliced-Wasserstein Flows (SWF)** move the data particles themselves directly in the space to learn a target distribution ν.

Entropic Regularization and Particle Dynamics: The cost functional in SWF typically includes the SW distance and an **entropy term** $H(\mu) = \int \mu \log \mu$:

$$F(\mu) = SW_2^2(\mu, \nu) + \lambda H(\mu).$$

The **entropic regularization** term $\lambda H(\mu)$ introduces a repulsive force between particles (diffusion), ensuring they do not collapse into a single point, while the SW_2^2 term acts as an attractive force pulling the particles toward the structure of the target distribution ν.

The Flow Equation: The evolution of each particle X^i follows a McKean-Vlasov type dynamics:

$$\frac{dX_t^i}{dt} = -\nabla_{X^i} SW_2^2(\mu_t, \nu) + \sqrt{2\lambda} W_t^i,$$

where W_t^i is Brownian motion. The gradient of the squared Sliced-Wasserstein distance with respect to a particle's position X^i is calculated by summing the gradients of its 1D projections:

$$\nabla_{X^i} SW_2^2(\mu, \nu) \approx \frac{2}{L} \sum_{k=1}^{L} \theta_k \left(\langle X^i, \theta_k \rangle - \text{proj}_\nu^k(\langle X^i, \theta_k \rangle) \right),$$

where proj_ν^k represents the sorted sample value from the target distribution ν in the direction θ_k. This allows a set of initial noise samples to be iteratively "morphed" into the target distribution.

Chapter 2

Generative Modeling and Optimal Transport

2.1 Introduction

The first chapter established the theoretical foundations of Optimal Transport, focusing on the Monge–Kantorovich formulation, the geometry of Wasserstein spaces, and the structure of optimal transport maps, providing a rigorous mathematical framework for comparing and transforming probability distributions.

In this chapter, we move from the classical theory to its modern computational and machine learning perspectives, exploring how Optimal Transport ideas are incorporated into generative modeling frameworks. A central theme is the interpretation of probability transport as a dynamic process, which naturally leads to the study of normalizing flows and continuous-time generative models.

We cover Wasserstein-based learning objectives, sliced-Wasserstein methods (SWAEs and SWFs), and Wasserstein Flow Matching, which together form the conceptual bridge between the geometric theory of Optimal Transport and the dynamical systems perspective developed in the following chapter. As the main focus of this thesis is on the latter topic, the general framework of Wasserstein generative modeling is presented in a concise manner, emphasizing the conceptual ideas rather than the technical details.

Taxonomy of Models: Explicit and Implicit Density Following the taxonomy formalized by Kobyzev et al. (2020), generative models can be fundamentally categorized based on their mathematical representation of, and operational interaction with, the underlying data distribution. This yields two main paradigms:

1. **Explicit Density Models**: The goal is to determine a mathematical expression fro the provability distribution. The model is trained through the Maximum Likelihood Estimation. The Normalizing Flows consist a perfect example of this category, because they allow the exact calculation of $p(x)$

2. **Implicit Density Models**: In that case, the model does not provide direct formula for the density, but a sample generation mechanism e.g. GANs. The training is based on a comparison of samples through a discrete or metric distance.

The Manifold Hypothesis and Limitations of Classical Metrics A foundational assumption in high-dimensional data modeling|such as the financial time series analyzed by Ericson et al. (2024)|is that the data effectively concentrate on submanifolds of significantly lower dimensionality than the ambient space. Under this geometric setting, classical divergence measures, such as the *Kullback-Leibler* or *Jensen-Shannon* divergences, present severe limitations when the model and target distributions possess disjoint supports. In such cases, the resulting gradients vanish almost everywhere, rendering gradient-based optimization impossible.

The Necessity of a Geometric Approach: Optimal Transport As discussed by Huang & Malik (2025), adopting the Wasserstein distance (or Earth Mover's Distance) provides a robust solution to these limitations. In contrast to the Kullback–Leibler divergence, the Wasserstein metric accounts for the geometry of the underlying space, providing a meaningful training signal even when the distributions have entirely disjoint supports.

The connection between generative modeling and optimal transport theory allows the formulation of model training as a gradient flow within the space of probability measures. This geometric framework constitutes the cornerstone for the modern architectures examined in the following sections, beginning with Normalizing Flows, which bridge statistical modeling with geometric insight.

2.2 Normalizing Flows

Normalizing flows - NFs constitute a class of explicit density generative models, that they allow the exact calculation of data's likelihood, as well as efficient sampling. According to Kobyzev et al. (2020), the NFs main idea is the construction of a complex probability distribution via a sequence of invertible and differentiable transformations applied to a simple base distribution.

2.2.1 Mathematical Foundation and the Change of Variables Formula

Let $zon\mathbb{R}^d$ be a random variable that follows a simple base distribution $p_z(z)$, such as the standard multivariate normal distribution $\mathcal{N}(0, I)$. We consider an invertible transformation $f : \mathbb{R}^d \to \mathbb{R}^d$ $x = f(z)$. The distribution of the new variable x, which represents the real data, is given by the change of variables formula:

$$p_x(x) = p_z\big(f^{-1}(x)\big) \left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right| = p_z(z) \left|\det\left(\frac{\partial f(z)}{\partial z}\right)\right|^{-1}.$$

Where $\frac{\partial f}{\partial z}$ is the Jacobian matrix of the transformation f. In order for the above density to be tractable in high-dimensional problems, the architecture of f must be chosen such that:

- The transformation is easily invertible (f^{-1}), allowing both sampling and density estimation.
- The determinant of the Jacobian matrix can be computed with low computational cost, avoiding the $\mathcal{O}(d^3)$ complexity of general matrices.

Composition of Transformations One of the most powerful properties of Normalizing Flows (NFs) is that multiple transformations can be composed in sequence. If $f = f_K \circ f_{K-1} \circ \cdots \circ f_1$, then the final variable x is obtained through the successive application of the transformations f_k. The overall log-likelihood is expressed as the sum of the logarithms of the individual determinants:

$$\log p_x(x) = \log p_z(z) - \sum_{k=1}^{K} \log \left|\det\left(\frac{\partial f_k(z_{k-1})}{\partial z_{k-1}}\right)\right|.$$

This approach allows the gradual "warping" of a simple base distribution into a highly complex structure, capable of capturing multimodal distributions.

2.2.2 Flow Architectures and Invertibility Strategies

Following the classification formalized by Kobyzev et al. (2020), normalizing flows are primarily structured based on their approach to ensuring efficient computation of the Jacobian determinant:

- **Autoregressive Flows:** In this framework (e.g., MAF, IAF), each dimension x_i is conditioned strictly on the preceding dimensions $z_{1:i-1}$. This structural constraint yields a triangular Jacobian matrix, simplifying the determinant calculation to the product of its diagonal elements.

- **Coupling Flows:** The variable z is split into two parts $(z_{1:m}, z_{m+1:d})$. The first part remains unchanged, while the second is transformed via a function that depends parametrically on the first. Architectures such as RealNVP and Glow use this strategy to ensure determinants that can be computed in linear time $\mathcal{O}(d)$.

- **Linear Flows:** Transformations based on matrix operations (e.g., 1×1 convolutions), where invertibility is ensured via LU decomposition.

2.2.3 Continuous Normalizing Flows (CNF)

An alternative approach is to define the flow in continuous time. Instead of discrete steps f_k, the flow is described by an Ordinary Differential Equation (ODE):

$$\frac{dz(t)}{dt} = v(z(t), t, \theta),$$

where v is a neural network that defines the vector field, which is to be chosen so that the corresponding flow defines a transport map between the chosen measures. The change in density along the flow is given by the continuity equation and the Instantaneous Change of Variables formula:

$$\frac{\partial}{\partial t} \log p(z(t)) = -\mathrm{div}(v(z(t), t, \theta)),$$

where div denotes the divergence of the field. This approach is directly connected to Flow Matching, which will be examined later.

Connections to Optimal Transport and Geometric Constraints Although Normalizing Flows provide a powerful statistical framework, traditional architectures often map density paths along geometrically inefficient trajectories. Incorporating Optimal Transport theory and the Wasserstein-2 metric constrains these flows to follow optimal transport paths, corresponding to constant-speed geodesics. Minimizing the total transport cost along the velocity field not only enhances computational efficiency but also provides a strong inductive bias that improves generalization performance, particularly in data-constrained settings such as high-dimensional financial time series.

2.3 Generative Modeling via Wasserstein-2 Loss

While traditional generative modeling paradigms rely on likelihood maximization or adversarial training, recent frameworks focus on the direct minimization of the Wasserstein distance. This

approach allows model training to be formulated as a continuous dynamical process over the space of probability measures.

2.3.1 The W2 Loss Objective

The goal of generative modeling under the Optimal Transport framework is to find a model distribution μ that minimizes the second-order Wasserstein distance (W_2) from the true data distribution μ_d. The W_2 distance is defined as:

$$W_2^2(\mu, \mu_d) = \inf_{\pi \in \Pi(\mu, \mu_d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\pi(x, y).$$

The choice of W_2 over other metrics (such as W_1 or the KL divergence) is due to its excellent geometric properties and its close connection to gradient flows in probability space (Wasserstein space).

2.3.2 Distribution-Dependent ODEs and Gradient Flows

A central framework in this paradigm involves formulating a distribution-dependent Ordinary Differential Equation (ODE) that drives samples continuously toward the target distribution. Here, X_t denotes a stochastic process on $\mathbb{R}^d$, where $X_0 \sim \mu_0$ is drawn from a simple base distribution (e.g. a Gaussian), and the law of X_t at each time $t \in [0, 1]$ is the intermediate distribution μ_t. In particular, $X_1 \sim \mu_d$ corresponds to a random variable with the desired target distribution. The underlying dynamics are defined as:

$$\frac{dX_t}{dt} = \nabla \phi_t(X_t),$$

where ϕ_t denotes the Kantorovich potential arising from the dual formulation of the optimal transport problem between the evolving distribution μ_t and the target data distribution μ_d. The flow generated by this ODE thus defines a sequence of intermediate transport maps $T_t : \mathbb{R}^d \to \mathbb{R}^d$, with $T_t = \nabla \varphi_t$, pushing μ_0 forward to μ_t at each time t.

The fundamental theoretical guarantee establishes that the time-marginal laws of this ODE constitute a gradient flow minimizing the Wasserstein-2 (W_2) loss. This geometric property implies that the model distribution converges systematically toward μ_d, ensuring structural and optimization stability that is typically lacking in traditional adversarial frameworks.

2.3.3 The Role of Kantorovich Potentials in Training

In practice, the potential ϕ_t is unknown and must be estimated. Training involves:

- **Persistent Training:** A technique in which the network estimating the Kantorovich potential is continuously updated as samples evolve.

- **Euler Scheme:** For computational implementation, the ODE is discretized as:

$$X_{n+1} = X_n + \gamma \nabla \phi_n(X_n),$$

  where γ is the step size.

Advantages over Adversarial Methods In contrast to Wasserstein GANs (WGANs), which require enforcing a Lipschitz constraint (e.g., via gradient penalty), minimizing the W_2 loss via ODEs:

- Eliminates the necessity for adversarial training (min-max optimization), which is inherently unstable and prone to training pathology.

- Leverages the mathematical structure of Input Convex Neural Networks (ICNNs) to enforce the convexity of the potential ϕ, strictly adhering to the theoretical requirements of Brenier's theorem.

- Demonstrates robust performance across both low- and high-dimensional settings, effectively mitigating the risk of mode collapse.

2.3.4 Implementation within the Autoencoder Framework

This approach can also be integrated into autoencoder architectures, where the Wasserstein distance is used to enforce the structure of the latent space, ensuring that the distribution of encoded samples matches a predefined prior distribution. This forms the basis for Wasserstein Autoencoders (WAE) and their extensions, which will be discussed in the next section.

2.4 Sliced-Wasserstein Autoencoders (SWAE)

An **autoencoder** is a neural network architecture consisting of two components: an **encoder** $E_\phi : \mathcal{X} \to \mathcal{Z}$ that maps input data to a lower-dimensional latent space $\mathcal{Z}$, and a **decoder** $G_\psi : \mathcal{Z} \to \mathcal{X}$ that reconstructs the original data from the latent representation. The network is trained to minimize a reconstruction loss, encouraging $G_\psi(E_\phi(x)) \approx x$.

A **Wasserstein Autoencoder** extends this framework by imposing a geometric constraint on the latent space: the distribution of encoded samples $E_\phi(X)$ is required to match a fixed prior distribution μ_Z (e.g. a Gaussian) in the sense of the Wasserstein distance. Intuitively,

this ensures that the latent space is well-structured and that new samples can be generated by drawing from μ_Z and passing them through the decoder.

Despite the theoretical superiority of the Wasserstein-2 distance, its computation in high-dimensional spaces remains computationally expensive. Kolouri et al. (2018) introduce Sliced-Wasserstein Autoencoders (SWAE), which exploit the property that the Wasserstein distance admits a closed-form solution in one dimension, offering a model that combines geometric stability with fast training.

The Slicing Property of Wasserstein Distance The central idea is based on the Radon theorem, which states that a probability distribution in multiple dimensions can be fully represented by the set of its one-dimensional projections. The Sliced-Wasserstein (SW) distance is defined as the expectation (average) of the W_2 distances between projections of the two distributions along random directions:

$$SW_2^2(\mu, \nu) = \int_{S^{d-1}} W_2^2\big(R_\theta \mu, R_\theta \nu\big)\, d\theta,$$

where S^{d-1} is the unit sphere on which projection directions θ are defined, and R_θ is the Radon transform (projection). In one dimension, the W_2 distance has a closed form and is computed simply via sorting the samples, making SW highly efficient.

2.4.1 SWAE Architecture and Objective Function

The SWAE architecture consists of an encoder E_ϕ, a decoder D_ψ, and a predefined prior distribution p_z in the latent space. The loss function consists of two terms:

- **Reconstruction Loss:** The distance (typically L_2) between the input x and its reconstruction $D_\psi(E_\phi(x))$.
- **Distribution Matching Penalty:** The Sliced-Wasserstein distance between the distribution of encoded samples $q_z = E_\phi(x)$ and the prior distribution p_z.

$$\mathcal{L}(\phi, \psi) = \mathbb{E}_{x \sim P_{\text{data}}}\left[\|x - D_\psi(E_\phi(x))\|^2\right] + \lambda\, SW_2^2(q_z, p_z).$$

Numerical Implementation via Monte Carlo In practice, the integral over the sphere S^{d-1} is approximated via Monte Carlo sampling of L random directions θ_l:

$$SW_2^2(q_z, p_z) \approx \frac{1}{L} \sum_{l=1}^{L} W_2^2\big(R_{\theta_l} q_z, R_{\theta_l} p_z\big).$$

For each direction, the projected samples are sorted, and the W_2 distance is computed as the mean squared error between the ordered samples. This process allows SWAE to impose any desired distribution in the latent space (e.g., uniform, Gaussian mixtures), as long as sampling from it is feasible.

2.4.2 Comparison with VAE and WAE

Several architectural advantages characterize the Sliced Wasserstein Autoencoder (SWAE) relative to traditional Variational Autoencoders (VAEs) and Wasserstein Autoencoders (WAEs).

First, SWAEs eliminate the necessity for adversarial training, a stark contrast to WAE-GAN frameworks. This bypasses the inherent instabilities and non-convergence issues typically encountered in discriminator-based optimization. Furthermore, unlike VAEs, which rely heavily on a closed analytical form of the latent prior distribution p_z to compute the Kullback–Leibler divergence, SWAEs operate without this constraint, offering greater distributional flexibility.

Additionally, the geometric regularization introduced via the Sliced Wasserstein distance ensures that the latent representations preserve the underlying topological structure of the original data manifold more robustly.

These geometric and statistical properties carry significant utility in financial modeling applications. The capacity of SWAEs to shape the latent space without rigid analytical assumptions permits the integration of heavy-tailed latent priors. Such priors are crucial for capturing extreme market anomalies and tail-risk phenomena, ultimately enhancing the precision of Value at Risk (VaR) estimations in high-dimensional financial time series.

2.5 Sliced-Wasserstein Flows (SWF)

In contrast to parametric methods based on fixed neural network architectures, a non-parametric approach can be formulated via Sliced-Wasserstein Flows. This method defines the generative modeling paradigm as a continuous particle dynamics problem over the space of probability measures.

2.5.1 Functional Optimization and Gradient Flows

The problem is formulated as finding a probability measure μ that minimizes the Sliced-Wasserstein distance from the data distribution μ_d. The evolution of the distribution μ_t is described by the gradient flow:

$$\frac{\partial \mu_t}{\partial t} = -\nabla_{W_2} \mathcal{D}_{SW}(\mu_t, \mu_d),$$

where ∇_{W_2} denotes the gradient in Wasserstein space. This approach enables the gradual transformation of an initial distribution (e.g., noise) into the target distribution via the solution of the McKean{Vlasov (1.7.3) equation.

2.5.2 Connection to Stochastic Differential Equations (SDEs)

The flow is implemented by simulating a system of particles following the stochastic differential equation:

$$dX_t = -\nabla_x \left(\int_{S^{d-1}} V_\theta(X_t, \mu_t)\, d\theta \right) dt + \sqrt{2\beta^{-1}}\, dB_t,$$

where V_θ is the potential induced by the one-dimensional projection along direction θ, and dB_t is Brownian motion introducing diffusion to avoid local minima. This method provides strong theoretical convergence guarantees (finite-time error guarantees), which are typically absent in most parametric models.

2.6 Wasserstein Generative Learning of Conditional Distributions

The need for conditional data generation is critical for complex prediction and forecasting tasks. A powerful framework for learning the conditional distribution $P_{Y|X}$ involves minimizing the Wasserstein distance between the joint distributions, transforming conditional generation into a joint distribution alignment problem.

2.6.1 The Conditional Generator Mechanism

Let $G_\theta(\eta, x)$ be a generator, where η is latent noise and x is the conditioning variable (predictor). The goal is for the distribution of $G_\theta(\eta, X)$ to match $P_{Y|X}$. The optimization is performed over joint distributions:

$$\min_\theta W_1\big(P_{X,G_\theta(\eta,X)},\, P_{X,Y}\big).$$

Mitigating the Curse of Dimensionality A key result of the work is that using the Wasserstein distance allows the model to overcome the curse of dimensionality, provided that the data lie on a set with low intrinsic dimension d_α. The non-asymptotic error bound scales as n^{-1/d_α}, making the method highly effective in complex financial environments with many features.

2.7 Wasserstein Flow Matching (WFM)

Flow Matching (FM) represents a significant development in continuous flow-based generative models. By extending this framework to the space of probability distributions, Wasserstein Flow

Matching integrates optimal transport principles directly into the flow-matching paradigm.

2.7.1 Lifting Flow Matching to Families of Distributions

WFM transports not just individual points but entire families of distributions. A vector field v_t is defined to transport a source μ_0 to a target μ_1 along Wasserstein geodesics:

$$\mu_t = \big((1-t)\mathrm{id} + tT\big)_{\#}\mu_0,$$

where T is the optimal transport map (Brenier map). The training objective is:

$$\mathcal{L}_{\mathrm{WFM}}(\theta) = \mathbb{E}_{t,q_t}\left[\|v_\theta(x,t) - u_t(x)\|^2\right].$$

Applications in High-Dimensional Shapes The key innovation of WFM lies in its ability to handle data represented either analytically (e.g., Gaussians) or empirically (e.g., point clouds). By leveraging attention mechanisms and entropic optimal transport estimates, the model can synthesize complex structures (e.g., 3D shapes or cellular environments) while respecting the intrinsic Wasserstein geometry of the data.

Chapter 3

Approximation of Optimal Transport Maps via Linear-Control Neural ODEs

3.1 Introduction

Building on the previous chapters, which established the connection between Optimal Transport, Wasserstein geometry, and dynamical formulations of generative modeling via Neural ODEs and continuous flows, we now focus on a more structural question: whether optimal transport maps can be approximated within restricted classes of controlled dynamical systems.

In particular, this chapter investigates linear-control Neural ODEs as a parametrized family of continuous-time dynamical systems capable of generating transport maps through their flow. Within this framework, probability measures are evolved along trajectories induced by time-dependent vector fields with linear control structure.

A key aspect of the analysis lies in understanding the expressive power of these controlled systems. To this end, we employ tools from Lie algebra theory in order to characterize their controllability and geometric richness. These algebraic properties determine the extent to which the induced flows can approximate arbitrary transformations of the state space.

Following the framework introduced by A. Scagliotti and S. Farinelli [2025], we show that under suitable assumptions, linear-control dynamical systems exhibit a strong approximation property. In particular, W_2-optimal transport maps can be approximated arbitrarily well by the flows generated by such systems.

The chapter is organized as follows. Section 3.2 introduces the dynamical system formulation of linear-control Neural ODEs. Section 3.3 develops the associated Lie algebraic structure. Section 3.4 establishes the main approximation results, and Sections 3.5{3.7 connect these findings to optimal transport theory and construct the approximation of optimal transport

maps.

3.2 Linear-Control Neural ODEs as Dynamical Systems

3.2.1 The Dynamical Formulation of the Problem

The main idea is that a transport transformation can be generated as the terminal state of a continuous dynamical system. For this reason, we consider the following Ordinary Differential Equation:

$$\dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t), \tag{3.1}$$

where:

- $x(t) \in \mathbb{R}^n$ is the state of the system at time t,
- $F_i : \mathbb{R}^n \to \mathbb{R}^n$ are smooth vector fields,
- $u_i(t)$ are time-dependent controls.

This system can be written compactly as:

$$\dot{x}(t) = F(x(t))u(t), \tag{3.1}$$

where:

$$F(x) = \begin{bmatrix} F_1(x) & \cdots & F_k(x) \end{bmatrix}, \qquad u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_k(t) \end{bmatrix}. \tag{3.2}$$

We observe that the system is linear with respect to the control u, but generally nonlinear with respect to the state x. This form arises naturally in control theory and allows for the application of tools from geometric control theory and Lie algebras.

Open-loop control One of the fundamental characteristics of the considered model is that the control variable depends exclusively on time:

$$u = u(t), \tag{3.3}$$

and not on the current state of the system:

$$u \neq u(x(t), t). \tag{3.4}$$

In the terminology of control theory, such systems are called *open-loop control systems*. More specifically, the control signal is predetermined before the evolution of the system and remains independent of the trajectory followed by the state variable during the dynamics.

This is in contrast with *closed-loop* or *feedback* control systems, where the control depends explicitly on the current state of the system:

$$u = u(x(t), t). \tag{3.5}$$

In feedback control, the system continuously adapts its behavior according to its current position in the state space. Although feedback systems generally provide increased flexibility and robustness, the open-loop formulation adopted in the present framework possesses several substantial theoretical advantages which are crucial for the mathematical analysis developed in the paper.

First, the use of open-loop controls significantly simplifies the analytical structure of the problem. Since the controls are functions only of time, the associated Ordinary Differential Equation can be studied using classical tools from geometric control theory and the theory of flows of vector fields. In particular, the flow generated by the system can be described explicitly in terms of the chosen controls, allowing a rigorous investigation of approximation properties.

Secondly, the open-loop formulation establishes a direct connection with Continuous Normalizing Flows (CNFs) and Neural ODE architectures. In these models, the transformation of data distributions is generated through the integration of a time-dependent differential equation:

$$\dot{x}(t) = f(x(t), t), \tag{3.6}$$

where the evolution of the particles is determined by a velocity field depending continuously on time. The controls $u(t)$ in the present framework play a role analogous to trainable parameters evolving along the temporal direction.

Furthermore, from the optimization perspective, the learning problem becomes an optimization problem over admissible controls. More precisely, instead of optimizing directly over maps between probability measures, one optimizes over the space of control functions:

$$u : [0, 1] \to \mathbb{R}^k. \tag{3.7}$$

Consequently, the approximation of an optimal transport map is reformulated as the search for a control function whose induced flow transports the initial distribution close to the target distribution.

An additional important aspect concerns the regularity assumptions imposed on the controls. In the paper, the admissible controls are chosen in the Hilbert space:

$$u \in L^2([0,1], \mathbb{R}^k). \tag{3.8}$$

The space $L^2([0,1], \mathbb{R}^k)$ consists of square-integrable functions, namely functions satisfying:

$$\int_0^1 |u(t)|^2 dt < +\infty. \tag{3.9}$$

This assumption is sufficiently weak to allow a large class of admissible controls, while simultaneously being strong enough to guarantee the well-posedness of the associated dynamical system.

Indeed, under suitable regularity assumptions on the vector fields

$$F_i : \mathbb{R}^n \to \mathbb{R}^n, \tag{3.10}$$

classical existence and uniqueness theorems for ODEs imply that, for every initial condition $x_0 \in \mathbb{R}^n$, the controlled system

$$\begin{cases} \dot{x}(t) = \sum_{i=1}^{k} F_i(x(t)) u_i(t), \\ x(0) = x_0, \end{cases} \tag{3.11}$$

admits a unique absolutely continuous solution.

The corresponding solution defines a flow map:

$$\Phi_u(x_0) := x_u(1, x_0), \tag{3.12}$$

which associates to every initial point its final position at time $t = 1$.

Geometrically, the control function determines how the particles evolve continuously in time inside the ambient space. Different choices of controls generate different flows and therefore different deformations of the underlying probability distribution.

This observation is particularly important in the context of Optimal Transport. Instead of constructing directly an optimal transport map

$$T : \mathbb{R}^n \to \mathbb{R}^n, \tag{3.13}$$

the paper proposes to approximate T through flows generated by suitable controls. In this way, the transport problem is transformed into a dynamical approximation problem.

Finally, the open-loop formulation plays a central role in the approximation theorems proved later in the paper. The density results obtained through Lie algebra techniques rely heavily on the fact that the admissible controls are time-dependent functions. This structure allows the generation of highly expressive flows capable of approximating broad classes of diffeomorphisms, including optimal transport maps under suitable regularity assumptions.

3.2.2 Existence and Uniqueness of Solutions

For every initial condition $x_0 \in \mathbb{R}^n$, we consider the following Cauchy problem associated with the controlled dynamical system:

$$\begin{cases} \dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t), \\ x(0) = x_0. \end{cases} \tag{2.2}$$

The previous system describes the temporal evolution of a particle whose motion is determined by the vector fields F_i and the corresponding time-dependent controls $u_i(t)$. For every admissible control function u, the differential equation generates a trajectory in the state space $\mathbb{R}^n$.

A fundamental question concerns the well-posedness of the system. More precisely, one must determine whether the system admits solutions, whether these solutions are unique, and whether they depend continuously on the initial data. These properties are essential both from the mathematical and computational viewpoints, since they guarantee that the dynamics are stable and deterministic.

The existence and uniqueness of solutions follow from the classical Picard–Lindelöf theorem. Recall that the theorem applies to Ordinary Differential Equations of the form

$$\dot{x}(t) = f(t, x(t)), \tag{3.14}$$

provided that the vector field f satisfies suitable regularity assumptions, typically continuity with respect to time and local Lipschitz continuity with respect to the state variable.

In the present framework, the vector field is given by

$$f(t, x) = \sum_{i=1}^{k} F_i(x)u_i(t). \tag{3.15}$$

Since the vector fields

$$F_i : \mathbb{R}^n \to \mathbb{R}^n \tag{3.16}$$

are assumed to be smooth (C^∞) functions, they are in particular locally Lipschitz continuous. Moreover, the controls

$$u_i \in L^2([0,1]) \tag{3.17}$$

are measurable functions of time. Therefore, the resulting vector field satisfies the hypotheses required by the Carathéodory theory for ODEs, ensuring the existence of absolutely continuous solutions.

Consequently, for every initial condition $x_0 \in \mathbb{R}^n$, there exists a unique trajectory

$$x_u(\cdot, x_0) : [0,1] \to \mathbb{R}^n \tag{3.18}$$

satisfying the differential equation almost everywhere.

More precisely, the following properties hold:

- For every admissible control u and every initial condition x_0, there exists a unique solution to the system.
- The solution depends continuously on the initial condition x_0. Hence, small perturbations of the initial data produce only small perturbations of the trajectory.
- The dependence of the solution on the initial condition is smooth whenever the vector fields F_i are smooth.

The continuous dependence on initial conditions is particularly important in applications related to probability measures and transport problems. Indeed, when transporting an entire distribution, one studies simultaneously the evolution of infinitely many particles. Stability of trajectories guarantees that nearby particles evolve coherently under the flow.

Furthermore, the smooth dependence on the initial condition implies that the dynamics generate smooth deformations of the ambient space. This property is fundamental in the theory of diffeomorphic flows and continuous normalizing flows.

The existence and uniqueness theorem allows one to define the flow associated with the control u. Denoting by

$$x_u(t, x_0) \tag{3.19}$$

the unique solution of the system starting from x_0, one defines the flow map

$$\Phi_u^t(x_0) := x_u(t, x_0). \tag{3.20}$$

In particular, the time-one flow is given by

$$\Phi_u(x_0) := \Phi_u^1(x_0) = x_u(1, x_0). \tag{3.21}$$

The map

$$\Phi_u : \mathbb{R}^n \to \mathbb{R}^n \tag{3.22}$$

associates to every initial point its final position after one unit of time under the dynamics generated by the control u. We note that the choice of $T = 1$ as the terminal time is an arbitrary scaling convention; the framework extends naturally to any terminal time $T > 0$.

Geometrically, the flow can be interpreted as a continuous deformation of the space induced by the vector field of the system. Each point evolves along a trajectory determined by the controlled dynamics, and the collection of all trajectories defines the global deformation generated by the ODE.

Under suitable regularity assumptions on the vector fields, the flow map Φ_u is a diffeomorphism of $\mathbb{R}^n$. In particular, it is invertible and both the map and its inverse are smooth. This property plays a crucial role in the approximation results proved later in the paper, since optimal transport maps are approximated through such diffeomorphic flows.

The existence of a well-defined flow is therefore one of the key mathematical ingredients connecting controlled dynamical systems with Optimal Transport and Continuous Normalizing Flows. Indeed, instead of constructing directly a transport map between probability measures, the paper proposes to generate such maps dynamically through the evolution induced by the controlled ODE.

3.2.3 The Flow of the Dynamical System

For a fixed admissible control function

$$u \in L^2([0, 1], \mathbb{R}^k), \tag{3.23}$$

the controlled Ordinary Differential Equation

$$\begin{cases} \dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t), \\ x(0) = x_0, \end{cases} \tag{3.24}$$

generates a unique trajectory starting from the initial condition $x_0 \in \mathbb{R}^n$.

We denote this solution by

$$x_u(t, x_0), \tag{3.25}$$

in order to emphasize both its dependence on the control u and on the initial state x_0.

The trajectory

$$t \mapsto x_u(t, x_0) \tag{3.26}$$

describes the continuous evolution of a particle moving inside the ambient space under the action of the controlled vector field. For each initial point x_0, the ODE determines a unique path in $\mathbb{R}^n$, and the collection of all such trajectories provides a global description of the dynamics induced by the system.

A central object in the analysis developed in the paper is the associated flow map. More precisely, the final state of the trajectory at time $t = 1$ defines the mapping

$$\Phi_u(x_0) := x_u(1, x_0). \tag{3.3}$$

Thus, the map

$$\Phi_u : \mathbb{R}^n \to \mathbb{R}^n \tag{3.27}$$

associates to every initial point its position after one unit of time under the dynamics generated by the control u.

More generally, one may define the time-dependent flow

$$\Phi_u^t(x_0) := x_u(t, x_0), \tag{3.28}$$

so that

$$\Phi_u = \Phi_u^1. \tag{3.29}$$

Geometrically, the flow can be interpreted as a continuous deformation of the ambient space. Every point evolves along a trajectory determined by the controlled dynamics, and the entire space is smoothly transported through time.

An essential property of the flow map is its regularity. Under the smoothness assumptions imposed on the vector fields F_i, the paper proves that the flow map Φ_u is a diffeomorphism of $\mathbb{R}^n$. Consequently, the map satisfies the following properties:

- **Invertibility:** every final point corresponds to a unique initial point.
- **Smoothness:** the map Φ_u is differentiable with smooth derivatives.
- **Smooth inverse:** the inverse map

$$\Phi_u^{-1} : \mathbb{R}^n \to \mathbb{R}^n \tag{3.30}$$

  is also smooth.

The invertibility of the flow is a fundamental feature of the model. Since trajectories of the ODE cannot intersect, the evolution generated by the system preserves the individuality of particles. In particular, two distinct initial points cannot evolve into the same final point under the same dynamics.

The diffeomorphic structure of the flow is crucial for its connection with Continuous Normalizing Flows and invertible generative models. In normalizing flow theory, one transforms probability densities through invertible maps. If

$$y = \Phi_u(x), \tag{3.31}$$

then the transformed density satisfies the change-of-variables formula:

$$\rho_\nu(y) = \rho_\mu(\Phi_u^{-1}(y)) \left|\det D\Phi_u^{-1}(y)\right|. \tag{3.32}$$

The validity of this formula requires the invertibility and differentiability of the flow map, making the diffeomorphic property of Φ_u indispensable for transporting probability measures in a mathematically consistent way.

Moreover, by varying the control functions $u(t)$, one generates different diffeomorphic deformations of the ambient space. The central idea of the paper is precisely that suitable choices of controls allow these flows to approximate optimal transport maps arbitrarily well.

3.3 Approximation Properties and Lie Algebras

One of the central mathematical ingredients of the paper is the connection between geometric control theory and approximation properties of controlled dynamical systems. More specifically,

the expressive power of the considered Neural ODE model is studied through the Lie algebra generated by the underlying vector fields.

The main idea is that, although the system is directly driven only by a finite number of vector fields

$$F_1, \dots, F_k, \tag{3.33}$$

the nonlinear interaction between these fields generates additional directions of motion. These new directions arise through Lie brackets and play a fundamental role in the controllability and approximation capabilities of the system.

The theory developed in the paper shows that, under suitable assumptions, the flows generated by the controlled system are sufficiently rich to approximate broad classes of diffeomorphisms, including optimal transport maps.

3.3.1 Lie Brackets and Geometric Interpretation

Let

$$X, Y : \mathbb{R}^n \to \mathbb{R}^n \tag{3.34}$$

be two smooth vector fields. The Lie bracket of X and Y is defined by

$$[X, Y](x) = DY(x)X(x) - DX(x)Y(x). \tag{3.35}$$

Here,

$$DX(x), \quad DY(x) \tag{3.36}$$

denote the Jacobian matrices of the vector fields evaluated at the point x.

The Lie bracket is one of the most fundamental concepts in differential geometry and geometric control theory. Intuitively, it measures the noncommutativity of the flows generated by the two vector fields.

To understand this geometrically, consider the following procedure:

- move for a short time along the vector field X,
- then move along Y,
- then move backwards along X,
- and finally move backwards along Y.

If the flows generated by X and Y commuted perfectly, the particle would return exactly to

its initial position. However, in general, this does not happen. The discrepancy produced by this infinitesimal commutator motion is precisely described by the Lie bracket $[X, Y]$.

Therefore, the Lie bracket captures new infinitesimal directions of motion that are not explicitly present in the original system.

This observation is extremely important in control theory. Even if the system can move directly only along a finite number of vector fields, repeated combinations of motions may generate entirely new directions.

Consequently, the dynamical system may possess a much richer geometric structure than what initially appears from the original vector fields alone.

As a simple illustration, suppose that the admissible motions are generated only by two vector fields F_1 and F_2. Although direct motion is restricted to combinations of these two directions, alternating rapidly between them can produce an effective displacement along the bracket direction

$$[F_1, F_2]. \tag{3.37}$$

Iterating this procedure further generates higher-order directions such as

$$[F_1, [F_1, F_2]], \qquad [F_2, [F_1, F_2]], \tag{3.38}$$

and so on.

In general, this iterative process does not necessarily terminate in a finite number of steps. One may continue generating higher-order Lie brackets indefinitely. However, since all vector fields are defined on a finite-dimensional space $\mathbb{R}^n$, the span of all iterated brackets can contain at most n linearly independent directions at each point. In special cases, such as nilpotent Lie algebras, the process does terminate after finitely many steps because sufficiently deep brackets vanish identically.

This mechanism lies at the heart of geometric controllability theory and explains why relatively simple controlled systems may exhibit extremely expressive dynamical behavior.

3.3.2 Lie Algebra of the System

The collection of vector fields

$$F_1, \ldots, F_k \tag{3.39}$$

generates a Lie algebra, denoted by

$$\mathrm{Lie}(F_1, \ldots, F_k). \tag{3.40}$$

More precisely, this Lie algebra consists of:

- all linear combinations of the vector fields,
- all Lie brackets between them,
- all iterated (nested) Lie brackets.

For example, the Lie algebra contains vector fields of the form

$$[F_i, F_j], \tag{3.41}$$

as well as higher-order terms such as

$$[F_i, [F_j, F_l]], \tag{3.42}$$

and arbitrarily deep nested commutators.

The Lie algebra generated by the system represents all infinitesimal directions that can be dynamically produced through admissible controls and repeated interactions of the vector fields.

Geometrically, one may think of the Lie algebra as describing the effective tangent directions accessible to the system. Although the original controls act only along the vector fields F_i, the nonlinear structure of the dynamics enlarges the set of reachable directions through Lie brackets.

This enlargement mechanism is fundamental for approximation theory. Indeed, the expressive power of the controlled Neural ODE depends not only on the original vector fields but also on the richness of the Lie algebra they generate.

The paper exploits precisely this idea. Under suitable assumptions, the Lie algebra generated by the vector fields becomes sufficiently rich to approximate arbitrary smooth vector fields on compact sets.

This property is referred to in the paper as the *Lie algebra strong approximating property* and constitutes one of the central assumptions behind the universal approximation results.

More specifically, the richness of the Lie algebra implies that the flows generated by the controlled system are dense in large classes of diffeomorphisms. Consequently, by choosing suitable controls $u(t)$, one can generate flows that approximate complex nonlinear transformations with arbitrary precision.

This observation provides the mathematical foundation for interpreting controlled Neural ODEs as universal approximators of transport maps.

In particular, the approximation results proved later in the paper show that the flows generated by the system can approximate optimal transport maps associated with the Wasserstein distance W_2. Therefore, the Lie algebra structure becomes the key mechanism connecting geometric control theory with Optimal Transport and continuous generative modeling.

From the machine learning perspective, this result is particularly significant. It implies that relatively simple dynamical architectures may possess extremely high expressive power, provided that the underlying vector fields generate a sufficiently rich Lie algebra.

Thus, the approximation capabilities of the model are deeply connected with the geometric structure of the controlled dynamics rather than merely with the dimensionality of the parameter space.

3.4 Strong Approximating Property

One of the main theoretical assumptions introduced in the paper is the so-called *Lie algebra strong approximating property*. This property is the fundamental mechanism that allows the controlled dynamical system to generate highly expressive flows and, ultimately, to approximate optimal transport maps.

The approximation framework developed in the paper is based on the Lie algebra generated by the vector fields

$$F_1, \dots, F_k. \tag{3.43}$$

Although the controlled system evolves directly only along these vector fields, the nonlinear interactions created through Lie brackets generate a much richer family of admissible dynamics. The key question is therefore whether this generated family is sufficiently large to approximate arbitrary smooth vector fields.

This requirement is formalized through the following density assumption.

Assumption

Assumption 3.4.1. The Lie algebra generated by $F_1, \dots, F_k$ satisfies the *strong approximation property*, i.e., for every smooth vector field

$$Y : \mathbb{R}^n \to \mathbb{R}^n, \tag{3.44}$$

every compact set

$$K \subset \mathbb{R}^n, \tag{3.45}$$

and every tolerance

$$\varepsilon > 0, \tag{3.46}$$

there exists a vector field

$$X \in \mathrm{Lie}(F_1, \ldots, F_k) \tag{3.47}$$

such that

$$\sup_{x \in K} |X(x) - Y(x)| < \varepsilon. \tag{3.48}$$

In other words, the Lie algebra generated by the system is dense in the space of smooth vector fields with respect to uniform convergence on compact subsets.

Geometrically, this means that the controlled system can reproduce arbitrarily well the local behavior of any smooth dynamical system. Even though the model is constructed from a finite number of vector fields, the repeated Lie bracket interactions dramatically enlarge the set of dynamically realizable motions.

This property is the core mathematical ingredient behind the approximation capabilities of the model. Indeed, if arbitrary smooth vector fields can be approximated, then the associated flows generated by these vector fields can also be approximated through suitable controls.

Consequently, the dynamics generated by the controlled Neural ODE become sufficiently expressive to approximate broad classes of nonlinear geometric deformations.

The strong approximating property therefore plays a role analogous to the classical universal approximation principle in neural network theory. However, in the present framework, the approximation concerns dynamical flows and diffeomorphic transformations rather than static functions.

A classical example arises from polynomial vector fields constructed using Hermite polynomials. Consider a one-dimensional controlled system of the form

$$\dot{x} = F_1(x)u_1(t) + F_2(x)u_2(t), \tag{3.49}$$

where $F_1(x) = 1$ and $F_2(x) = x$. Iterated Lie brackets generate higher-order polynomial terms, and in particular one obtains a hierarchy of vector fields corresponding to Hermite-type structures after suitable orthogonalization. This yields a basis of polynomial deformations capable of approximating smooth nonlinear transformations on compact sets.

3.4.1 Universal Approximation Theorem

Under the strong approximating property, the paper proves a universal approximation theorem for the flows generated by the controlled system.

More precisely, let

$$\Psi : \mathbb{R}^n \to \mathbb{R}^n \tag{3.50}$$

be a smooth diffeomorphism that is isotopic to the identity.

The isotopy assumption is important because it ensures that the target transformation can be continuously generated through a smooth time-dependent deformation. This is naturally compatible with the flow-based structure of Neural ODEs.

Theorem 3.4.1. *Theorem For every compact set*

$$K \subset \mathbb{R}^n \tag{3.51}$$

and every

$$\varepsilon > 0, \tag{3.52}$$

there exists an admissible control function

$$u : [0, 1] \to \mathbb{R}^k \tag{3.53}$$

such that the associated flow map satisfies

$$\sup_{x \in K} |\Psi(x) - \Phi_u(x)| \leq \varepsilon. \tag{3.54}$$

Therefore, the flow generated by the controlled Neural ODE approximates the target diffeomorphism uniformly on compact subsets.

This result constitutes one of the main theoretical contributions of the paper. It establishes that flows generated by linear-control Neural ODEs possess a universal approximation property at the level of smooth diffeomorphic transformations.

In particular, the theorem shows that the expressive power of the model is not merely a consequence of parameter dimensionality, but rather a consequence of the geometric richness of the Lie algebra generated by the vector fields.

Theorem 3.4.2. *Theorem Approximation of Isotopic Diffeomorphisms*

The central approximation theorem proved in the paper establishes that the flows generated by the controlled system are dense in the class of smooth diffeomorphisms isotopic to the identity.

More precisely, assume that the family of vector fields

$$F_1, \dots, F_k$$

satisfies the Lie algebra strong approximating property.

Then, for every smooth diffeomorphism

$$\Psi : \mathbb{R}^n \to \mathbb{R}^n$$

which is isotopic to the identity, for every compact set

$$K \subset \mathbb{R}^n,$$

and for every

$$\varepsilon > 0,$$

there exists an admissible control

$$u \in L^2([0,1], \mathbb{R}^k)$$

such that the corresponding flow map satisfies

$$\sup_{x \in K} |\Psi(x) - \Phi_u(x)| < \varepsilon.$$

This theorem constitutes the fundamental approximation result of the paper. It shows that controlled Neural ODE flows possess a universal approximation property at the geometric level of smooth diffeomorphic transformations.

The significance of this result is substantial. Instead of approximating static functions, as in classical neural network approximation theory, the present framework approximates entire dynamical deformations of the ambient space.

The theorem also highlights the fundamental role played by the Lie algebra generated by the vector fields. The approximation capability of the model is not simply a consequence of parameter dimensionality, but rather of the geometric richness of the admissible dynamics generated through Lie brackets and iterated commutators.

Consequently, sufficiently expressive controlled dynamical systems are capable of reproducing highly complex nonlinear geometric transformations through suitable choices of time-dependent controls.

3.5 Connection with Optimal Transport Maps

The approximation framework developed in the paper becomes particularly important when applied to Optimal Transport theory.

As discussed in the previous chapter, Brenier's theorem implies that, under suitable assumptions on the probability measures and for the quadratic Wasserstein cost, the optimal transport map can be represented as

$$T = \nabla\varphi, \tag{3.55}$$

where φ is a convex potential.

Moreover, under standard regularity assumptions on the densities and the domains, regularity theory for the Monge–Ampère equation implies that the optimal transport map possesses strong smoothness properties. In particular,

$$T \in C^\infty. \tag{3.56}$$

Remark on Variational Approximation and Γ-Convergence The framework also employs variational approximation arguments related to the regularization of transport maps. In practice, the underlying probability measures are typically accessed through empirical approximations, which converge only in the weak-* sense to the target distributions. As a consequence, variational problems are effectively solved on sequences of approximating functionals defined at the empirical level. In this setting, classical pointwise notions of convergence are not sufficient to guarantee stability of minimizers. In this context, Γ-convergence provides a natural framework for studying the convergence of approximation schemes and their associated minimization problems.

More precisely, Γ-convergence is a notion of variational convergence particularly suited for the analysis of asymptotic behavior of sequences of energy functionals. One of its main advantages is that minimizers of approximating functionals converge to minimizers of the limiting functional under suitable compactness assumptions.

Within this framework, these variational ideas are used to justify the approximation of transport maps through smoother dynamical constructions compatible with the controlled flow formulation.

Although the detailed theory of Γ-convergence is beyond the scope of the present thesis, its role is conceptually important in connecting variational optimal transport problems with dynamical approximation schemes.

Hence, the optimal transport map belongs to the class of smooth diffeomorphic transforma-

tions considered in the universal approximation theorem.

This observation provides the key connection between Optimal Transport and the controlled dynamical framework studied in the paper.

Indeed, since the flows generated by the controlled system are dense in the class of smooth diffeomorphisms isotopic to the identity, it follows that optimal transport maps can be approximated arbitrarily well by suitable controlled flows.

More precisely, for every compact set

$$K \subset \mathbb{R}^n \tag{3.57}$$

and every

$$\varepsilon > 0, \tag{3.58}$$

one can construct a control function u such that the associated flow map Φ_u satisfies

$$\sup_{x \in K} |T(x) - \Phi_u(x)| < \varepsilon. \tag{3.59}$$

Therefore, the optimal transport problem can be reformulated dynamically: instead of constructing directly the transport map T, one seeks a controlled evolution whose terminal flow approximates the desired transport transformation.

This result constitutes the main conceptual bridge between Optimal Transport theory and controlled Neural ODE architectures developed in the paper.

3.6 The OT Map as a Flow of a Dynamical System

One of the central ideas developed in the paper is that the optimal transport map can be interpreted dynamically as the terminal flow of a suitable time-dependent differential equation.

This point of view is fundamental because it establishes the bridge between Optimal Transport theory and controlled Neural ODE architectures. Instead of viewing the transport map as a static transformation between probability measures, the paper reformulates transport as a continuous evolution generated by a dynamical system.

More precisely, the goal is to construct a smooth deformation connecting the identity map with the optimal transport map. Once such a deformation is obtained, the transport map can be realized as the endpoint of a flow, allowing the approximation theory developed previously to be applied.

3.6.1 Construction of the Interpolation

Let

$$\widetilde{T} : \Omega_1 \to \Omega_2 \tag{3.60}$$

denote the regularized optimal transport map considered in the paper.

The first step consists in constructing a continuous interpolation between the identity transformation and the map $\widetilde{T}$.

The paper introduces the family of maps

$$\widetilde{T}_t = (1-t)Id + t\widetilde{T}, \tag{3.6}$$

for

$$t \in [0,1]. \tag{3.61}$$

This interpolation satisfies the boundary conditions

$$\widetilde{T}_0 = Id, \tag{3.62}$$

and

$$\widetilde{T}_1 = \widetilde{T}. \tag{3.63}$$

Therefore, the family $\widetilde{T}_t$ defines a continuous deformation from the identity map to the optimal transport map.

Geometrically, each intermediate map $\widetilde{T}_t$ describes a partial deformation of the ambient space. When $t = 0$, no deformation has yet occurred, while at $t = 1$, the full transport transformation is recovered.

The interpolation plays a crucial role because it converts the transport problem into a time-dependent geometric evolution. Instead of constructing directly the map $\widetilde{T}$, one studies the continuous trajectory of deformations connecting the identity to the target transformation.

This dynamical interpretation is precisely what allows the transport map to be embedded into the flow framework developed earlier in the paper.

3.6.2 Construction of a Time-Varying Vector Field

After constructing the interpolation, the framework defines a time-dependent vector field whose flow generates the desired transport map.

The vector field is given by

$$F(t,y) = -\widetilde{T}_t^{-1}(y) + \widetilde{T}\big(\widetilde{T}_t^{-1}(y)\big). \tag{3.64}$$

This construction constitutes one of the key analytical steps of the proof.

The idea is to define a vector field that is compatible with the interpolation $\widetilde{T}_t$. More precisely, the vector field is chosen in such a way that the associated flow follows exactly the path determined by the interpolation between the identity and the transport map.

Observe that the inverse map

$$\widetilde{T}_t^{-1} \tag{3.65}$$

appears explicitly in the definition of the vector field. This reflects the fact that the dynamics must be expressed in Eulerian coordinates, namely in terms of the current position y of the evolving particle.

The term

$$\widetilde{T}\big(\widetilde{T}_t^{-1}(y)\big) \tag{3.66}$$

represents the target position associated with the particle currently located at y, while

$$\widetilde{T}_t^{-1}(y) \tag{3.67}$$

corresponds to its original configuration under the interpolation.

Consequently, the vector field measures the instantaneous displacement required to move the evolving configuration toward the final transport map.

Proof of Isotopy The framework then analyzes the flow generated by the non-autonomous differential equation

$$\frac{d}{dt}\Psi(t,x) = F(t,\Psi(t,x)). \tag{3.68}$$

The main result is that the solution of this system satisfies

$$\Psi(1,x) = T(x). \tag{3.69}$$

Thus, the optimal transport map is realized as the terminal flow of the constructed time-dependent vector field.

This result has a major geometric consequence: the optimal transport map is isotopic to the identity.

Indeed, the family of intermediate maps

$$\Psi(t, \cdot) \tag{3.70}$$

provides a continuous path of diffeomorphisms connecting

$$Id \tag{3.71}$$

to

$$T. \tag{3.72}$$

The isotopy property is one of the decisive ingredients required for the application of the universal approximation theorem proved earlier in the paper.

Since the approximation theorem applies to diffeomorphisms isotopic to the identity, proving that the optimal transport map belongs to this class allows one to transfer the approximation results directly to the Optimal Transport setting.

Therefore, the paper establishes a complete conceptual chain:

Optimal Transport Map $\Longrightarrow$ Flow of a Smooth Dynamical System $\Longrightarrow$ Approximable by Controlled Flows.

Proposition 3.6.1 (Isotopy of the Optimal Transport Map)**.** *Let $\widetilde{T}$ denote the regularized optimal transport map. Then there exists a smooth time-dependent vector field $F(t, \cdot)$ generating a flow $\Psi(t, \cdot)$ such that*

$$\Psi(0, x) = x, \qquad \Psi(1, x) = \widetilde{T}(x),$$

for all x in the domain. In particular, $\widetilde{T}$ is isotopic to the identity through a smooth family of diffeomorphisms.

Proof. Consider the interpolation

$$\widetilde{T}_t = (1 - t)Id + t\widetilde{T}, \qquad t \in [0, 1].$$

Define the associated time-dependent vector field

$$F(t, y) = -\widetilde{T}_t^{-1}(y) + \widetilde{T}\big(\widetilde{T}_t^{-1}(y)\big).$$

Let $\Psi(t,x)$ be the solution of the non-autonomous ODE

$$\begin{cases} \dfrac{d}{dt}\Psi(t,x) = F(t,\Psi(t,x)), \\ \Psi(0,x) = x. \end{cases}$$

We claim that $\Psi(t,x) = \widetilde{T}_t(x)$. Indeed, differentiating the interpolation yields

$$\frac{d}{dt}\widetilde{T}_t(x) = -x + \widetilde{T}(x).$$

Now set $y = \widetilde{T}_t(x)$, so that $x = \widetilde{T}_t^{-1}(y)$. Then

$$F(t,\widetilde{T}_t(x)) = -\widetilde{T}_t^{-1}(\widetilde{T}_t(x)) + \widetilde{T}\big(\widetilde{T}_t^{-1}(\widetilde{T}_t(x))\big) = -x + \widetilde{T}(x).$$

Hence,

$$\frac{d}{dt}\widetilde{T}_t(x) = F(t,\widetilde{T}_t(x)).$$

By uniqueness of solutions of ODEs, it follows that

$$\Psi(t,x) = \widetilde{T}_t(x) \quad \text{for all } t \in [0,1].$$

Evaluating at $t = 1$ gives

$$\Psi(1,x) = \widetilde{T}_1(x) = \widetilde{T}(x).$$

Therefore, $\widetilde{T}$ is realized as the terminal flow of a smooth time-dependent dynamical system, and is isotopic to the identity. □

3.7 Approximation of the OT Map via Linear-Control Neural ODEs

After establishing that the optimal transport map can be represented as a smooth flow isotopic to the identity, the paper combines this result with the universal approximation theorem for controlled Neural ODEs.

More precisely, combining 3.4.1 with 3.6.1 yields the main approximation corollary of this framework.

Corollary. For every

$$\varepsilon > 0, \tag{3.73}$$

there exists a control function

$$u : [0, 1] \to \mathbb{R}^k \tag{3.74}$$

such that the associated controlled flow satisfies

$$\sup_{x \in \Omega_1} |T(x) - \Phi_u(x)| \leq \varepsilon. \tag{3.75}$$

In other words, the flow generated by the linear-control Neural ODE approximates the optimal transport map uniformly on compact subsets.

This result constitutes the central conclusion of the theoretical analysis.

Indeed, it proves rigorously that optimal transport maps can be approximated arbitrarily well through flows generated by controlled dynamical systems of the form

$$\dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t). \tag{3.76}$$

The significance of this theorem is both mathematical and computational.

From the mathematical perspective, it establishes a deep connection between:

- Optimal Transport theory,
- geometric control theory,
- Lie algebra approximation,
- and dynamical systems.

From the machine learning perspective, the result provides a rigorous theoretical foundation for flow-based generative models and continuous normalizing flows.

In particular, the theorem shows that controlled Neural ODE architectures are capable of learning highly complex transport transformations through continuous dynamical evolutions.

The approximation result can therefore be summarized schematically as

$$\boxed{\text{Optimal Transport} \approx \text{Controlled Neural ODE Flow}} \tag{3.77}$$

This viewpoint is one of the main conceptual contributions of the paper. It shows that optimal transport maps can be interpreted not merely as static geometric objects, but as dynamical transformations generated through continuous flows.

Consequently, the theory developed in the paper provides a rigorous mathematical framework connecting Optimal Transport with modern invertible generative modeling and continuous

deep learning architectures.

3.8 Conclusion

In this chapter, we studied the approximation of optimal transport maps through controlled dynamical systems and linear-control Neural ODEs. The analysis established the theoretical bridge between Optimal Transport theory, geometric control theory, and continuous deep learning architectures.

We first introduced the controlled dynamical system

$$\dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t), \tag{3.78}$$

where the controls depend only on time. The associated flow maps generated smooth diffeomorphic deformations of the ambient space, providing a natural dynamical framework for modeling transport transformations.

A central role was played by the Lie algebra generated by the vector fields $F_1, \ldots, F_k$. Through Lie brackets and iterated commutators, the system acquires a significantly richer geometric structure than what is initially visible from the original vector fields alone.

Under the Lie algebra strong approximating property introduced in the paper, the flows generated by the controlled system become dense in the class of smooth diffeomorphisms isotopic to the identity. This led to a universal approximation theorem for controlled Neural ODE flows.

The chapter then connected these approximation results with Optimal Transport theory. Using the regularity properties of Brenier maps and the interpolation construction developed in the paper, the optimal transport map was interpreted as the terminal flow of a smooth time-dependent dynamical system.

This observation allowed the application of the universal approximation theorem to the Optimal Transport setting. As a consequence, the paper proves that optimal transport maps can be approximated arbitrarily well by flows generated through linear-control Neural ODEs.

Therefore, the approximation framework developed in the paper provides a rigorous mathematical justification for the use of continuous normalizing flows and controlled Neural ODE architectures in transport-based generative modeling.

More broadly, the results presented in this chapter illustrate how ideas from geometric control theory and dynamical systems can be combined with Optimal Transport to construct expressive and mathematically well-founded continuous generative models.

Chapter 4

Normalizing Flows via Linear-Control Neural ODEs

4.1 Introduction

Chapter 3 established theoretical results showing that optimal transport maps can be approximated by linear-control dynamical systems via Neural ODEs and Lie-algebraic controllability arguments.

Building on these results, the present chapter focuses on the constructive and algorithmic realization of this framework. Instead of addressing only approximation properties, we now consider how optimal transport maps can be explicitly learned and implemented through a variational formulation of controlled Neural ODEs.

More precisely, we interpret the generation of normalizing flows as an optimal control problem, where the dynamics are governed by linear-control neural ordinary differential equations. In this setting, the objective is to learn the control functions that induce a flow transporting a source distribution toward a target distribution while minimizing a suitable cost functional.

This perspective leads naturally to a variational learning framework, in which discrete optimal transport plans, data fidelity terms, and regularization components are combined to define an optimization problem over the space of controls. The resulting formulation is closely related to classical optimal control theory and can be solved using techniques such as the Pontryagin Maximum Principle, adjoint equations, and gradient-based updates.

The chapter is structured as follows. Section 4.2 presents the motivation and general idea behind the algorithmic construction. Section 4.3 introduces linear-control Neural ODEs in detail. Section 4.4 connects the framework to optimal transport map approximation. Section 4.5 formulates the variational learning problem, while Section 4.6 derives the optimization proce-

dure and training algorithm. Section 4.7 provides an algorithmic and geometric interpretation, and Section 4.8 discusses the advantages of the proposed approach. Finally, Section 4.9 summarizes the complete algorithmic pipeline and presents the final practical implementation of the proposed method.

4.2 Approximation of Optimal Transport Maps

4.2.1 Practical Scenario

In practical machine learning applications, the exact probability distributions are usually unknown.

Instead, only finite collections of samples are available:

$$\{x_1, \dots, x_N\} \sim \mu,$$

$$\{y_1, \dots, y_M\} \sim \nu.$$

Consequently, the continuous optimal transport problem must be replaced by a discrete approximation.

The proposed framework first computes a discrete optimal transport plan between the empirical distributions and subsequently learns a Neural ODE flow approximating the associated transport map.

4.3 Variational Learning Framework

4.3.1 Discrete Optimal Transport Plan

Let γ_N denote a discrete optimal transport plan between the empirical distributions.

The transport plan establishes correspondences between source samples and target samples and provides the information required for training the controlled flow.

4.3.2 Objective Functional

The optimization problem is formulated through the following cost functional:

$$\mathcal{F}_{N,\beta}(u) = \int_{\mathbb{R}^n \times \mathbb{R}^n} |\Phi_u(x) - y|^2 \, d\gamma_N(x, y) + \frac{\beta}{2} \|u\|_{L^2}^2.$$

The first term measures the transportation error between the transformed samples and the target samples.

The second term acts as a regularization mechanism controlling the energy of the control functions.

4.3.3 Data Fidelity Term

The quantity

$$|\Phi_u(x) - y|^2$$

measures how close the transformed point $\Phi_u(x)$ is to the desired target point y.

Minimizing this term forces the learned flow to reproduce the geometry of the optimal transport map.

4.3.4 Regularization Term

The regularization term

$$\frac{\beta}{2}\|u\|_{L^2}^2$$

penalizes excessively large control signals.

This regularization is essential for several reasons:

- stability of the optimization problem,
- coercivity of the functional,
- smoothness of the learned flow,
- prevention of overfitting.

The parameter $\beta > 0$ controls the trade-off between transport accuracy and control complexity.

4.4 Optimization Procedure

Learning the Controls: The objective of training is to determine the control functions $u(t)$ that minimize the cost functional.

The optimization procedure iteratively updates the controls in order to reduce the discrepancy between transformed samples and target samples.

Pontryagin Maximum Principle: The optimization algorithm is based on the Pontryagin Maximum Principle (PMP), which is a fundamental tool in optimal control theory.

The PMP is applied to the linear-control dynamical system underlying the Neural ODE,

$$\dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t),$$

where $u(t) \in \mathbb{R}^k$ denotes the control input.

The PMP plays a role analogous to backpropagation in deep learning.

The procedure involves:

1. solving the forward linear-control ODE,
2. solving the adjoint backward equation associated with the linear-control system,
3. computing gradients with respect to the control functions $u(t)$,
4. updating the controls iteratively.

This framework allows the optimization of continuous-time linear-control dynamical systems in a principled mathematical manner.

Forward Dynamics: Starting from an initial sample x_0, the state trajectory is computed by integrating the controlled ODE forward in time:

$$\dot{x}(t) = F(x(t))u(t).$$

The terminal point defines the transformed sample.

Backward Adjoint Equation: The adjoint variable $\lambda(t)$ is defined through the Pontryagin Maximum Principle and satisfies the backward-in-time adjoint equation associated with the linear-control system

$$\dot{x}(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t).$$

More precisely, let the Hamiltonian be defined by

$$H(x, \lambda, u) = \left\langle \lambda, \sum_{i=1}^{k} F_i(x)u_i \right\rangle.$$

Then the adjoint equation is given by

$$\dot{\lambda}(t) = -\left(D_x \sum_{i=1}^{k} F_i(x(t))u_i(t) \right)^* \lambda(t),$$

with terminal condition

$$\lambda(T) = \nabla_x \mathcal{J}(x(T)),$$

where $\mathcal{J}$ denotes the objective functional and $(\cdot)^*$ denotes the transpose (adjoint) Jacobian.

This backward equation is coupled with the forward state equation, forming the classical forward{backward system of the Pontryagin Maximum Principle.

This mechanism is conceptually analogous to gradient backpropagation in deep neural networks.

Control Update: After computing the gradients, the controls are updated iteratively through gradient-based optimization methods.

The process is repeated until convergence of the objective functional.

4.5 Algorithmic Interpretation

4.5.1 Geometric Interpretation

The proposed model can be interpreted geometrically as a continuous deformation of one probability distribution into another.

Each sample evolves smoothly through time under the influence of the learned vector field.

The learned dynamics define a transportation flow minimizing both:

- transportation error,
- control energy.

4.5.2 Generative Modeling Interpretation

From a generative modeling perspective, the framework operates as follows:

1. Sample points from a simple base distribution.
2. Propagate the samples through the learned Neural ODE flow.
3. Obtain samples approximating the target distribution.

Therefore, once the flow has been trained, it can generate new samples from the target distribution efficiently.

4.6 Advantages of the Proposed Framework

The proposed methodology presents several important advantages.

- **Continuous-Time Modeling:** The use of Neural ODEs enables smooth continuous transformations of probability distributions.
- **Invertibility:** The generated flows are invertible, which is a fundamental property of normalizing flows.
- **Theoretical Guarantees:** The framework provides rigorous approximation guarantees based on Optimal Transport theory and controllability results.
- **Connection with Optimal Control:** The formulation naturally incorporates tools from optimal control theory, enabling mathematically principled optimization procedures.
- **Generative Capabilities:** The learned flows can be used as generative models capable of producing realistic samples from complex distributions.

4.6.1 Numerical Implementation and Experimental Evaluation

The algorithm described in Algorithm 2 was implemented in `Python` using the `PyTorch` framework for automatic differentiation and neural network optimization, while the discrete optimal transport plan was computed using the `POT` (Python Optimal Transport) library.

The implementation follows exactly the theoretical framework introduced in the previous sections, combining:

- discrete optimal transport computation,
- controlled neural ODE dynamics,
- and gradient-based optimization of the control functions $u(t)$ and vector fields F_l.

Experimental Setup. To evaluate the performance of the proposed method, three different numerical experiments were conducted on synthetic datasets generated from the `moons` (source distribution μ) and `circles` (target distribution ν) datasets.

Each experiment corresponds to a different choice of model capacity and optimization hyperparameters, as summarized below.

Algorithm 2 Approximation of Optimal Transport Maps via Linear-Control Neural ODEs

Input: Source samples $\{x_i\}_{i=1}^N \sim \mu$, target samples $\{y_j\}_{j=1}^M \sim \nu$, initial controls $u^{(0)}(t)$, vector fields $\{F_1, \dots, F_k\}$, regularization parameter $\beta > 0$, learning rate η.

Compute the empirical distributions:

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \qquad \nu_M = \frac{1}{M}\sum_{j=1}^{M} \delta_{y_j}$$

Compute the discrete optimal transport plan:

$$\gamma_N \in \Pi(\mu_N, \nu_M)$$

Initialize iteration counter:

$$m \leftarrow 0$$

while stopping criterion is not satisfied **do**

  for each source sample x_i **do**

    Solve the controlled ODE forward in time:

$$\dot{x}(t) = \sum_{l=1}^{k} F_l(x(t)) u_l^{(m)}(t)$$

    Compute the terminal state:

$$\Phi_{u^{(m)}}(x_i) = x_i(1)$$

  end for

  Compute the loss functional:

$$\mathcal{F}_{N,\beta}(u) = \int |\Phi_u(x) - y|^2 \, d\gamma_N(x, y) + \frac{\beta}{2}\|u\|_{L^2}^2$$

  Solve the adjoint backward equations

  Compute gradients with respect to the controls

  Update the controls:

$$u^{(m+1)} = u^{(m)} - \eta \nabla_u \mathcal{F}_{N,\beta}$$

  Update iteration counter:

$$m \leftarrow m + 1$$

end while

Output: Learned flow map Φ_u

Experiment 1: Baseline Configuration

Hyperparameters:

- Number of vector fields: $k = 4$
- Regularization: $\beta = 10^{-3}$
- Time steps: $T = 25$
- Learning rate: $\eta = 0.01$
- Hidden dimension: 32
- Iterations: 120

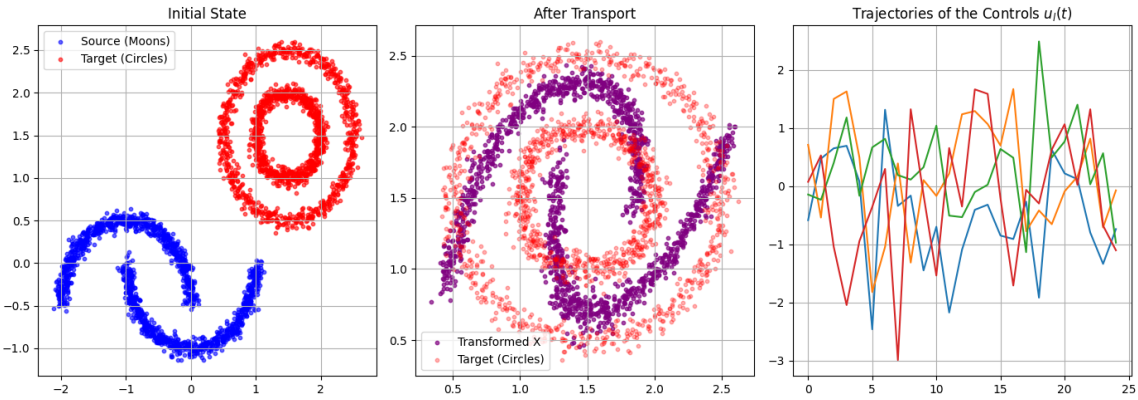

Figure 4.1: Experiment 1: Source distribution, transported samples, and control trajectories.

Discussion.

In this baseline setting, the model is already able to learn a meaningful transport structure between the two distributions. The source moons distribution is successfully deformed into a configuration close to the target circles distribution.

However, the transported samples still exhibit noticeable dispersion and local geometric distortions. This indicates that the model capacity and temporal resolution are still limited.

The control trajectories $u_l(t)$ show high variability, suggesting that the optimization process is still in an exploratory regime and has not fully converged to a stable control strategy.

Experiment 2: Increased Model Capacity

Hyperparameters:

- Number of vector fields: $k = 6$
- Regularization: $\beta = 10^{-4}$
- Time steps: $T = 30$
- Learning rate: $\eta = 0.01$
- Hidden dimension: 64
- Iterations: 300

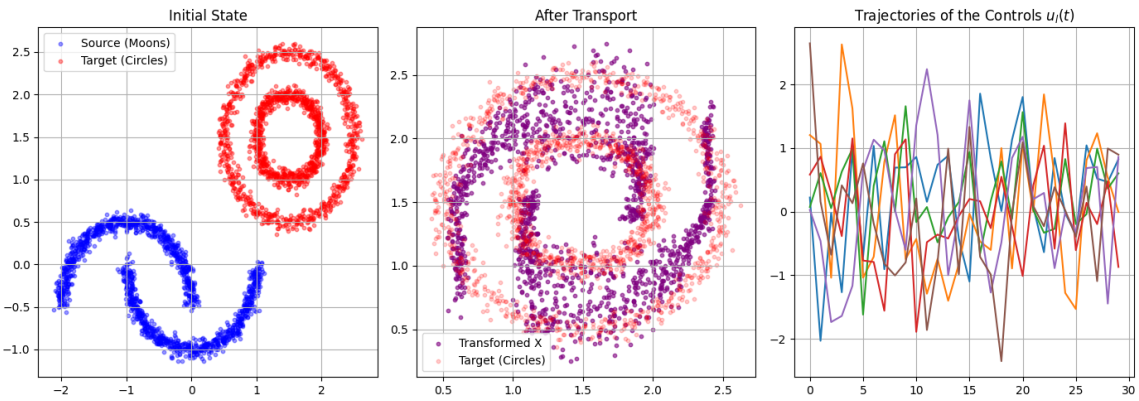

Figure 4.2: Experiment 2: Improved capacity leads to more structured transport.

Discussion.

Increasing both the number of vector fields and the number of integration steps significantly improves the quality of the learned transport map.

The transformed distribution now better aligns with the geometry of the target circles, exhibiting reduced noise and improved global structure preservation.

The control signals become more structured, indicating that the system starts to exploit the temporal dimension more efficiently. This suggests a transition from an unstructured exploratory regime to a more organized control strategy.

Experiment 3: Improved Optimization Stability

Hyperparameters:

- Number of vector fields: $k = 6$
- Time steps: $T = 30$
- Hidden dimension: 64
- Regularization: $\beta = 5 \cdot 10^{-5}$
- Learning rate: $\eta = 0.005$
- Iterations: 450

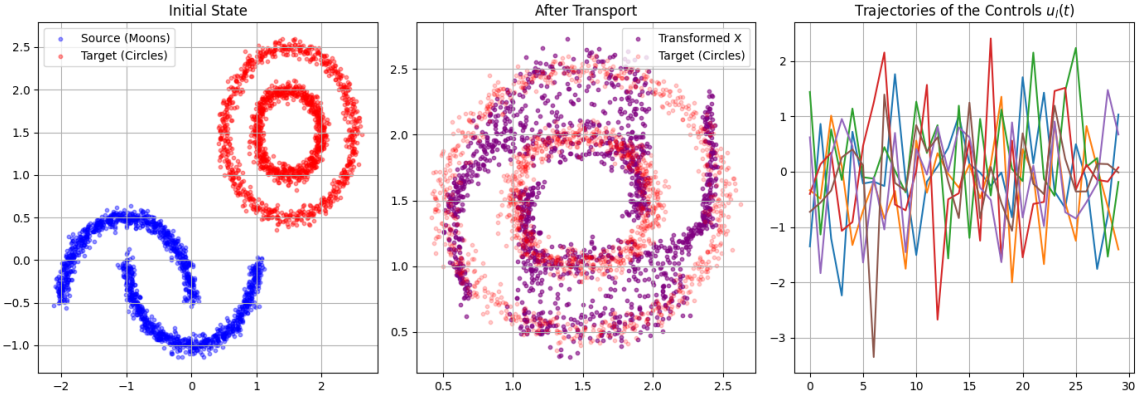

Figure 4.3: Experiment 3: Best performing configuration with improved stability.

Discussion.

The third experiment provides the most stable and geometrically consistent transport behavior.

By reducing the learning rate and the regularization strength while increasing the number of iterations, the optimization process is allowed to converge more smoothly.

The resulting transported distribution closely matches the target geometry, with significantly reduced variance compared to the previous experiments. The control trajectories also appear more stable, indicating a better balance between expressivity and regularization.

Comparison of Experiments

Across the three experiments, a clear trend emerges:

- Increasing the number of vector fields improves the expressivity of the dynamical system.
- Increasing the number of time steps improves the smoothness of the learned transport.
- Stronger optimization (more iterations and lower learning rate) improves stability and convergence.

Overall, the results demonstrate that the proposed linear-control Neural ODE framework is capable of progressively approximating the structure of optimal transport maps when sufficient model capacity and optimization stability are provided.

The evolution from Experiment 1 to Experiment 3 clearly shows a transition from a noisy exploratory flow to a well-structured transport dynamics closely aligned with the geometry of the target distribution.

4.7 Conclusion

This chapter presented the framework of normalizing flows generated through linear-control Neural ODEs for the approximation of optimal transport maps.

The methodology combines Optimal Transport theory, continuous-time deep learning, and control theory into a unified variational framework. The central idea consists of representing transport maps as flows of controlled dynamical systems and learning these flows through optimal control techniques.

The resulting model possesses several desirable properties, including smoothness, invertibility, theoretical approximation guarantees, and generative capabilities.

Furthermore, the framework establishes a mathematically rigorous connection between optimal transport and modern generative modeling, highlighting the growing importance of dynamical systems and control-theoretic approaches in machine learning.

Chapter 5

Discussion and Conclusion

This thesis investigated the interplay between Optimal Transport theory, geometric analysis, and modern generative modeling frameworks. Starting from the classical Monge–Kantorovich problem, we developed the mathematical structure of Wasserstein spaces and examined how transport-based geometry naturally leads to continuous dynamical formulations for probability distributions.

A central theme throughout this work has been the interpretation of generative modeling as a transport problem evolving over the space of probability measures. Within this perspective, several modern architectures—including Normalizing Flows, Continuous Normalizing Flows, Sliced-Wasserstein methods, and Flow Matching techniques—can be understood as different mechanisms for approximating transport dynamics between probability distributions.

The geometric viewpoint provided by Wasserstein theory offers several important advantages. Unlike classical divergence-based approaches, transport-based objectives remain meaningful even when distributions possess disjoint supports, while simultaneously incorporating the geometry of the underlying space. This leads to more stable optimization procedures and provides a natural framework for continuous-time generative dynamics.

Particular emphasis was placed on the approximation of transport maps through controlled Neural Ordinary Differential Equations and Lie-composition-based constructions. These formulations highlight the deep connection between Optimal Transport, control theory, and continuous-time neural architectures. In this setting, generative modeling is no longer interpreted merely as density estimation, but rather as the construction of geometrically structured flows capable of transforming probability measures through learned dynamical systems.

Despite the theoretical strengths of the presented framework, several challenges remain open. Exact Optimal Transport computations remain computationally demanding in high-dimensional settings, while the numerical approximation of Wasserstein gradient flows and

Kantorovich potentials may become unstable for complex data distributions. Furthermore, many theoretical guarantees rely on regularity assumptions that are difficult to fully satisfy in practical machine learning applications.

Future research directions include the development of computationally efficient transport approximations, the study of more expressive flow architectures, and the integration of transport-based methods with stochastic and diffusion-driven generative models. Additional work may also focus on the theoretical analysis of controllability and approximation properties for Neural ODE systems under geometric constraints.

Overall, Optimal Transport provides a mathematically rich and geometrically coherent framework for understanding modern generative modeling. The interaction between transport geometry, dynamical systems, and machine learning continues to reveal deep theoretical connections while simultaneously motivating new computational methodologies for high-dimensional probabilistic modeling.

Chapter 6

Appendix

Implementation Code

Listing 6.1: Neural Optimal Transport Control Framework and Experiments

```
import matplotlib.pyplot as plt
import numpy as np
import ot
import torch
import torch.nn as nn
import torch.optim as optim
from sklearn.datasets import make_circles, make_moons

class NeuralVectorField(nn.Module):
        def __init__(self, input_dim=2, hidden_dim=64):
                super(NeuralVectorField, self).__init__()
                self.net = nn.Sequential(
                        nn.Linear(input_dim, hidden_dim),
                        nn.Tanh(),
                        nn.Linear(hidden_dim, hidden_dim),
                        nn.Tanh(),
                        nn.Linear(hidden_dim, input_dim),
                )

        def forward(self, x):
                return self.net(x)
```

```
def solve_controlled_ode_rk4(x_init, u, vector_fields, num_steps=20):
        dt = 1.0 / num_steps
        x = x_init.clone()

        for step in range(num_steps):
                u_t = u[step]

                def compute_dx(current_x):
                        dx = 0
                        for l, F_l in enumerate(vector_fields):
                                dx += F_l(current_x) * u_t[l]
                        return dx

                k1 = compute_dx(x)
                k2 = compute_dx(x + 0.5 * dt * k1)
                k3 = compute_dx(x + 0.5 * dt * k2)
                k4 = compute_dx(x + dt * k3)

                x = x + (dt / 6.0) * (k1 + 2 * k2 + 2 * k3 + k4)

        return x

def train_advanced_ot_control(
        X_src,
        Y_tgt,
        k_fields=4,
        num_steps=30,
        beta=0.005,
        eta=0.01,
        max_iters=150,
):
        device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
```

```
X_src = X_src.to(device)
Y_tgt = Y_tgt.to(device)

N, d = X_src.shape
M = Y_tgt.shape[0]

vector_fields = [
        NeuralVectorField(input_dim=d).to(device) for _ in range(k_fields)
]

a, b = np.ones((N,)) / N, np.ones((M,)) / M
M_cost = ot.dist(X_src.cpu().numpy(), Y_tgt.cpu().numpy(),
metric="sqeuclidean")
gamma = ot.emd(a, b, M_cost)
gamma_tensor = torch.tensor(gamma, dtype=torch.float32, device=device)

u = torch.randn(num_steps, k_fields, device=device, requires_grad=True)

all_params = [u]
for F_l in vector_fields:
        all_params.extend(list(F_l.parameters()))

optimizer = optim.Adam(all_params, lr=eta)

for m in range(max_iters):
        optimizer.zero_grad()

        X_terminal = solve_controlled_ode_rk4(X_src, u, vector_fields,
        num_steps)

        dist_matrix = torch.cdist(X_terminal, Y_tgt, p=2) ** 2
        transport_loss = torch.sum(dist_matrix * gamma_tensor)

        dt = 1.0 / num_steps
```

```
            reg_loss = (beta / 2.0) * torch.sum(u**2) * dt

            total_loss = transport_loss + reg_loss

            total_loss.backward()
            optimizer.step()

    return u, vector_fields

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

np.random.seed(42)
X_np, _ = make_moons(n_samples=1500, noise=0.05)
Y_np, _ = make_circles(n_samples=1500, noise=0.05, factor=0.5)

X_np = X_np - np.array([1.0, 0.5])
Y_np = Y_np + np.array([1.5, 1.5])

X_src = torch.tensor(X_np, dtype=torch.float32)
Y_tgt = torch.tensor(Y_np, dtype=torch.float32)

HIDDEN_DIM = 64
K_FIELDS = 6
NUM_STEPS = 30
BETA = 0.00005
ETA = 0.005
MAX_ITERS = 450

NeuralVectorField.__init__ = lambda self, input_dim=2, hidden_dim=HIDDEN_DIM: \
    super(NeuralVectorField, self).__init__() or setattr(
        self,
        'net',
            nn.Sequential(
            nn.Linear(input_dim, hidden_dim),
```

```
                nn.Tanh(),
                nn.Linear(hidden_dim, hidden_dim),
                nn.Tanh(),
                nn.Linear(hidden_dim, input_dim)
            )
        )

learned_u, trained_fields = train_advanced_ot_control(
        X_src, Y_tgt,
        k_fields=K_FIELDS,
        num_steps=NUM_STEPS,
        beta=BETA,
        eta=ETA,
        max_iters=MAX_ITERS
)

with torch.no_grad():
        X_transformed = solve_controlled_ode_rk4(
                X_src.to(device),
                learned_u,
                trained_fields,
                num_steps=NUM_STEPS
        ).cpu()

plt.figure(figsize=(14, 5))

plt.subplot(1, 3, 1)
plt.scatter(X_np[:, 0], X_np[:, 1], c="blue", alpha=0.5, s=10)
plt.scatter(Y_np[:, 0], Y_np[:, 1], c="red", alpha=0.5, s=10)
plt.title("Initial State")
plt.legend()
plt.grid(True)

plt.subplot(1, 3, 2)
```

```
plt.scatter(X_transformed[:, 0], X_transformed[:, 1], c="purple", alpha=0.6, s=10)
plt.scatter(Y_np[:, 0], Y_np[:, 1], c="red", alpha=0.2, s=10)
plt.title("After Transport")
plt.legend()
plt.grid(True)

plt.subplot(1, 3, 3)
plt.plot(learned_u.detach().cpu().numpy())
plt.title("Trajectories of the Controls $u_1(t)$")
plt.grid(True)

plt.tight_layout()
plt.show()
```

Bibliography

[1] A. Scagliotti and S. Farinelli, *Normalizing flows as approximations of optimal transport maps via linear-control neural ODEs*, Nonlinear Analysis: Theory, Methods & Applications, vol. 257, p. 113811, 2025. doi: 10.1016/j.na.2025.113811.

[2] R. Jordan, D. Kinderlehrer, and F. Otto, *The Variational Formulation of the Fokker-Planck Equation*, SIAM Journal on Mathematical Analysis, vol. 29, no. 1, pp. 1–17, 1998.

[3] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, *Normalizing Flows: An Introduction and Review of Current Methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[4] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Birkhäuser, 2015.

[5] A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F.-R. Stöter, *Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions*, arXiv:1905.13117, 2020.

[6] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde, *Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model*, arXiv:1804.01947, 2019.

[7] Y.-J. Huang and Z. Malik, *Generative Modeling by Minimizing the Wasserstein-2 Loss*, 2025.

[8] AMS Word, *A Simple Introduction on Sinkhorn Distances*, Medium, 2020. Available at: `https://amsword.medium.com/a-simple-introduction-on-sinkhorn-distances-d01a4ef4f085`

[9] P.W.Y. Lee, *On the Jordan-Kinderlehrer-Otto Scheme*, 2015.

[10] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, Advances in Neural Information Processing Systems (NeurIPS), 2013.

[11] D. Haviv, A.-A. Pooladian, D. Pe'er, and B. Amos, *Wasserstein Flow Matching: Generative Modeling Over Families of Distributions*, arXiv preprint, 2024.

[12] S. Liu, X. Zhou, Y. Jiao, and J. Huang, *Wasserstein Generative Learning of Conditional Distribution*, 2021.

[13] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, Revised Edition, CRC Press, 2015.