NUMERICAL COMPUTATION AND ANALYSIS RELATED TO OPTIMAL TRANSPORT THEORY

A Dissertation
Presented to
The Academic Faculty

By

Shu Liu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics
& School of Computational Science and Engineering

Georgia Institute of Technology

May 2022

NUMERICAL COMPUTATION AND ANALYSIS RELATED TO OPTIMAL TRANSPORT THEORY

Thesis committee:

Dr. Luca Dieci
School of Mathematics
Georgia Institute of Technology

Dr. Xiaojing Ye
Department of Mathematics and Statistics
Georgia State University

Dr. Sungha Kang
School of Mathematics
Georgia Institute of Technology

Dr. Haomin Zhou, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Molei Tao
School of Mathematics
Georgia Institute of Technology

Date approved: March 23rd, 2022

To *my parents* and *grandparents*.

ACKNOWLEDGMENTS

During the years at Georgia Tech, I have conducted researches in applied and computational mathematics under the support from many friends. This thesis won't be completed without the great help of the people around me.

Firstly, I would like to express my deepest gratitude to my advisor, professor Haomin Zhou, who leads me to the wonderful mathematical world related to optimal transport. I feel really fortunate to get the chance to explore mathematical problems that I have great interests under the guide of Prof. Zhou. I still remember the days that I spent the entire afternoon in his office discussing all kinds of challenging problems encountered in my research. Those enjoyable discussions have become one of my greatest memories of PhD study. I am also very grateful for receiving many beneficial academic suggestions from him. In addition, I should also thank his consistent care and encouragement during the pandemic time.

I should also express my sincere thank to Prof. Hongyuan Zha. Interaction with him exposes me to the vibrant machine learning community and brings a lot of inspirations to my research; I also thank Prof. Wuchen Li who had shared a lot of his unique understanding and ideas about optimal transport with me. I am also grateful to his nice suggestions on writing scientific articles and academic presention. I also thank Prof. Yongxin Chen and Prof. Xiaojing Ye for their interests in my research and collaborations. I have learned a lot from the insightful discussions with them.

I would like to thank Prof. Luca Dieci, Sung Ha Kang, Molei Tao and Xiaojing Ye for serving as committee members of my thesis defense. I want to thank all the professors who have taught me, as well as all the staff members who have assisted me during my Ph.D. study. I should also thank to Dr. Klara Grodzinsky and Dr. Mo Burke who have supported my teaching in the School of Math.

I also feel fortunate to cooperate with several graduate students and postdocs at Georgia

Tech, I learned a lot about Hamiltonian systems and stochastic analysis from Dr. Jianbo Cui through our cooperation; I also benefited a lot from the cooperation with Haodong Sun, Shaojun Ma and Jiaojiao Fan, who have shared a lot of their precious experiences on coding and deep learning with me.

I also would like to thank Adrian Bustamante, Yibin Ding, Ruilin Li, Hao Wu, Xin Xing, Yaohua Zang and Haoyan Zhai for having discussions on mathematics during my Ph.D. study. I thank Prof. Molei Tao for suggesting, and Renyi Chen, Yuchen He, Ben Ide, Hao Liu and many more for organizing and contributing to the applied math seminar for graduate students, which has dramatically expended my research horizons.

I am also happy to make friends during my graduate study. In addition to the aforementioned people, I should also thank Tongzhou Chen, Guangyu Cui, He Guo, Salem Jad, Yu Jing, Daniyar Omarov, Yichen Sun, Yuqing Wang, Shijie Xie, Jiaqi Yang, Yian Yao, Weiwei Zhang, and many more for their friendships. I especially thank Guangyu Cui and Salem Jad as well as Shaojun Ma for providing me with a temporal residence during the pandemic. I also express my gratitude to Guangyu Cui, Salem Jad, Haodong Sun and Hao Wu for their kind help during the pandemic. I wish my friends all the best in the future.

Last but not least, I sincerely thank my parents and grandparents for their selfless support and love. I always feel guilty of not having the chance to repay their love, but I get the chance this time: To them, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis presents several research projects related to optimal transport (OT) theory. The main part of this thesis consists of three sections.

- In the first section, we focus on solving OT problems from three different perspectives: (1) direct approximation of the optimal transport map in high dimensions; (2) particle evolving method for generating samples from the optimal transport plan; (3) learning high dimensional geodesics joining two known distributions. These three perspectives focus on various aspects of OT problems, which may find their own applications under distinct settings in diverse branches of data science and machine learning. We derive sample-based algorithms for each project. Our methods are supported by theoretical guarantees and numerical justifications on both synthetic and realistic data sets.

- In the second section, we develop and analyze a sampling-friendly method for high dimensional Fokker-Planck equations by leveraging the generative models from deep learning. By utilizing the fact that the Fokker-Planck equation can be viewed as gradient flow on probability manifold equipped with certain OT distance, we derive an ordinary differential equation (ODE) on parameter space whose parameters are inherited from the generative models. We design a variational semi-implicit scheme for solving the proposed ODE. Moreover, we establish bounds for both the convergence analysis and error analysis for our method. Several numerical examples are provided to illustrate the performance of the proposed algorithms and analysis.

- In the third section, we present a definition of Hamiltonian process on finite graph via its corresponding density dynamics on Wasserstein manifold. We demonstrate the existence of such Hamiltonian process in many classical discrete problems, such as the OT problem, Schrödinger equation and Schrödinger bridge problem (SBP). The stationary and periodic properties of Hamiltonian processes are investigated in the framework of SBP.

CHAPTER 1
INTRODUCTION

Optimal transport (OT) problem was initially introduced as a constrained optimization problem [1, 2] seeking for the optimal transport plan to move mass from initial to target positions with minimum cost. Since then, OT problems become a classical topic in optimization, probability theory and economics. In recent decades, mathematicians had discovered nice geometric structures of optimal transport [3, 4]. This leads to an elegant interplay between optimal transport, partial differential equations [5, 4], fluid dynamics [6], and differential geometry [7]. On the other hand, optimal transport distance itself plays a significant role on measuring the discrepancy between distributions due to its symmetric and robust properties. Because of this reason, in recent years, optimal transport has found its widespread applications in various disciplines like data science [8], economy[9], imaging science [10], and seismology [11].

1.1 Computational problems related to optimal transport

Due to the great importance of OT distance, we are motivated to develop computational tools for OT problems. Although there exist series of publications [12, 13, 14, 15, 16, 17, 18, 19] on computing both discrete and continuous OT problems, they are still challenging tasks in many applications. In the first part of the thesis, we propose three novel OT-related algorithms with different purposes under distinct problem settings.

- **Neural network-based approximation of the Monge map T_***

  The original version of OT, which is known as the Monge problem, aims at finding the cost-minimizing map T_* (known as the Monge map) that transports a given distribution μ to the desired distribution ν. In [20], we present a scalable algorithm to directly ap-

proximate T_* via neural networks. By introducing Lagrange multiplier, we formulate a max-min saddle scheme that only requires samples from μ, ν. Such saddle problem can be well resolved by Stochastic Gradient Descent (SGD)-typed methods. The algorithm is capable of computing OT problems with general cost c between μ, ν with different dimensions. The numerical error of such method can also be estimated via certain duality gaps that come from our algorithm.

This is a work in collaboration with Jiaojiao Fan, Shaojun Ma, Yongxin Chen and Haomin Zhou. I mainly worked on the derivation and the theoretical analysis of the method. I also worked on several low dimensional experiments.

- **Sample-based approximation of the optimal coupling γ_***

  A crucial relaxed formulation of the Monge problem is known as the Kantorovich problem. Instead of the map T_*, it seeks for the optimal coupling γ_* whose marginals are μ, ν. In [21, 22], we propose an innovative algorithm that computes samples from γ_* which encodes rich statistical information. Our algorithm uses the 2-Wasserstein gradient flow derived from the Entropy Transport Problem, which can be treated as the Kantorovich problem with soft marginal constraints. We realize our algorithm by evolving an associated interacting particle system so that the empirical distribution of the particles gradually converges to an approximation of γ_*. Our method is supported by theoretical justification and numerical verification.

  This is a work in collaboration with Haodong Sun, and Hongyuan Zha. I mainly focused on the derivation and the theoretical analysis of the method.

- **Computing geodesic between probability distributions μ and ν**

  Most of the existing treatments for OT problem focus on the static aspect. We are interested in the problem of interpolating between μ and ν using the action-minimizing curve: In [23], We consider an optimal control problem on probability space whose solution $\{\rho_t\}$ can be regarded as the geodesic joining μ and ν. By applying Lagrange

multiplier method and deriving a min-max saddle problem, we design a deep learning strategy that further leads to a sample based algorithm for solving geodesic joining μ, ν in high dimensional space. Our proposed method enables sampling from the geodesic $\{\rho_t\}$ at arbitrary time t. The algorithm also computes the OT distance together with the Monge map as by-products. The performance of our method is demonstrated through a series of experiments on both synthetic and real-world data.

This is a work in collaboration with Shaojun Ma, Yongxin Chen and Haomin Zhou. I mainly worked on the derivation and the theoretical analysis of the method. I also worked on several 2D-3D experiments including color transfer, as well as several experiments on the MNIST data set.

1.2 Computation of high dimensional Fokker-Planck equations via parametrized pushforward map

In the work [24, 25], we develop and analyze numerical methods for high dimensional Fokker-Planck equations by leveraging generative models from deep learning. Our starting point is a formulation of the Fokker-Planck equation as a system of ordinary differential equations (ODEs) on finite-dimensional parameter space with the parameters inherited from generative models such as normalizing flows. We call such ODEs *neural parametric Fokker-Planck equations*. The fact that the Fokker-Planck equation can be viewed as the 2-Wasserstein gradient flow of relative entropy functional (also known as Kullback–Leibler (KL) divergence) allows us to derive the ODEs as the constrained 2-Wasserstein gradient flow of relative entropy on the set of probability densities generated by neural networks. For numerical computation, we design a novel bi-level minimization strategy for semi-implicit time discretization scheme of the proposed ODE. Such an algorithm is sampling-based, which can readily handle the Fokker-Planck equations in higher dimensional spaces. Moreover, we also establish bounds for the asymptotic convergence analysis of the neural parametric Fokker-Planck equation as well as the error analysis for both the continuous and

discrete versions. Several numerical examples are presented to illustrate the performance of the proposed algorithms and numerical analysis results.

This is a work in collaboration with Wuchen Li, Hongyuan Zha and Haomin Zhou, I worked on the majority part of this research project including the derivation, numerical analysis and experiments of the proposed method.

1.3 Hamiltonian process on finite graphs via Wasserstein Hamiltonian theory

Hamiltonian system is ubiquitous in the physical world and has been well studied for the continuous space. However, very few work has been done on its counterpart in discrete space such as graphs due to the lack of geometric structures. In [26], motivated by the previous researches on Markovian process on graphs [27, 28], we present a strategy to tackle this challenge by using the Wasserstein Hamiltonian flows on the cotangent bundle of the probability space $\mathcal{P}(G)$ defined on graph G. We define a stochastic process as Hamiltonian process on G if its time varying density can be written as a Wasserstein Hamiltonian flows. Furthermore, we demonstrate the existence of such Hamiltonian process in many classical discrete problems.

This is a work in collaboration with Jianbo Cui and Haomin Zhou. I participated in proposing the concept of Hamiltonian process on graph, I also worked on the examples as well as the theoretical results in this article.

CHAPTER 2

PRELIMINARY KNOWLEDGE

In this chapter, we briefly introduce some preliminary knowledge needed for the thesis. This contains two parts: (1) An overview of the optimal transport problems, and (2) Basic knowledge of Wasserstein manifold.

2.1 Optimal transport problems

Before our discussion, we first introduce several common notations that we will use throughout the thesis. Suppose X is a measurable space. We denote $\mathcal{P}(X)$ as the space of probability distributions defined on X. For measurable spaces X, Y, $\mu \in \mathcal{P}(X)$, and measurable map $T : X \to Y$, we define $T_\sharp \mu \in \mathcal{P}(Y)$ as

$$T_\sharp \mu(E) = \mu(T^{-1}(E)), \quad \text{for any measurable set } E \subset Y. \tag{2.1}$$

We will call $T_\sharp \mu$ as the **pushforward of μ by T**. We will also call T as the **pushforward map**.

2.1.1 Monge problem

Optimal transport (OT) problem was initially formalized by the mathematician Gaspard Monge in [1]. To formulated this problem, we consider X, Y as two measurable spaces. Assume $c(\cdot, \cdot) : X \times Y \to \mathbb{R}$ is a measurable function defined on $X \times Y$. We treat it as the **cost function**: $c(x, y)$ quantifies the effort of moving one unit of mass from location x to location y. Now given two distributions $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, we consider the following

problem on minimizing the average cost

$$C_{\text{Monge}}(\mu,\nu) \triangleq \min_{T} \left\{ \int c(x,T(x))d\mu(x) \right\} \quad \text{(MP)} \tag{2.2}$$

over the set of all measurable maps $T : X \to Y$ such that $T_\sharp \mu = \nu$.

We call (Equation 2.2) as **Monge problem(MP)**. An intuitive illustration of (MP) is shown

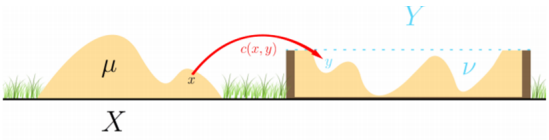


Figure 2.1: Illustration of Monge problem: Filling the pit (with distribution ν) by the pile of sand (with distribution μ) while minimizing the total transport cost.
Source of the image: https://medium.com/analytics-vidhya

in Figure 2.1. In our research, we will mainly focus on the optimal transport on Euclidean space, i.e., we will treat $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ for our future discussion.

It is natural to ask about the existence and the uniqueness of the optimal solution to (Equation 2.2). Such problems have been studied by many scholars, and we refer readers to chapter 9 and chapter 10 of [7] and the references therein for detailed discussions. For the sake of completeness in this thesis, we state the following useful result, which is a simplified version of a series of theorems quoted from Chapter 10 of [7]. Let us consider the following conditions on cost function $c(\cdot,\cdot)$.

$$\text{There exists } a \in L^1(\mu),\ b \in L^1(\nu),\ \text{such that}\ \ c(x,y) \geq a(x) + b(y); \tag{2.3}$$

$$c(\cdot,\cdot) \text{ is } \textbf{locally Lipschitz} \text{ and } \textbf{superdifferentiable} \text{ everywhere;} \tag{2.4}$$

$$\partial_x c(x,\cdot) \ \text{ is injective for any } x \in \mathbb{R}^n. \tag{2.5}$$

We state the definitions for locally Lipschitz and superdifferentiablity in the appendix. We then have the following result.

Theorem 2.1.1 (Existence and uniqueness of the Monge map)**.** *Suppose the cost function* $c(\cdot,\cdot)$ *satisfies* (Equation 2.3)*,* (Equation 2.4)*,* (Equation 2.5)*, we assume that* μ *and* ν *are compactly supported and* μ *is absolute continuous with respect to the Lebesgue measure* $\mathscr{L}^n$ *on* $\mathbb{R}^n$ *(In the following discussion of this thesis, we use the notation* $\mu \ll \mathscr{L}^n$ *for absolute continuity.). Then there exists a unique transport map* T_* *solving the Monge problem. We call* T_* *the* ***Monge map*** *for the problem* (Equation 2.2)*.*

2.1.2 Kantorovich problem

The Monge Problem aims at finding an optimal map that maps μ to ν. But this requirement is too restrictive, one can design many counterexamples in which such maps don't exist. For example, if we choose $\mu = \delta_0$ as the Dirac distribution concentrated at 0 and $\nu = \mathcal{N}(0, I)$ as standard Gaussian distribution. Since one cannot break particles, we can never find T that pushforward the point measure δ_0 to $\mathcal{N}(0, I)$.

In order to generalize the Monge problem, Leonid Kantorovich proposed the following relaxed formulation [2]

$$C(\mu,\nu) \triangleq \min_{\gamma} \left\{ \iint_{X\times Y} c(x,y)\, d\gamma(x,y) \right\} \quad \text{(KP)} \tag{2.6}$$

over the set of joint probability distributions $\gamma \in \Pi(\mu,\nu)$.

Here we define $\Pi(\mu,\nu)$ as the set of joint distributions with fixed marginals

$$\Pi(\mu,\nu) = \left\{ \gamma \in \mathcal{P}(X\times Y) \;\middle|\; \begin{array}{l} \gamma(A\times Y) = \mu(A),\ \gamma(X\times B) = \nu(B) \\ \text{for any measurable set } A \subset X, B \subset Y. \end{array} \right\}.$$

In stead of searching for the map, we compute an **optimal coupling** γ_* with marginal distributions μ, ν. Such γ_* is also called **optimal transport plan** between μ, ν. We call this relaxed formulation as **Kantorovich problem(KP)**. The formulation allows us to break the single particle into pieces and then transport each piece to certain positions according

to γ_*. Thus we are able to deal with OT problem under much more general settings. Some examples of optimal couplings are shown in Figure 2.2.

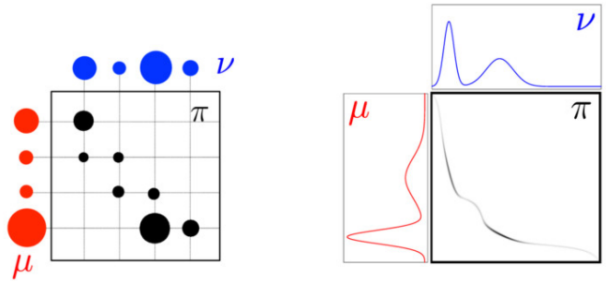


Figure 2.2: Optimal coupling γ_* for discrete point measure (left), and continuous measure (right). One can tell $\gamma_* = (\mathrm{Id}, T_*)_\sharp \mu$ for the continuous case, i.e., the support of γ_* belongs to the graph of Monge map T_*. The figure is taken from [10].

We summarize the existence result of optimal coupling and its relation with Monge map in the following theorem. Its proof can be found in [29, 7].

Theorem 2.1.2 (Existence & uniqueness of optimal coupling, relation with Monge map)**.** *Suppose the cost function $c(\cdot, \cdot)$ is lower semi-continuous and satisfies* (Equation 2.3)*. Then there exists an optimal coupling $\gamma_* \in \Pi(\mu, \nu)$ that solves (KP)* (Equation 2.6)*.*

If we assume $X = Y = \mathbb{R}^d$ and the cost function takes the form $c(x, y) = h(x - y)$, where h is a strictly convex function, suppose $\mu \ll \mathscr{L}^d, \nu \ll \mathscr{L}^d$. Then the optimal solution to (Equation 2.6) *exists and is also unique.*

If we assume that the Monge map T_* exists for the corresponding (MP) (Equation 2.2)*, then $(Id, T_*)_\sharp \mu \in \Pi(\mu, \nu)$ is an optimal coupling of (KP)* (Equation 2.6)*, and $C(\mu, \nu) = C_{Monge}(\mu, \nu)$.*

We will denote the minimum value of the Kantorovich problem (Equation 2.6) as $C(\mu, \nu)$. We call $C(\mu, \nu)$ as the **optimal transport (OT) distance** or **Wasserstein distance** between μ and ν with respect to the cost function c.

Theorem 2.1.3 ($C(\cdot, \cdot)$ as distance function)**.** *Suppose $X = Y = \mathbb{R}^d$ and $c(\cdot, \cdot)$ is distance function on $\mathbb{R}^d \times \mathbb{R}^d$. Then $C(\cdot, \cdot)$ is also a distance function on $\mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$.*

We give the definition of distance function in Definition A.0.3 in the Appendix. A direct proof to this theorem can be found in chapter 6 of [7].

Example 2.1.1 (p-Wasserstein distance)**.** *Suppose $X = Y = \mathbb{R}^d$, if we set $c(x, y) = |x-y|^p$ with $p \geq 1$, then we denote the corresponding OT distance as $W_p(\mu, \nu)^p$, we call W_p as p-**Wasserstein distance**. Among these distance functions, W_1 distance (also known as Earth mover's distance) has found many important applications in image processing [30] and machine learning [8]; W_2 distance can imply Riemannian geometry structure on the probability manifold. This will be discussed in details in subsection 2.2.1.*

2.1.3 Kantorovich dual problem

We can treat Kantorovich problem (Equation 2.6) as a linear optimization problem with two marginal constraints and nonnegative constraint of γ. It is thus natural to consider the Lagrange multiplier method for (Equation 2.6), i.e., we introduce Lagrange multiplier functions ψ, ϕ, and $\sigma \geq 0$ for the constraints

$$\int_Y \gamma(\cdot, y)\, dy = \mu(\cdot), \quad \int_X \gamma(x, \cdot)\, dx = \nu(\cdot), \quad \gamma \geq 0$$

and we obtain the max-min problem of (KP) as

$$\max_{\psi,\, \phi,\, \sigma \geq 0} \min_{\gamma} \left\{ \iint_{X \times Y} c(x, y)\gamma(x, y)\, dxdy + \iint_{X \times Y} \psi(x)(\gamma(x, y) - \mu(x))\, dxdy \right. \tag{2.7}$$

$$\left. - \iint_{X \times Y} \phi(y)(\gamma(x, y) - \nu(y))\, dxdy - \iint_{X \times Y} \sigma(x, y)\gamma(x, y)\, dxdy \right\} \tag{2.8}$$

This max-min problem is equivalent to the maximization problem

$$\begin{aligned}
&\max_{\psi,\ \phi,\ \sigma\geq 0}\ \min_{\gamma}\left\{\iint_{X\times Y}(c(x,y)+\psi(x)-\phi(y)-\sigma(x,y))\gamma(x,y)\ dxdy\right.\\
&\left.+\int_Y \phi(y)d\nu(y)-\int_X \psi(x)d\mu(x)\right\}\\
=&\max_{\substack{\psi,\ \phi,\ \sigma\geq 0\\ c(x,y)+\psi(x)-\phi(y)-\sigma(x,y)=0}}\left\{\int_Y \phi(y)\nu(x)\ dy-\int_X \psi(x)d\mu(x)\right\}
\end{aligned}$$

We can then reformulate it as

$$K(\mu,\nu)\triangleq\max_{\phi(y)-\psi(x)\leq c(x,y)}\left\{\int_Y \phi(y)d\nu(y)-\int_X \psi(x)d\mu(x)\right\}\quad\text{(dual-KP)}. \tag{2.9}$$

We call the maximization problem (Equation 2.9) as **Kantorovich dual problem** (dual-KP). Actually, for arbitrary ψ, we can fix function ϕ in (Equation 2.9) as the optimal one, i.e., we denote

$$\psi^{c,+}(y)=\inf_{x\in X}\ \{\psi(x)+c(x,y)\}\quad\forall\ y\in Y. \tag{2.10}$$

And (Equation 2.9) can be reformulated as

$$\max_{\psi}\left\{\int_Y \psi^{c,+}(y)d\nu(y)-\int_X \psi(x)d\mu(x)\right\}. \tag{2.11}$$

Similarly, we denote

$$\phi^{c,-}(x)=\sup_{y\in Y}\ \{c(x,y)-\phi(y)\}\quad\forall\ x\in X.$$

And (Equation 2.9) can also be reformulated as

$$\max_{\phi}\left\{\int_Y \phi(y)d\nu(y)-\int_X \phi^{c,-}(x)d\mu(x)\right\}. \tag{2.12}$$

One can actually prove the equivalence between (dual-KP) and (KP) [29, 7]. This is summarized in the following theorem.

Theorem 2.1.4 (Equivalence between primal and dual problem)**.** *Suppose c is a cost function defined on $X \times Y$ and satisfies* (Equation 2.3)*. Recall $C(\mu, \nu)$, $K(\mu, \nu)$ defined as the optimal values in* (Equation 2.6)*,* (Equation 2.9)*. We have $C(\mu, \nu) = K(\mu, \nu)$.*

Furthermore, we can characterize the Monge map T_* as well as the optimal coupling γ_* by the optimal dual variables ψ_*, ϕ_*, σ_*. To simplify our discussion, we consider the case in which the Monge map T_* and optimal coupling γ_* uniquely exist. And thus by Theorem 2.1.2, $\gamma_* = (\mathrm{Id}, T_*)_\sharp \mu$. We now consider the Karush–Kuhn–Tucker (KKT) condition [31] of (Equation 2.8), this gives

$$c(x, y) + \psi_*(x) - \phi_*(y) - \sigma_*(x, y) = 0,$$
$$\sigma_* = 0, \text{ on } \mathrm{Spt}(\gamma_*).$$

Combining both conditions gives

$$c(x, y) + \psi_*(x) - \phi_*(y) = 0 \quad \text{on } \mathrm{Spt}(\gamma_*). \tag{2.13}$$

Now since $\gamma_* = (\mathrm{Id}, T_*)_\sharp \mu$, thus $\mathrm{Spt}(\gamma_*) = \{(x, T_*(x)) | x \in \mathrm{Spt}(\mu)\}$. Use (Equation 2.13), for any $x \in \mathrm{Spt}(\mu)$, $\phi_*(T_*(x)) - \psi_*(x) = c(x, T_*(x))$. On the other hand, since $\phi_*(y) - \psi_*(x) \leq c(x, y)$, for any fixed $x \in \mathrm{Spt}(\mu)$, $y = T_*(x)$ is the maximizer of $\phi_*(y) - \psi_*(x) - c(x, y)$(w.r.t. y). Under certain differentiable assumption, the gradient w.r.t. y should vanish at $y = T_*(x)$. This leads to

$$\nabla \phi_*(T_*(x)) - \partial_y c(x, T_*(x)) = 0, \quad \text{for } x \in \mathrm{Spt}(\mu). \tag{2.14}$$

In addition, for any fixed $y = T_*(x)$, using the similar argument, we obtain

$$-\nabla\psi_*(x) - \partial_x c(x, T_*(x)) = 0, \quad \text{for } x \in \text{Spt}(\mu). \tag{2.15}$$

One can also verify the equivalence between (Equation 2.14) and (Equation 2.15) using the fact that for optimal dual pair (ψ_*, ϕ_*) satisfies $\phi_* = \psi_*^{c,+}$ or $\psi_* = \phi_*^{c,-}$. Now (Equation 2.15) directly characterizes the Monge map T_* via

$$T_*(x) = \partial_x c(x, \cdot)^{-1}(-\nabla\psi_*(x)), \quad \text{for } x \in \text{Spt}(\mu). \tag{2.16}$$

The above derivation can be made rigorous in the following theorem. It is a simplified version of Theorem 10.28 combined with Remark 10.33 taken from [7].

Theorem 2.1.5 (Characterization of Monge map and optimal coupling)**.** *Under the same conditions stated in Theorem 2.1.1, there exists unique Monge map T_* that solves (MP)* (Equation 2.2) *and unique γ_* solving (KP)* (Equation 2.6) *with $\gamma_* = (Id, T_*)_\sharp\mu$. Furthermore, there exists unique ψ_*, ϕ_* solving the (dual-KP)* (Equation 2.9)*, or equivalently, ψ_*, ϕ_* uniquely solve* (Equation 2.11)*,* (Equation 2.12)*. Then ψ_*, ϕ_* are differentiable on Spt(μ), Spt(ν). And T_*, ψ_*, ϕ_* satisfies* (Equation 2.15)*,* (Equation 2.14)*.*

Example 2.1.2. *When we pick $X = Y = \mathbb{R}^d$ and $c(x, y) = \frac{1}{2}|x - y|^2$. Assume one of the optimal duals is ψ_*, then by* (Equation 2.16)*, the Monge map has the form*

$$T_*(x) = x + \nabla\psi_*(x). \tag{2.17}$$

2.1.4 Monge map and Monge-Ampère equation

We consider $X = Y = \mathbb{R}^d$. Let us recall under suitable conditions, the Monge map T_* always possess the form (Equation 2.16). On the other hand, we know T_* must pushforward distribution μ to ν, i.e. $T_{*\sharp}\mu = \nu$. Suppose the density function of μ, ν are ρ_a, ρ_b, then the

density function of $T_{*\sharp}\mu$ can be written as $\frac{\rho_a(T_*^{-1}(\cdot))}{\det(DT_*(T_*^{-1}(\cdot)))}$, where DT_* denotes the Jacobian of map T_*. Then we obtain the equation

$$\frac{\rho_a(T_*^{-1}(y))}{\det(DT_*(T_*^{-1}(y)))} = \rho_b(y).$$

Now replace y by $T_*(x)$ and use $T_*(x) = \partial_x c(x,\cdot)^{-1}(-\nabla\psi_*(x))$, we arrive at the following equation of ψ_*

$$\rho_a(x) = \det(D(\partial_x c(x,\cdot)^{-1}(-\nabla\psi_*(x)))) \cdot \rho_b(\partial_x c(x,\cdot)^{-1}(-\nabla\psi_*(x))). \tag{2.18}$$

This is a nonlinear second order partial differential equation known as the Monge-Ampère equation [32, 33]. From the discussion above, combining the equation (Equation 2.18) for ψ_* and (Equation 2.16) together, one can obtain the Monge map T_*.

Example 2.1.3. *When $c(x,y) = \frac{1}{2}|x-y|^2$, the equation* (Equation 2.18) *reduce to the classical form*

$$det(I_d + \nabla^2\psi(x))\rho_b(x + \nabla\psi(x)) = \rho_a(x). \tag{2.19}$$

Here I_d denotes the $d \times d$ identity matrix.

2.1.5 Dynamical formulation of optimal transport problem

From now on, let us restrict our discussion on optimal transport problems on the same space $X = \mathbb{R}^d$. Both Monge problem (Equation 2.2) and Kantorovich problem (Equation 2.6) can be treated as static optimal transport problems, which do not involve any time evolutional dynamics. It is insightful to generalize the static OT problem to dynamical versions.

Let us consider the following optimal control problem with boundary constraints

$$C_{\text{Dym}}(\rho_a, \rho_b) \triangleq \min_{\rho, v} \left\{ \int_0^1 \int_{\mathbb{R}^d} L(v(x,t))\ \rho(x,t)\ dxdt \right\}, \quad \text{(Dym-OT)} \tag{2.20}$$

$$\text{subject to: } \frac{\partial \rho(x,t)}{\partial t} + \nabla \cdot (\rho(x,t) v(x,t)) = 0, \tag{2.21}$$

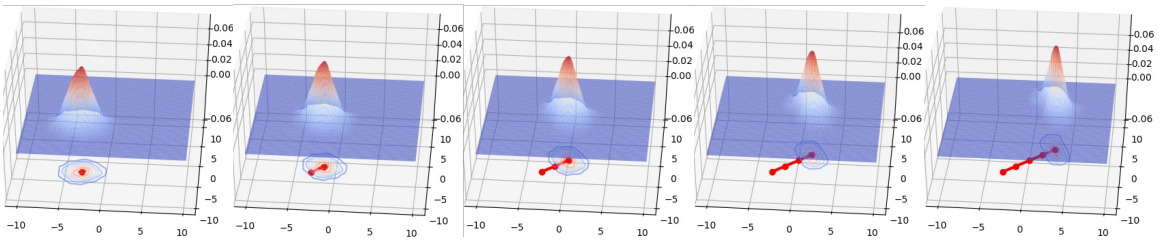
$$\text{and } \rho(\cdot, 0) = \rho_a,\ \rho(\cdot, 1) = \rho_b. \tag{2.22}$$

This is an optimal control problem on $\mathcal{P}(\mathbb{R}^d)$. It is first proposed by Jean-David Benamou and Yann Brenier in [6]. We call (Equation 2.20) as the **Dynamical OT problem (Dym-OT)**. Here we define the cost function $L(\cdot)$ as

Definition 2.1.1. *We define $L(\cdot) \in C^1(\mathbb{R}^d)$ as the **Lagrangian** of the control problem* (Equation 2.20)*. We always assume that L satisfies $L(-u) = L(u)$ for arbitrary $u \in \mathbb{R}^d$ and is strictly convex and superlinear, i.e., $L(\cdot)$ is super linear if $\lim_{u\to\infty} \frac{L(u)}{|u|} = \infty$.*

One can either treat L as the transporting cost or as the kinetic energy of the transport motion. The first constraint (Equation 2.21) is the **continuity equation** of ρ, which describe the density evolution of ρ along the flow field $v(\cdot, t)$. Thus the goal of Dynamical OT problem is to find an optimal way to continuously transfer density ρ_a to ρ_b with minimum average transport cost. The optimal solution $\{\rho(\cdot, t)\}_{0\le t\le 1}$ can be treated as an cost-minimizing interpolation between ρ_a and ρ_b. We also call the optimal $\{\rho(\cdot, t)\}$ as the **geodesic** (w.r.t. the cost L) joining ρ_a and ρ_b. An example of dynamical OT problem between Gaussian distributions is presented in Figure 2.3.

In addition to the PDE formulation (Equation 2.20)(Equation 2.21)(Equation 2.22) on (Dym-OT), we also have the equivalent particle control formulation on Dynamical OT



(a) t=0 (b) t=0.25 (c) t=0.5 (d) t=0.75 (e) t=1

Figure 2.3: The optimal solution to a dynamical OT problem between two Gaussian distributions with $L(v) = \frac{|v|^2}{2}$; One can tell that each particle is moving in straight line with constant velocity.

problem

$$\min_{v} \left\{ \int_0^1 \mathbb{E}[L(v(\boldsymbol{X}_t, t))]\, dt \right\}, \quad \text{(p-Dym-OT)} \tag{2.23}$$
$$\text{subject to: } \frac{d}{dt}\boldsymbol{X}_t = v(\boldsymbol{X}_t, t),$$
$$\text{and } \boldsymbol{X}_0 \sim \rho_a,\ \boldsymbol{X}_1 \sim \rho_b.$$

In order to study the optimal solution to Dynamical OT problem (Equation 2.20), we introduce Lagrange Multiplier $\Phi(\cdot, t)$ for constraint (Equation 2.21), and $\Psi_a,\ \Psi_b$ for the boundary constraints (Equation 2.22). Then we consider the functional

$$\begin{aligned}
&\mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b) \qquad (2.24)\\
&= \int_0^1 \int L(v)\,\rho + \left(\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v)\right) \Phi(x,t)\, dxdt \\
&\quad + \int \Psi_a(x)(\rho(x,0) - \rho_a(x))\, dx + \int \Psi_b(x)(\rho(x,1) - \rho_b(x))\, dx \\
&= \int_0^1 \int \left(L(v) - \frac{\partial \Phi}{\partial t} - \nabla\Phi \cdot v\right)\, \rho(x,t)\, dxdt + \int \Phi(x,1)\rho(x,1) - \Phi(x,0)\rho(x,0)\, dx \\
&\quad + \int \Psi_a(x)(\rho(x,0) - \rho_a(x))\, dx + \int \Psi_b(x)(\rho(x,1) - \rho_b(x))\, dx. \\
&= \int_0^1 \int \left(L(v) - \frac{\partial \Phi}{\partial t} - \nabla\Phi \cdot v\right)\, \rho(x,t)\, dxdt + \int (-\Psi_a\rho_a - \Psi_b\rho_b)\, dx \\
&\quad + \int (\Psi_a(x) - \Phi(x,0))\rho(x,0)\, dx + \int (\Psi_b(x) + \Phi(x,1))\rho(x,1)\, dx.
\end{aligned}$$

For the second equality, we apply integration by parts on $[0, 1]$. Solving the constrained optimization problem (Equation 2.20) is equivalent to investigating the following saddle point optimization problem

$$\max_{\Phi,\Psi_a,\Psi_b} \min_{\rho,v} \mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b), \tag{2.25}$$

$$\text{or} \quad \min_{\rho,v} \max_{\Phi,\Psi_a,\Psi_b} \mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b). \tag{2.26}$$

The optimality condition is given by the Karush–Kuhn–Tucker (KKT) conditions [31]

$$\frac{\partial \mathfrak{L}}{\partial \Phi} = 0, \frac{\partial \mathfrak{L}}{\partial \Psi_a} = 0, \frac{\partial \mathfrak{L}}{\partial \Psi_b} = 0, \frac{\partial \mathfrak{L}}{\partial \rho} = 0, \frac{\partial \mathfrak{L}}{\partial v} = 0. \tag{2.27}$$

The first three conditions lead to the constraints (Equation 2.21) and (Equation 2.22). The fourth condition in (Equation 2.27) yields

$$-\frac{\partial \Phi}{\partial t} - (\nabla \Phi(x,t) \cdot v(x,t) - L(v(x,t))) = 0, \tag{2.28}$$

$$\Psi_b(x) + \Phi(x, 1) = 0, \tag{2.29}$$

$$\Psi_a(x) - \Phi(x, 0) = 0. \tag{2.30}$$

The last condition in (Equation 2.27) yields $\nabla L(v(x,t)) - \nabla \Phi(x,t) = 0$, which can be rewritten as

$$v(x,t) = \nabla L^{-1}(\nabla \Phi(x,t)). \tag{2.31}$$

Now combine (Equation 2.31) and (Equation 2.28) we obtain the equation

$$\frac{\partial \Phi}{\partial t} + (\nabla \Phi(x,t) \cdot \nabla L^{-1}(\nabla \Phi(x,t)) - L(\nabla L^{-1}(\nabla \Phi(x,t)))) = 0. \tag{2.32}$$

Now we define

Definition 2.1.2. *We define the **Hamiltonian** $H(\cdot)$ as the Legendre transform of the **La-***

grangian $L(\cdot)$ *as*

$$H(p) = \max_{v\in\mathbb{R}^d}\{p\cdot v - L(v)\} = \nabla L^{-1}(p)\cdot p - L(\nabla L^{-1}(p)) \quad \forall v \ \in \mathbb{R}^d. \tag{2.33}$$

Furthermore, one can verify that when L *is strictly convex, then*

$$\nabla L^{-1}(p) = \nabla H(p), \quad \nabla H^{-1}(v) = \nabla L(v), \quad \forall\, p, v \in \mathbb{R}^d. \tag{2.34}$$

Then (Equation 2.32) becomes the following **Hamilton-Jacobi equation**

$$\frac{\partial \Phi}{\partial t} + H(\nabla\Phi(x,t)) = 0. \tag{2.35}$$

Now we combine all the KKT conditions together and obtain the following PDE system as the coupling of continuity equation and Hamilton-Jacobi equation with boundary conditions.

$$\frac{\partial \rho(x,t)}{\partial t} + \nabla\cdot(\rho(x,t)v(x,t)) = 0, \tag{2.36}$$

$$\text{here } \ v(x,t) = \nabla L^{-1}(\nabla\Phi(x,t)),$$

$$\frac{\partial \Phi(x,t)}{\partial t} + H(\nabla\Phi(x,t)) = 0. \tag{2.37}$$

$$\text{Such that: } \ \rho(\cdot,0) = \rho_a, \ \rho(\cdot,1) = \rho_b.$$

We denote $\{\rho_*(\cdot,t), v_*(\cdot,t)\}$ as the optimal solution to (Equation 2.20). Then we have $\rho_*(\cdot,t) = \rho(\cdot,t), \ v_*(\cdot,t) = \nabla L^{-1}(\nabla\Phi(\cdot,t))$, where $(\rho(\cdot,t),\Phi(\cdot,t))$ solve the PDE system (Equation 2.36)(Equation 2.37).

Example 2.1.4 (p-Wasserstein geodesic)**.** *We consider* $L(v) = \frac{|v|^p}{p}$ *with* $p > 1$*. We call the optimal* $\{\rho(\cdot,t)\}$ *to (Dym-OT)*(Equation 2.20) *as the* p-***Wasserstein geodesic*** *joining* ρ_a, ρ_b*.*

Specifically, when $p = 2$, the geodesic equation for 2-Wasserstein geodesic is

$$\frac{\partial \rho(x,t)}{\partial t} + \nabla \cdot (\rho(x,t)\nabla\Phi(x,t)) = 0, \tag{2.38}$$

$$\frac{\partial \Phi(x,t)}{\partial t} + \frac{|\nabla\Phi(x,t)|^2}{2} = 0. \tag{2.39}$$

Such that: $\rho(\cdot,0) = \rho_a,\ \rho(\cdot,1) = \rho_b$.

Remark 1 (Duality of Dynamical OT problem)**.** *One can verify that the max-min saddle problem* (Equation 2.25) *is further equivalent to the following maximization problem*

$$K_{Dym}(\rho_a,\rho_b) \triangleq \max_{\Phi} \left\{ \int_{\mathbb{R}^d} \Phi(x,1)\ \rho_b(x)\ dx - \int_{\mathbb{R}^d} \Phi(x,0)\rho_a(x)\ dx \right\}, \quad \textit{(dual-Dym-OT)} \tag{2.40}$$

$$\textit{subject to:}\ \ \frac{\partial \Phi(x,t)}{\partial t} + H(\nabla\Phi(x,t)) = 0 \quad 0 \le t \le 1. \tag{2.41}$$

(Equation 2.40) *can be treated as the dual problem of dynamical OT problem* (Equation 2.20). *It is worth pointing out the equivalence between* (Equation 2.40) *and the Kantorovich dual problem* (Equation 2.11). *Actually, if we set the cost function $c(x,y) = L(x-y)$. Then by Hopf-Lax formula [34] of Hamilton Jacobi equation* (Equation 2.41)*, one can verify $\Phi(\cdot,1) = \inf_x\{\Phi(x,0) + c(x,\cdot)\} = \Phi^{c,+}(\cdot,0)$. This is exactly the constraint condition of* (Equation 2.11).

It is also worth studying the optimal solution to the equivalent problem (Equation 2.23) from particle point of view. Due to the equivalence between (Equation 2.23), (Equation 2.20), the optimal vector field $v_*(\cdot,t)$ of problem (Equation 2.23) should also be $\nabla L^{-1}(\nabla\Phi(\cdot,t))$, where Φ is solved from (Equation 2.37). Now denote $\{\boldsymbol{X}_t^*\}$ as the trajectory obeying the optimal vector field $v_*(\cdot,t)$, i.e., $\boldsymbol{X}_t^*$ solves

$$\dot{\boldsymbol{X}}_t^* = v_*(\boldsymbol{X}_t^*,t) = \nabla L^{-1}(\nabla\Phi(\boldsymbol{X}_t^*,t)). \tag{2.42}$$

we analyze the behavior of the movement of particle $\boldsymbol{X}_t^*$ by considering its second order dynamic: by denoting $\boldsymbol{p}_t = \nabla\Phi(\boldsymbol{X}_t^*, t)$, then under certain smooth assumption on Φ, one can verify

$$\begin{aligned}\dot{\boldsymbol{p}}_t &= \nabla^2\Phi(\boldsymbol{X}_t^*, t)\dot{\boldsymbol{X}}_t^* + \frac{\partial}{\partial t}\nabla\Phi(\boldsymbol{X}_t^*, t) \\ &= \nabla^2\Phi(\boldsymbol{X}_t^*, t)\nabla L^{-1}(\nabla\Phi(\boldsymbol{X}_t^*, t)) + \frac{\partial}{\partial t}\nabla\Phi(\boldsymbol{X}_t^*, t) \\ &= \nabla^2\Phi(\boldsymbol{X}_t^*, t)\nabla H(\nabla\Phi(\boldsymbol{X}_t^*, t)) + \frac{\partial}{\partial t}\nabla\Phi(\boldsymbol{X}_t^*, t). \end{aligned} \tag{2.43}$$

Here for the third equality, we use the fact (Equation 2.34) stated in Definition 2.1.2. On the other hand, taking gradient of x on both sides of Hamilton Jacobi equation (Equation 2.37) yields

$$\frac{\partial}{\partial t}\nabla\Phi(x, t) + \nabla^2\Phi(x, t)\nabla H(\nabla\Phi(x, t)) = 0. \tag{2.44}$$

Thus, combine (Equation 2.43) and (Equation 2.44) one derives $\dot{\boldsymbol{p}}_t = 0$. This indicates $\ddot{\boldsymbol{X}}_t^* = 0$ and $\boldsymbol{X}_t^*$ possess constant velocity. Such property is also reflected in Figure 2.3. In general, we have the following theorem

Theorem 2.1.6 (Trajectory of (Dym-OT))**.** *Suppose $\{\boldsymbol{X}_t^*\}_{t=0}^1$ is the trajectory obeying the optimal vector field of* (Equation 2.23) *with strictly convex Lagrangian L, then $\ddot{\boldsymbol{X}}_t^* = 0$, and thus*

$$\boldsymbol{X}_t^* = \boldsymbol{X}_0^* + tv_*(\boldsymbol{X}_0^*, 0) = \boldsymbol{X}_0^* + t\nabla L^{-1}(\nabla\Phi(\boldsymbol{X}_0^*, 0)). \tag{2.45}$$

This result was discussed by [6] for quadratic Lagrangian. More general results are presented in Theorem 5.5 of [29]. The following corollary is the natural result of Theorem 2.1.6.

Corollary 2.1.6.1. *Suppose the initial distribution of $\boldsymbol{X}_0^*$ equals ρ_a and $\{\boldsymbol{X}_t^*\}$ solves* (Equation 2.42)*. Then the optimal $\rho_*(\cdot, t)$ to* (Equation 2.20) *equals to the probability den-*

sity of Law$(\boldsymbol{X}_t^*)$[1]*. Using* (Equation 2.45)*, we deduce that*

$$\rho_*(\cdot,t) = (Id + tv_*(\cdot,0))_\sharp \rho_a = (Id + t\nabla L^{-1}(\nabla\Phi(\cdot,0)))_\sharp \rho_a. \tag{2.46}$$

(Equation 2.46) justifies that the optimal $\rho_*(\cdot,t)$ is obtained by pushforwarding the initial distribution ρ_a along the geodesic (straight lines) with the initial velocity $v_*(,0) = \nabla L^{-1}(\nabla\Phi(\cdot,0))$.

Remark 2. *One can also interpret Theorem 2.1.6 from another perspective: each particle* $\boldsymbol{X}_t^*$ *should choose its own optimal trajectory by minimizing its cost along the path* $\int_0^1 L(\dot{\boldsymbol{X}}_t^*)\,dt$*, thus* $\boldsymbol{X}_t^*$ *solves the corresponding Euler-Lagrange equation* $-\frac{d}{dt}(\nabla_v L(\dot{\boldsymbol{X}}_t^*)) + \nabla_x L(\dot{\boldsymbol{X}}_t^*)) = 0$*. In our discussion,* L *is independent of* x *and is convex w.r.t.* v*, thus the E-L equation directly leads to* $\ddot{\boldsymbol{X}}_t^* = 0$*.*

2.1.6 Equivalence among different versions of OT problem

We conclude this section by briefly stating the equivalence among different versions of optimal transport problems. We should first assume L as defined in Definition 2.1.1, and we further assume the cost function c is compatible with L in the sense of

$$c(x,y) \triangleq \min_{\{\boldsymbol{X}_t\},\boldsymbol{X}_0=x,\boldsymbol{X}_1=y} \left\{ \int_0^1 L(\dot{\boldsymbol{X}}_t)\,dt \right\} = L(x-y). \tag{2.47}$$

Let us consider the optimal transport between μ, ν. We assume both μ, ν possess densities ρ_a, ρ_b. We list four different versions of OT problems in the following table.

	(Dym-OT) (Equation 2.20)	(MP) (Equation 2.2)	(KP) (Equation 2.6)	(dual-KP) (Equation 2.9)
Optimal solution	$\{\rho_*(\cdot,t), v_*(\cdot,t)\}$ $= \{\rho(\cdot,t), \nabla L^{-1}(\nabla\Phi(\cdot,t))\}$	T_*	γ_*	(ψ_*, ϕ_*)
Optimal value	$C_{\text{Dym}}(\rho_a,\rho_b)$	$C_{\text{Monge}}(\mu,\nu)$	$C(\mu,\nu)$	$K(\mu,\nu)$

[1]We denote Law$(\boldsymbol{X})$ as the probability distribution of the random variable $\boldsymbol{X}$.

Under the assumption that these optimal solutions uniquely exist, relationships among these four versions of OT problems are listed as following

- The optimal values of (Dym-OT), (MP), (KP) and (dual-KP) are equal;
- (MP) & (KP): As stated in Theorem 2.1.2, $\gamma_* = (\mathrm{Id}, T_*)_\sharp \mu$;
- (MP) & (dual-KP): As stated in (Equation 2.14) (Equation 2.15) (Equation 2.16),

$$\nabla\phi_*(T_*(x)) - \partial_y c(x, T_*(x)) = 0, \qquad -\nabla\psi_*(x) - \partial_x c(x, T_*(x)) = 0, \quad \text{on Spt}(\mu).$$
$$T_*(x) = \partial_x c(x, \cdot)^{-1}(-\nabla\psi_*(x)), \quad \text{on Spt}(\mu).$$

- (Dym-OT) & (dual-KP): As stated in Remark 1, we have $(\psi_*, \phi_*) = (\Phi(\cdot, 0), \Phi(\cdot, 1))$ up to a constant number;
- (Dym-OT) & (MP): Combine the discussion in previous two bullets, we obtain

$$T_*(x) = \partial_x c(x, \cdot)^{-1}(-\nabla\Phi(x, 0)) = x + \nabla L^{-1}(\nabla\Phi(x, 0)) \quad \text{on Spt}(\mu). \tag{2.48}$$

2.2 Wasserstein manifold

In this section, we briefly present some of the basic concepts and results regarding Wasserstein manifold. We only provide in section 2.2 a heuristic but informal discussion on Wasserstein manifold and Wasserstein gradient flow. More rigorous treatments on these topics can be found in [7],[35].

2.2.1 Wasserstein metric

Recall the 2-Wasserstein distance W_2 introduced in Example 2.1.1, this distance function will imply Riemannian geometry structure on certain probability space[3][4]. To be more specific, we denote the probability space supported on $\mathbb{R}^d$ with positive densities and finite

second order moments as

$$\mathcal{P}_2 = \left\{ \rho \;\middle|\; \int \rho(x)dx = 1,\ \rho(x) > 0,\ \int |x|^2 \rho(x)\, dx < \infty \right\}. \tag{2.49}$$

Here and for the following discussion in this section, if not specified, we always treat $\int$ as $\int_{\mathbb{R}^d}$ for simplicity.

If we treat $\mathcal{P}_2$ as an infinite dimensional manifold, the Wasserstein distance W_2 can induce a metric g^W defined on the tangent bundle $\mathcal{T}\mathcal{P}_2$, with which $\mathcal{P}_2$ becomes a Riemannian manifold. For simplicity, here we directly give the definition of g^W. One can identify the tangent space at ρ as:

$$\mathcal{T}_\rho \mathcal{P}_2 = \left\{ f \;\middle|\; \int f(x)dx = 0 \right\}.$$

We now present a less-rigorous, but more heuristic derivation that can motivate the definition of metric g^W. Suppose we fix certain $\rho \in \mathcal{P}_2$, consider an arbitrary tangent vector $f \in \mathcal{T}_\rho \mathcal{P}$, and a very short time stepsize $h > 0$. Suppose at time $t = 0$, we start from $\rho_0 = \rho$, and we let ρ move along the tangent vector f, at time $t = h$, we should have $\rho_h = \rho + hf + o(h)$. The $2-$Wasserstein distance between ρ_0 and ρ_h is $W_2(\rho_0, \rho_h)$, the compatibleness between W_2 distance and g^W leads to

$$W_2(\rho_0, \rho_h)^2 = h^2 g^W(f, f) + o(h^2). \tag{2.50}$$

Suppose the Monge map between ρ_0 and ρ_h is T_*. Then

$$W_2(\rho_0, \rho_h)^2 = \int_{\mathbb{R}^d} |T_*(x) - x|^2 \rho(x)dx. \tag{2.51}$$

On the other hand, according to the discussion in subsection 2.1.4, the Monge map $T_*(x)$ takes the form $\partial_x c(x, \cdot)^{-1}(-\nabla \psi_*(x))$, where ψ_* solves the Monge-Ampère equation in

(Equation 2.18). In the L^2 case, $c(x, y) = |x - y|^2$, then we can write

$$T_*(x) = x + \frac{1}{2}\nabla\psi_*(x).$$

Furthermore, since ρ_h differs from ρ by an $O(h)$ term, it is then reasonable to recast the Monge map as

$$T_*(x) = x + \frac{h}{2}\nabla\widetilde{\psi_*}(x), \tag{2.52}$$

here we are rescaling ψ_* to $h\widetilde{\psi_*}$. Now by (Equation 2.18), one can verify that $\widetilde{\psi_*}$ solves

$$\rho(x) = \det(I_d + \frac{h}{2}\nabla^2\widetilde{\psi_*}(x))) \cdot \rho_h(x + \frac{h}{2}\nabla\widetilde{\psi_*}(x)).$$

Now replace ρ_h by $\rho + hf + o(h)$, the previous equation becomes

$$\rho(x) = \det(I_d + \frac{h}{2}\nabla^2\widetilde{\psi_*}(x))) \cdot (\rho(x + \frac{h}{2}\nabla\widetilde{\psi_*}(x)) + hf(x + \frac{h}{2}\nabla\widetilde{\psi_*}(x))) + o(h). \tag{2.53}$$

Now we expand $\det(I_d + \frac{h}{2}\nabla^2\widetilde{\psi_*}(x))) = 1 + \frac{h}{2}\Delta\widetilde{\psi_*}(x) + o(h)$, and $\rho(x + \frac{h}{2}\nabla\widetilde{\psi_*}(x)) = \rho(x) + \frac{h}{2}\nabla\rho(x) \cdot \nabla\widetilde{\psi_*}(x) + o(h)$, $f(x + \frac{h}{2}\nabla\widetilde{\psi_*}(x)) = f(x) + O(h)$. Plug these into (Equation 2.53), we obtain

$$\rho(x) = (1+\frac{h}{2}\Delta\widetilde{\psi_*}(x)+o(h))(\rho(x)+\frac{h}{2}\nabla\rho(x)\cdot\nabla\widetilde{\psi_*}(x)+o(h)+h(f(x)+O(h)))+o(h). \tag{2.54}$$

This equation can be simplified as

$$0 = h\left(\frac{1}{2}\Delta\widetilde{\psi_*}(x)\rho(x) + \frac{1}{2}\nabla\rho(x) \cdot \nabla\widetilde{\psi_*}(x) + f(x)\right) + o(h). \tag{2.55}$$

Now we divide on both sides of (Equation 2.55) by h, and send $h \to 0$, we obtain the elliptical PDE

$$-\nabla \cdot \left(\rho(x)\nabla\frac{\widetilde{\psi_*}(x)}{2}\right) = f(x). \tag{2.56}$$

Now recall (Equation 2.50), (Equation 2.51), and (Equation 2.52), we obtain

$$h^2 \left(\int |\nabla \frac{\widetilde{\psi_*}(x)}{2}|^2 \rho(x) dx \right) = h^2 g^W(f, f) + o(h^2). \tag{2.57}$$

Now we divide (Equation 2.57) on both sides by h^2 and send $h \to 0$, we obtain

$$g^W(f, f) = \int |\nabla \frac{\widetilde{\psi_*}(x)}{2}|^2 \rho(x) dx, \quad \widetilde{\psi_*} \text{ solves the equation (Equation 2.56).}$$

The above calculations will motivate the following definition for Wasserstein metric g^W:

Definition 2.2.1 (Wasserstein metric)**.** *For a specific $\rho \in \mathcal{P}_2$ and $f_i \in \mathcal{T}_\rho \mathcal{P}_2$, $i = 1, 2$, we define the **Wasserstein metric** tensor g^W as [3, 4]*

$$g^W(\rho)(f_1, f_2) = \int \nabla \psi_1(x) \cdot \nabla \psi_2(x) \rho(x) \, dx, \tag{2.58}$$

where ψ_1, ψ_2 satisfies

$$-\nabla \cdot (\rho \nabla \psi_i) = f_i \quad i = 1, 2, \tag{2.59}$$

with boundary conditions

$$\lim_{x \to \infty} \rho(x) \nabla \psi_i(x) = 0 \quad i = 1, 2.$$

Use the above definition, we can also write

$$g^W(\rho)(f_1, f_2) = \int \psi_1(-\nabla \cdot (\rho \nabla \psi_2)) \, dx = \int (-\nabla \cdot (\rho \nabla))^\dagger (f_1) \cdot f_2 \, dx.$$

Here we denote $(-\nabla \cdot (\rho \nabla))^\dagger$ as the pseudo inverse of the negative weighted Laplacian operator $-\nabla \cdot (\rho \nabla)$. Thus, we can identify $g^W(\rho)$ as $(-\nabla \cdot (\rho \nabla))^{-1}$. Due to the positive definiteness of $-\nabla \cdot (\rho \nabla)$, one can also verify that $g^W(\rho)$ is a positive definite bilinear form defined on tangent bundle $\mathcal{T}\mathcal{P}_2 = \{(\rho, f) \colon \rho \in \mathcal{P}_2, \ f \in \mathcal{T}_\rho \mathcal{P}_2\}$. Hence we can treat $\mathcal{P}_2$

as a Riemannian manifold, which we call **Wasserstein manifold**, denoted as $(\mathcal{P}_2, g^W)$ [4]. In order to keep our notations concise, in future discussion, we denote $g^W(\rho)$ as g^W if no confusion is caused.

The Wasserstein metric g^W is compatible with W_2 distance in the sense that for any $\mu, \nu \in \mathcal{P}_2$ with densities ρ_a, ρ_b,

$$W_2^2(\mu, \nu) = \min_{\substack{\{\rho_t\}_{0\le t\le 1} \\ \rho_0 = \rho_a, \rho_1 = \rho_b}} \left\{ \int_0^1 g^W(\dot\rho_t, \dot\rho_t)\, dt \right\}.$$

Here we denote $\rho_t = \rho(\cdot, t)$ and $\dot\rho_t = \partial_t \rho(\cdot, t)$.

2.2.2 Wasserstein gradient flow

We denote the **Wasserstein gradient** grad_W as the manifold gradient on $(\mathcal{P}_2, g^W)$. In Riemannian geometry, the manifold gradient must be compatible with the metric, implying that for any smooth functional $\mathcal{F}$ defined on $\mathcal{P}_2$ and any $\rho \in \mathcal{P}_2$, consider an arbitrary differentiable curve $\{\rho_t\}_{t\in(-\delta,\delta)}$ with $\rho_0 = \rho$, we have

$$\frac{d}{dt}\mathcal{F}(\rho_t)\Big|_{t=0} = g^W(\rho)(\text{grad}_W \mathcal{F}(\rho),\ \dot\rho_0).$$

Since we can write

$$\frac{d}{dt}\mathcal{F}(\rho_t)\Big|_{t=0} = \int \frac{\delta \mathcal{F}(\rho)}{\delta \rho(x)}(x) \cdot \dot\rho_0(x)\, dx = \left\langle \frac{\delta \mathcal{F}(\rho)}{\delta \rho}, \dot\rho_0 \right\rangle_{L^2},$$

here $\frac{\delta \mathcal{F}(\rho)}{\delta \rho(x)}(x)$ is the L^2 variation of $\mathcal{F}$ at point $x \in \mathbb{R}^d$, we then have

$$\left\langle \frac{\delta \mathcal{F}(\rho)}{\delta \rho}, \dot\rho_0 \right\rangle_{L^2} = g^W(\rho)(\text{grad}_W \mathcal{F}(\rho),\ \dot\rho_0) \quad \forall\, \dot\rho_0 \in \mathcal{T}_\rho \mathcal{P}_2. \tag{2.60}$$

Recall $g^W(\rho) = (-\nabla \cdot (\rho \nabla))^\dagger$, (Equation 2.60) then leads to the following useful formula for computing the Wasserstein gradient of $\mathcal{F}$

$$\text{grad}_W \mathcal{F}(\rho) = g^W(\rho)^{-1} \left(\frac{\delta \mathcal{F}}{\delta \rho} \right)(x) = -\nabla \cdot \left(\rho(x) \nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho(x)}(x) \right). \tag{2.61}$$

Once we know how to compute Wasserstein gradient, we can also formulate the **Wasserstein gradient flow** as the following evolutional PDE of $\rho(x,t)$.

$$\frac{\partial \rho(x,t)}{\partial t} = -\text{grad}_W \mathcal{F}(\rho_t) = \nabla \cdot \left(\rho(x,t) \nabla \frac{\delta}{\delta \rho_t} \mathcal{F}(\rho_t) \right). \tag{2.62}$$

Wasserstein gradient flow has a close relation with many dissipative evolutional PDEs. Here are some examples.

Example 2.2.1 (Fokker-Planck equation)**.** *If $\mathcal{F}$ is taken as the **relative entropy** functional*

$$\mathcal{H}(\rho) = \int V(x)\rho(x) + D\rho(x) \log \rho(x) \, dx + \textit{Constant}, \tag{2.63}$$

we have $\nabla \frac{\delta \mathcal{H}(\rho)}{\delta \rho} = \nabla V + D \nabla \log \rho$. Using (Equation 2.61)*, and noticing $\nabla \log \rho = \frac{\nabla \rho}{\rho}$, then $\nabla \cdot (\rho \nabla \log \rho) = \nabla \cdot (\nabla \rho) = \Delta \rho$, the Wasserstein gradient flow of $\mathcal{H}$ is*

$$\frac{\partial \rho}{\partial t} = -\textit{grad}_W \mathcal{H}(\rho) = \nabla \cdot (\rho \nabla V) + D \nabla \cdot (\rho \nabla \log \rho)).$$

This is the Fokker-Planck equation used to describe the density evolution of certain stochastic dynamics[36][37].

Example 2.2.2 (Porous medium equation)**.** *If $\mathcal{F}$ is taken as the power functional*

$$\mathcal{F}(\rho) = \frac{1}{m-1} \int \rho^m \, dx, \quad m \neq 1 \tag{2.64}$$

we have $\nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho} = \nabla (\frac{m}{m-1} \rho^{m-1}) = m \rho^{m-2} \nabla \rho$. Then $\textit{grad}_W \mathcal{F}(\rho) = -\nabla \cdot (\rho \nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho}) =$

$-\nabla \cdot (m\rho^{m-1}\nabla\rho) = -\Delta\rho^m$. *The Wasserstein gradient flow of* $\mathcal{F}$ *is*

$$\frac{\partial \rho}{\partial t} = -grad_W \mathcal{F}(\rho) = \Delta\rho^m.$$

This is the Porous-Medium equation, which appears in a number of physical applications, such as to describe processes involving fluid flow, heat transfer, or diffusion [4][38].

CHAPTER 3

COMPUTATIONAL PROBLEMS RELATED TO OPTIMAL TRANSPORT

In this chapter, we mainly focus on solving optimal transport problems from three different perspectives. This chapter is organized as follows: we first provide in section 3.1 a comprehensive literature review on existing algorithms for optimal transport problems; Then in section 3.2, we derive a scalable algortihm for computing Monge maps with general cost functions [20, 39]; In section 3.3, we propose a particle evolving technique for approximating the optimal coupling of the Kantorovich problem [21, 22]; In section 3.4, we design a novel formulation and learning strategy for computing the Wasserstein geodesic between two probability distributions in high dimensional space [40].

3.1 Literature review

3.1.1 Algorithms for Monge problem

In [13], the authors propose a method for computing the optimal coupling as well as the Monge map of a general OT problem by introducing entropic regularization term to the primal problem (Equation 2.6) and then optimize the corresponding dual problem;

Algorithms proposed in [18, 19, 41] make use of the convex property of Kantorovich dual pairs, in conjuction with the artitecture of input convex neural network (ICNN) [42] to formulate certain optimization algorithms, which are able to approximate the optimal map for L^2 Monge problem (i.e. $c(x, y) = |x - y|^2$) under high dimensional settings.

There are also researches focusing on solving the Monge-Ampère equation [43, 44, 45, 46]. As mentioned in subsection 2.1.4, evaluating the Monge map is equivalent to solving the Monge-Ampère equation (Equation 2.18). However, due to curse of dimenaionality, the discretization methods proposed in the aforementioned references can not be applied to

high dimensional problems.

3.1.2 Algorithms for Kantorovich problem

The straightforward way to solve the Kantorovich problem (Equation 2.6) is by using linear programming solvers such the Hungarian algorithm and the auction algorithm. See [47, 48, 10] and the references therein.

A popular method known as *Sinkhorn algorithm* [49] for solving Kantorovich problem (Equation 2.6) is proposed by introducing the entropic regularization to the Kantorovich problem, and then computes for the optimal γ_* via iterations. This algorithm is capable of computing the OT distance between two sets of data points in high dimensional space [50, 51, 52, 53].

Another popular treatment of high dimensional W_1 OT problem was introduced in Wasserstein generative adversarial networks (WGAN) [8] by considering the dual formulation (Equation 2.9). In order to enforce the Lipshitz-1 constraint in the dual Kantorovich problem, people introduce regularization term such as gradient penalty [54] to improve the performance of the algorithm. Since then, various regularization-based OT problems have been formulated, such as OT algorithms with spectral normalization [55], entropic regularization [13], Laplacian regularization [56], Group-Lasso regularization [57], Tsallis regularization [58] and L^2 regularization [59]. In addition to the ordinary OT problem, there exists research [60] proposing self-defined dual Kantorovich problem in order to invent new type of discrepancy functions of probability distributions that may improve the performance of corresponding generative models.

In the research [15], the authors relax the marginal constraints by incorporating the Wasserstein marginal discrepancies into the Kantorovich problem. Such treatment leads to an optimization problem without constraints, which can then be efficiently resolved by classical optimization techniques in deep learning. Although our work [21, 22] share similar ideas with [15], both the purposes and computational methods are distinct from this

work.

3.1.3 Algorithms for dynamical version of optimal transport problem

The first reference that motivates the dynamical version of $2-$Wasserstein problem is [61]. Later, in [62], the authors consider and discretize the dynamical $1-$Wasserstein problem and come up with an efficient primal-dual solver. In addition to the dynamical problems only involving pure transportation, unbalanced transport as well as unnormalized transport problems have also drawn the interests of researcher in recent days [63, 64]. However, it is worth mentioning that all the references mentioned so far in this subsection use classical finite difference or finite element methods, which cannot be scaled up into high dimensions.

One recent work aim at computing high dimensional mean-field games is proposed in [17]. The algorithm is capable of computing dynamical OT problem by relaxing the terminal time constraint as a cost functional in the mean-field control functional.

3.2 Scalable computation of Monge maps with general costs

3.2.1 Introduction

Optimal transport (OT) based applications have achieved great success in machine learning research [8, 18, 65, 52, 66, 67, 41, 68] in recent years. As mentioned in Chapter 1, Wasserstein distance $C(\mu, \nu)$ (subsection 2.1.2) is a crucial measurement of the discrepancy between probability distributions due to its robustness.

By discretizing the space, one can treat Kantorovich problem (Equation 2.6) as a linear programming problem. Under high dimensional setting, we will face a large scale linear programming problem which could be very challenging due to the curse of dimensionality. To tackle with such difficulty, by introducing the entropic regularization, people develop the so-called Sinkhorn algorithm [49] that operates iteratively to computes for the approximation of Wasserstein distance between two sets of data points in high dimensional space[50, 51, 52, 53]. However, Sinkhorn algorithm also experiences drawbacks such as slow convergence speed when entropic regularization coefficient $\frac{1}{\lambda}$ is small, and not being suitable to handle continuous probability distributions. On the other hand, in the field of data science and machine learning research, people prefer the exact optimal map T_* to the optimal coupling γ_* since the map T_*, as a generative model, offers great convenience on sampling from the target distribution ν. In this research, instead of considering the Kantorovich problem, we mainly focus on solving the original Monge problem (Equation 2.2) for the optimal transport map T_*. To be more specific, we treat (Equation 2.2) as a constrained optimization problem. By introducing the Lagrange multiplier, we reformulate the Monge problem as a max-min saddle point problem. Then by setting the transport map as well as the Lagrange multiplier function as neural networks, we obtain a scalable algorithm for approximating the Monge map T_* by alternatively optimizing the parameters of two networks. The main contributions of our proposed method can be summarized as:

1. Our method is capable of computing the Monge map associated with *general* cost func-

tion between two distributions with given samples points in high dimensional space;

2. Our method possesses flexibility to deal with OT problems between μ, ν that are not absolute continuous to the Lebesgue measures, as well as OT problems between unequal dimensional spaces;

3. We provide theoretical guarantee on the correctness of our method; we also establish error analysis result based on duality gaps that come from our algorithm;

4. We verify the performance of our algorithm via series of examples varying from low-dimensional synthetic data to high-dimensional realistic data.

We refer the readers to subsection 3.1.1 for related references. It is worth mentioning that in addition to our work [20], several related research projects on computing the Monge map for general OT problems [69, 70, 71] were also proposed by another group of researchers recently.

We refer the reader to subsection 2.1.1, subsection 2.1.2, subsection 2.1.3 and subsection 2.1.6 for related mathematical backgrounds.

3.2.2 Proposed method

Let us recall the Monge problem (Equation 2.2).

$$C_{\text{Monge}}(\mu, \nu) \triangleq \min_{T} \left\{ \int c(x, T(x)) d\mu(x) \right\} \quad \text{(MP)} \qquad \text{(Equation 2.2)}$$

over the set of all measurable maps $T : X \to Y$ such that $T_\sharp \mu = \nu$.

In our following discussion in this section, we treat $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$. In order to formulate a tractable algorithm for (Equation 2.2) with general cost c, we first notice that (Equation 2.2) is a constrained optimization problem. Thus, it is natural to introduce the Lagrange multiplier f for the constraint $T_\sharp \mu = \nu$ and then reformulate (Equation 2.2) as a

max-min saddle point problem

$$\sup_{f} \inf_{T} \mathcal{L}(T, f) \tag{3.1}$$

with $\mathcal{L}$ defined as

$$\begin{aligned} \mathcal{L}(T, f) &= \int_{\mathbb{R}^n} c(x, T(x)) d\mu(x) + \int_{\mathbb{R}^m} f(y)(\nu(y) - T_\sharp \mu(y))\, dy \\ &= \int_{\mathbb{R}^n} \left[c(x, T(x)) - f(T(x))\right] d\mu(x) + \int_{\mathbb{R}^m} f(y)\nu(y)\, dy \end{aligned} \tag{3.2}$$

We can verify that the max-min scheme (Equation 3.1) is equivalent to the Kantorovich dual problem (Equation 2.9). To this end, one only need to verify:

$$\begin{aligned} \inf_{T} \mathcal{L}(T, f) &= -\int_{\mathbb{R}^n} \sup_{\xi} \{f(\xi) - c(x, \xi)\} d\mu(x) + \int_{\mathbb{R}^m} f(y)\nu(y) dy \\ &= \int_{\mathbb{R}^m} f(y)\nu(y) dy - \int_{\mathbb{R}^n} f^{c,-}(x) d\mu(x). \end{aligned} \tag{3.3}$$

Here, recall $f^{c,-}$ is defined in (Equation 2.10). The following theorem guarantees that the max-min scheme (Equation 3.1) will find the optimal Monge map.

Theorem 3.2.1 (Consistency)**.** *We assume that the optimal solution to the Monge problem* (Equation 2.2) *exists, and the optimal solution to the Kantorovich problem* (Equation 2.6) *is unique up to a constant number. Suppose the saddle point solution to* (Equation 3.1) *is* (T_*, f_*)*, then* T_* *is the optimal solution to* (Equation 2.2) *and* $f_* = \phi_* + C$*, where* ϕ_* *is the optimal solution to* (Equation 2.9) *and* C *is a constant number.*

Proof of Theorem 3.2.1. According to (Equation 3.3), and the uniqueness assumption on (Equation 2.6), we are able to tell that f_* equals $\phi_* + C$, where C is a constant number. Furthermore, at the saddle point (T_*, f_*), we have

$$T_{*\sharp}\mu = \nu, \quad T_*(x) \in \text{argmax}_{\xi \in \mathbb{R}^m} \{f_*(\xi) - c(x, \xi)\}.$$

The second equation leads to

$$f_*^{c,-}(x) = f_*(T_*(x)) - c(x, T_*(x)).$$

Then we have

$$\begin{aligned}\int_{\mathbb{R}^n} c(x, T_*(x))d\mu(x) &= \int_{\mathbb{R}^n} f_*(T_*(x))d\mu(x) - \int_{\mathbb{R}^n} f_*^{c,-}(x)d\mu(x) \\ &= \int_{\mathbb{R}^m} f_*(y)\nu(y)\, dy - \int_{\mathbb{R}^n} f_*^{c,-}(x)d\mu(x) \\ &\leq \int_{\mathbb{R}^n \times \mathbb{R}^m} [f_*(y) - f_*^{c,-}(x)]d\gamma(x,y) \leq \int_{\mathbb{R}^n \times \mathbb{R}^m} c(x,y)d\gamma(x,y)\end{aligned}$$

for any $\gamma \in \Pi(\mu, \nu)$. Here the second equality is due to $T_{*\sharp}\mu = \nu$, the last inequality is due to the definition of $f_*^{c,-}(x) = \sup_y\{f_*(y) - c(x,y)\}$.

Now we set γ to be the optimal γ_* of corresponding Kantorovich problem in the previous inequality. Since we assume that the Monge map exists, by Theorem 2.1.2,

$$\int_{\mathbb{R}^n \times \mathbb{R}^m} c(x,y)d\gamma(x,y) = C(\mu,\nu) = C_{\text{Monge}}(\mu,\nu),$$

thus $\int_{\mathbb{R}^n} c(x, T_*(x))d\mu(x) = C_{\text{Monge}}(\mu,\nu)$. As a result, T_* is the optimal solution to (Equation 2.2). □

In exact implementation, we will replace both the map T and the dual variable f by the neural networks T_θ, f_η, with θ, η being the parameters of the networks. We aim at solving the following saddle point problem. The algorithm is summarized in Algorithm 1.

$$\max_\eta \min_\theta \mathcal{L}(T_\theta, f_\eta) := \frac{1}{N}\sum_{k=1}^{N} c(X_k, T_\theta(X_k)) - f_\eta(T_\theta(X_k)) + f_\eta(Y_k) \tag{3.4}$$

where N is the batch size and $\{X_k\}, \{Y_k\}$ are samples generated from μ and ν separately.

Remark 3 (Relation with WGAN)**.** *Although the proposed saddle scheme* (Equation 3.4)

Algorithm 1 Computing Wasserstein distance and optimal map from μ to ν

Input: Marginal distributions μ and ν, Batch size N, Cost function $c(x, T(x))$.
Initialize T_θ, f_η.
for K steps **do**
    Sample $\{X_k\}_{k=1}^N$ from μ. Sample $\{Y_k\}_{k=1}^N$ from ν.
    for K_1 steps **do**
        Update (via gradient descent) θ to decrease (Equation 3.4)
    end for
    for K_2 steps **do**
        Update (via gradient ascent) η to increase (Equation 3.4)
    end for
end for

shares similarity with the Wasserstein Generative Adversarial Networks (WGAN) [8], both the designing purpose and mathematical logic behind these two methods are distinct. Detailed comparisons are provided in appendix subsection B.1.1.

3.2.3 Error Analysis via Duality Gaps

In this section, we assume that $m = n = d$, i.e. we consider Monge problem between Euclidean spaces with the same dimension d. Suppose we solve (Equation 3.1) to a certain stage and obtain the pair (T, f), inspired by [72] and [18], we want to estimate a weighted L^2 error between our computed map T and the optimal Monge map T_*.

Before we present our result, we introduce definition for c-concave functions. We mainly adopt the definition from Chapter 5 of [7].

Definition 3.2.1 (c-concavity)**.** *We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is c-concave if there exists a function φ such that $f = \varphi^{c,+}$. This definition is also equivalent to $(f^{c,-})^{c,+} = f$.*

In order to establish our result, we require μ, ν and $c(\cdot, \cdot)$ to satisfy the conditions mentioned in Theorem (Theorem 2.1.1). In addition, we assume that $c \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$ and

satisfies

$$\partial_{xy}c(x,y), \text{ as an } d\times d \text{ matrix, is invertible and self-adjoint.} \tag{3.5}$$

$$\partial_{yy}c(x,y) \text{ is independent of } x; \tag{3.6}$$

Theorem 3.2.2 (Posterior Error Analysis via Duality Gaps)**.** *Assume $f \in C^2(\mathbb{R}^d)$ is a c-concave function and assume that there exists $\varphi \in C^2(\mathbb{R}^d)$ such that $f(y) = \inf_x\{\varphi(x) + c(x,y)\}$. Suppose $\varphi(x) + c(x,y)$ has a unique minimizer $\hat{x}_y$ for arbitrary $y \in \mathbb{R}^d$. We further assume there exists function $\lambda(\cdot) > 0$ such that the Hessian of $\varphi(\cdot) + c(\cdot, y)$ at minimizer $\hat{x}_y$ is positive definite and bounded from above by $\lambda(\cdot)$:*

$$\lambda(y)I_n \succeq \nabla^2_{xx}(\varphi(x) + c(x,y))|_{x=\hat{x}_y} \succ O_n, \tag{3.7}$$

where I_n, O_n denotes $n \times n$ identity matrix and zero matrix.

We denote $\sigma(x,y)$ as the minimum singular value of the matrix $\partial_{xy}c(x,y)$.

Now denote the duality gaps as

$$\mathcal{E}_1(T,f) = \mathcal{L}(T,f) - \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},f), \quad \mathcal{E}_2(f) = \sup_{\widetilde{f}} \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},\widetilde{f}) - \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},f).$$

Recall T_* as the Monge map of (Equation 2.2)*. Then there exists a strict positive weight function $\beta(x) \geq \min_y \left\{\frac{\sigma(x,y)}{2\lambda(y)}\right\}$ (β depends on c, T_*, f, φ) such that the weighted L^2 error between computed map T and optimal map T_* is upper bounded by*

$$\|T - T_*\|_{L^2(\beta\mu)} \leq \sqrt{2(\mathcal{E}_1(T,f) + \mathcal{E}_2(f))}.$$

We prove this theorem in the appendix subsection B.1.2.

Remark 4. *We can verify that $c(x,y) = \frac{1}{2}|x-y|^2$ or $c(x,y) = -x\cdot y$ satisfy the conditions mentioned above. Then Theorem 3.2.2 recovers similar results proved in [72] and [18].*

Remark 5. *Suppose* $c(\cdot,\cdot)$ *satisfies* (Equation 3.5) (Equation 3.6)*, if* c *is also an analytical function, then* c *takes the form* $\Psi(x) + \nabla u(x)^T y + \Phi(y)$*, where* Ψ, u, Φ *are analytical functions on* $\mathbb{R}^d$*, and* u *is strictly convex.*

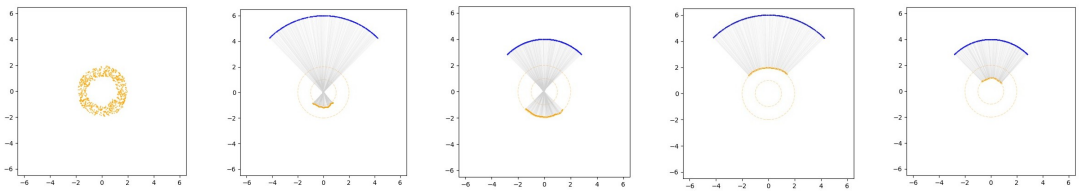
3.2.4 Experiments

In this section, we first conduct experiments to compute Monge maps under different cost functions in order to justify the correctness of our method; we then test the effectiveness of our algorithm on distributions that are either non absolute continuous or supported to spaces with unequal dimensions. We also test our method on high dimensional realistic data set to show the effective ness of the algorithm under different choices of cost functions. At last, we compare the accuracy of our algorithm quantitatively with existing methods on Gaussian examples. The details of the experiments including hyper parameter choices are provided in the appendix subsection B.1.3. We also refer the readers to more high dimensional examples discussed in [20, 39].

Effect of different costs

Next we test our algorithm with more general cost functions. We compare the results on the same set of problems but with different choice of costs, and illustrate the effects of different cost functions.

Inverse function as cost We consider the cost function $c(x,y) = \phi(|x-y|)$ with ϕ as a monotonic decreasing function. We test our algorithm for a specific example $\phi(s) = \frac{1}{s^2}$. In this example, we compute for the optimal Monge map from μ to ν with μ as a uniform distribution on Ω_a and ν as a uniform distribution on Ω_b, where we define $\Omega_a = \{(x_1,x_2) \mid 6^2 \geq x_1^2 + x_2^2 \geq 4^2\}$, $\Omega_b = \{(x,x_2) \mid 2^2 \geq x_1^2 + x_2^2 \geq 1^2\}$. We also compute the same problem for L^2 cost. Figure 3.1 shows the transported samples as well as the differences between two cost functions.



(a) (b) (c) (d) (e)

Figure 3.1: (a) samples of computed $T_\sharp \rho_a$; $c(x,y) = \frac{1}{|x-y|^2}$: Computed Monge map of quarter circles with radius 6 (subplot b) and radius 4 (subplot c); $c(x,y) = |x-y|^2$: Computed Monge map of quarter circles with radius 6 (subplot d) and radius 4(subplot e).

Monge problem on sphere For a given sphere S with radius R, for any two points $x, y \in S$, we define the distance $d(x,y)$ as the length of the geodesic joining x and y. Now for given ρ_a, ρ_b defined on S, we consider solving the following Monge problem on S

$$\min_{T,\ T_\sharp \rho_a = \rho_b} \int_S d(x, T(x)) \rho_a(x)\, dx. \tag{3.8}$$

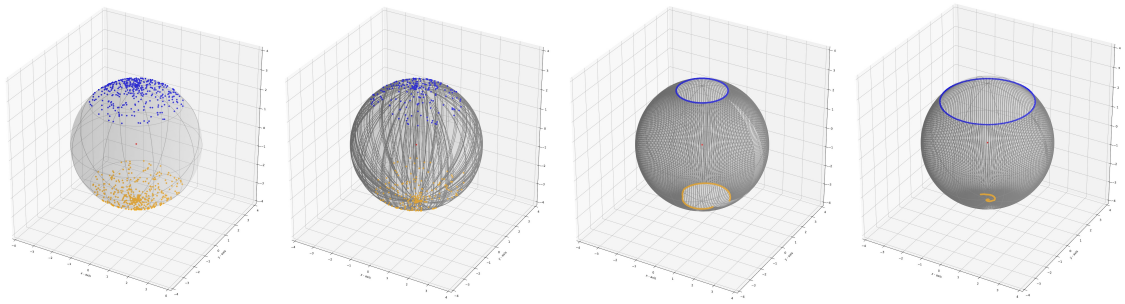
Such sphere OT problem can be transferred to an OT problem defined on angular domain $D = [0, 2\pi) \times [0, \pi]$, to be more specific, we consider (θ, ϕ) ($\theta \in [0, 2\pi)$, $\phi \in [0, \pi]$) as the azimuthal and polar angle of the spherical coordinates. For two points $x = (R\sin\phi_1 \cos\theta_1, R\sin\phi_1 \sin\theta_1, R\cos\phi_1)$, $y = (R\sin\phi_2 \cos\theta_2, R\sin\phi_2 \sin\theta_2, R\cos\phi_2)$ on S, the geodesic distance

$$d(x,y) = c((\theta_1, \phi_1), (\theta_2, \phi_2)) = R \cdot \arccos(\sin\phi_1 \sin\phi_2 \cos(\theta_2 - \theta_1) + \cos\phi_1 \cos\phi_2).$$

Denote the corresponding distribution of μ, ν on D as $\hat{\mu}, \hat{\nu}$, now (Equation 3.8) can also be formulated as

$$\min_{\hat{T}, \hat{T}_\sharp \hat{\mu} = \hat{\nu}} \int c((\theta, \phi), \hat{T}(\theta, \phi)) \hat{\mu}\, d\theta d\phi. \tag{3.9}$$

We set $\hat{\mu} = U([0, 2\pi]) \otimes U([0, \frac{\pi}{4}])$ and $\hat{\nu} = U([0, 2\pi]) \otimes U([\frac{3\pi}{4}, \pi])$. We apply our algorithm to solve (Equation 3.9) and then translate our computed Monge map back to the sphere S to obtain the following results.



(a) (b) (c) (d)

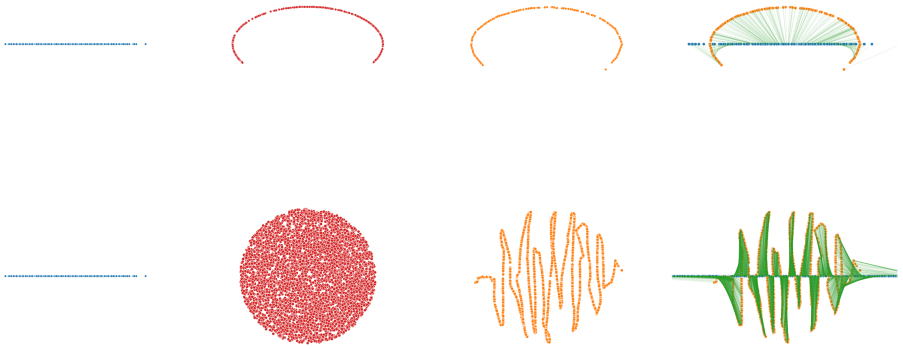
Figure 3.2: Monge map from μ to ν on the sphere: (a) blue samples from μ (corresponds to $\hat{\mu}$) and orange samples from ν (corresponds to $\hat{\nu}$); (b) blue samples from μ, orange samples are obtained from $\hat{T}_\sharp\hat{\nu}$, grey curves are geodesics connecting each transporting pairs; (c) our computed Monge map maps blue ring ($\phi = \frac{\pi}{8}$) to the orange curve (ground truth is $\phi = \frac{7}{8}\pi$); (d) our computed Monge map maps blue ring ($\phi = \frac{\pi}{4}$) to the orange curve (ground truth is the southpole)

Learning with unequal dimensions

Our algorithm framework enjoys a distinguishing quality that it can learn the map from a lower dimension space $\mathbb{R}^{d_x}$ to a manifold in a higher dimension space $\mathbb{R}^{d_y}(d_x \leq d_y)$. In this scenario, we make the input dimension of neural network T to be d_x and output dimension to be d_y. In case the cost function $c(x, y)$ requires dimensions are x and y are equal dimensional, we patch zeros behind each sample $X \sim \mu$ and complement to a counterpart sample $\widetilde{X} = [X; \mathbf{0}]$, where dimension of $\mathbf{0}$ is $d_y - d_x$. And the targeted min-max problem is replaced by

$$\max_{\theta} \min_{\eta} \frac{1}{N} \sum_{k=1}^{N} c(\widetilde{X}_k, T_\theta(X_k)) - f_\eta(T_\theta(X_k)) + f_\eta(Y_k).$$

In Figure 3.3, we conduct two experiments for $d_x = 1$ and $d_y = 2$. Similarly, each row is shown as an example. The incomplete ellipse is a 1D manifold and our algorithm is able to learn a symmetric map from $\mathcal{N}(0, 1)$ towards it. The second row is when the support of ν is in a higher dimension manifold than μ. In this case, our method pushforwards μ to samples that are attempting to fill the space of the ball.



(a) ρ_a (b) ρ_b (c) $T\sharp\rho_a$ (d) $T(\cdot)$ map

Figure 3.3: Qualitative results for learning unequal dimension maps. ρ_a for two examples are both $\mathcal{N}(0, 1)$, and ρ_b are uniformly distributed on a incomplete ellipse and a ball respectively.

Experiments in high dimensions

Next we test our algorithm for high dimensional data set. We compare the results on the same set of problem but with different choice of costs.

Examples in 256D space

KL divergence vs L^2 cost In this experiment, we study the digits transfer map from the data set taken from Modified National Institute of Standards and Technology database (MNIST) (μ, scaled to 16×16 dimensional) to the data taken from US Postal (USPS) ($\nu, 16 \times 16$ dimensional) handwritten digits data sets. The USPS data is derived from a project on recognizing handwritten digits on envelopes, mentioned in [73]. The MNIST dataset, one of the most famous in digit recognition, is created by [74]. As we see from Figure 3.4, the style of MNIST digit number is thinner in (a) while the style of USPS digit number is larger and rounder in (b). We choose cost functions as $|\cdot|_2^2$ and KL divergence respectively. To apply KL divergence, we force L^1 norm of each sample is equal to one by a softmax normalization.

(a) ρ_a (b) ρ_b (c) $T\sharp\rho_a$ (d) $T\sharp\rho_a$

Figure 3.4: MNIST maps to USPS in 256D (c): $\|\cdot\|_2^2$ cost, (d): KL divergence cost.

Learning the map between Gaussians with L^2 cost

In this section, we test our method on Gaussian marginal setting to investigate the quantitative performance. We follow the experiment setup exactly in W2GN [19]. The error is quantified as L^2-UVP $= 100\cdot[\|T-T_*\|^2_{\rho_a}/\mathrm{Var}(\rho_b))]\%$. The marginal μ, ν are two randomly generated centered Gaussian distributions. We refer to Section 5.1, Section C.4 of [19] for the performance of W2-OT and W2GN methods. In Figure 3.5, we rerun the experiment in each dimension for 5 times and report the error bar. When $d < 64$, the L^2-UVP is lower than 1%, which is on par with the performance of W2-OT and W2GN. And L^2-UVP in $d \geq 64$ is still bounded by 3%.

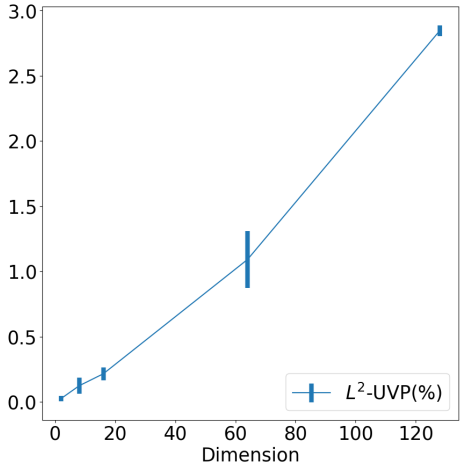


Figure 3.5: Quantitative comparison in Gaussian marginals setting with L^2 cost.

3.2.5 Conclusion

We present a novel method to compute the Monge map between two given distributions with general cost functions. By applying Lagrange multipliers to the Monge problem, we come up with a max-min saddle point problem. By introducing neural networks as the transport map as well as the multiplier, we propose a scalable algorithm that can handle most general costs and even the case where the dimensions of marginals are unequal.

Our method not only computes sample based Wasserstein distance, but also produces the Monge map. The correctness, effectiveness and accuracy of our scheme has been verified through a series of experiments varying from low dimensional examples to high dimensional data sets.

The proposed method may find its applications in machine learning research. It can serve as a useful tool for domain adaption that requires transforming data distributions; it may also find its diverse applications in generative models. Our method also has the potential to be applied to research fields such as computer vision and robotic control problems.

3.3 Approximating the Optimal Transport Plan via Particle-Evolving Method

3.3.1 Introduction

As we have mentioned in the previous section, optimal transport provides a flexible framework for comparing probability measures.

In this section, we aim at solving for the optimal coupling γ_* of the Kantorovich problem (Equation 2.6). However, instead of directly dealing with the standard Kantorovich problem with strict marginal constraints, we consider the so-called entropy transport (ET) problem, which can be treated as a relaxed Kantorovich problem with soft marginal constraints. Recently, the importance of ET problem has drawn researchers' attention due to its rich theoretical properties [75]. By restricting the ET problem to probability manifold and formulating the Wasserstein gradient flow of the target functional, we derive a time-evolutional Partial Differential Equation (PDE) that can be then realized by evolving an interacting particle system via Kernel Density Estimation techniques [76].

Our method directly computes for the **sample-wised approximation** of the optimal coupling γ_* to the OT problem (Equation 2.6) between two **known densities**. That is to say, given the density functions ρ_a, ρ_b (no need to be normalized[1]) of the marginal distributions, our algorithm is capable of generating samples that approximate the optimal coupling γ_*. This is very different from traditional methods such as Linear Programming [47, 48], Monge-Ampère Equation [43, 44, 45, 46], and dynamical scheme [6, 77], which all require discretization of the continuous space. Our method is also different from the Sinkhorn algorithm [12], which relies samples from marginals and computes for the optimal coupling on discrete data set; as well as methods involving neural network approximations [13, 78, 18, 20], which also require marginal samples and directly approximates the Monge map T_* or the Kantorovich dual pair (ψ_*, ϕ_*). We note that a recent independent work [79] on sampling algorithm for Wasserstein Barycenter problems shares similar ideas with our

[1]For example, if the probability densities of marginal distributions are $\rho_a(\cdot) = \frac{1}{Z_a} f_a(\cdot)$, $\rho_b(\cdot) = \frac{1}{Z_b} f_b(\cdot)$. Then our algorithm can still operate if we are only given the unnormalized densities f_a, f_b.

proposed method.

Our main contribution is to analyze the theoretical properties of the entropy transport problem constrained on probability space and derive its Wasserstein gradient flow. To be specific, we study the existence and uniqueness of the solution to ET problem and further study its Γ-convergence property to the classical OT problem. Then based on the gradient flow that we derive, we propose an innovative particle-evolving algorithm for obtaining the sample approximation of the optimal transport plan. Our method can deal with optimal transport problem between two known unnormalized densities. To the best of our knowledge, there is no method capable of solving this type of problem. We demonstrate the efficiency of our method by numerical experiments.

We refer the readers to subsection 3.1.1 and subsection 3.1.2 for related references.

We refer the readers to subsection 2.1.2, subsection 2.2.1 and subsection 2.2.2 for related mathematical backgrounds.

3.3.2 Constrained entropy transport as a regularized optimal transport problem

Optimal transport problem and its relaxation

In this work, we mainly consider the Kantorovich problem (Equation 2.6) with $X = Y = \mathbb{R}^d$. For the following discussion of section 3.3, we call the optimal solution of (Equation 2.6) as **optimal transport plan** and we denote it as γ_{OT}.

We can also reformulate (Equation 2.6) as $\min_{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \{\mathcal{E}_\iota(\gamma|\mu,\nu)\}$ where

$$\mathcal{E}_\iota(\gamma|\mu,\nu) = \iint c(x,y)d\gamma(x,y) + \int \iota\left(\frac{d\gamma_1}{d\mu}\right)d\mu + \int \iota\left(\frac{d\gamma_2}{d\nu}\right)d\nu \tag{3.10}$$

Here ι is defined as $\iota(1) = 0$ and $\iota(s) = +\infty$ when $s \neq 1$. We now derive a relaxed version of (Equation 2.6) by replacing $\iota(\cdot)$ with $\Lambda F(\cdot)$, where $\Lambda > 0$ is a tunable positive parameter

and we assume

$$F \text{ is a } \textbf{smooth convex} \text{ function with } F(1) = 0, \text{ and } 1 \text{ as the unique minimizer.} \tag{3.11}$$

$$F \text{ is superlinear, i.e., } \lim_{x\to\infty} \frac{F(x)}{|x|} = +\infty. \tag{3.12}$$

In our research, we mainly focus on $F(s) = s\log s - s + 1$ [75]. It is worth mentioning that after such relaxation, the constraint term $\int F(\frac{d\gamma_1}{d\mu})\, d\mu$ is known as the Kullback-Leibler (KL) divergence [80] and is denoted as $\mathcal{D}_{\mathrm{KL}}(\gamma_1\|\mu)$.

From now on, we should focus on the following functional involving cost

$$c(x, y) = h(x - y) \text{ with } h \text{ as a strictly convex function,} \tag{3.13}$$

and enforcing the marginal constraints by using KL-divergence:

$$\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) = \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\, d\gamma(x, y) + \Lambda D_{\mathrm{KL}}(\gamma_1\|\mu) + \Lambda D_{\mathrm{KL}}(\gamma_2\|\nu). \tag{3.14}$$

Constrained Entropy Transport problem and its properties

For the following discussions, we always assume that the marginal distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\mu \ll \mathscr{L}^d,\ \nu \ll \mathscr{L}^d$. We now focus on solving the following problem

$$\min_{\gamma\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)\}. \tag{3.15}$$

A similar problem

$$\min_{\gamma\in\mathcal{M}(\mathbb{R}^d\times\mathbb{R}^d)} \{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)\} \tag{3.16}$$

has been studied in [75] with $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ being replaced by the space of positive measures $\mathcal{M}(\mathbb{R}^d\times\mathbb{R}^d)$ and is named as **Entropy Transport (ET) problem** therein. In our case, since we restrict γ to probability space, we call (Equation 3.15) **constrained Entropy Transport**

(cET) problem and call $\mathcal{E}_{\Lambda,\mathrm{KL}}$ the Entropy Transport functional.

The following theorem shows the existence of the optimal solution to cET problem (Equation 3.15). It also describes the relationship between the solution to the cET problem (Equation 3.15) and the solution to the general ET problem (Equation 3.16):

Theorem 3.3.1. *The infimum value* $\mathcal{E}_{\min} = \inf_{\gamma\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)}\{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)\}$ *is finite. There always exits an optimal solution* $\tilde{\gamma}$ *to the ET problem* (Equation 3.16)*. We denote* $\gamma = \frac{1}{Z}\tilde{\gamma}$*, here* $Z = e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}$*. Then* $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ *is the optimal solution to cET problem* (Equation 3.15).

Proof of Theorem 3.3.1. First, we prove that $\mathcal{E}_{\min} = \inf\limits_{\gamma\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)$ is finite.

By choosing $\gamma = \mu\otimes\nu$, i.e. choose γ as the direct product of μ,ν, we have

$$\mathcal{E}_{\Lambda,\mathrm{KL}}(\mu\otimes\nu|\mu,\nu) = \iint c d\mu\otimes\nu \geq \min_{x\in\mathbb{R}^d}\{h(x)\},$$

which is finite value given that h defined in Equation 3.11 is convex. One can thus prove that

$$\inf_{\gamma\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) \geq \min_{x\in\mathbb{R}^d}\{h(x)\}.$$

Thus the infimum value is finite.

Second, the existence of $\tilde{\gamma}$ is guaranteed in Theorem B.2.1 stated in appendix .

Then, for any $\tilde{\sigma}\in\mathcal{M}(\mathbb{R}^d\times\mathbb{R}^d)$, we can write it as:

$$\tilde{\sigma} = M\sigma$$

with $M = \tilde{\sigma}(\mathbb{R}^d \times \mathbb{R}^d)$ and $\sigma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. Now one can write $\mathcal{E}_{\Lambda,\mathrm{KL}}(\tilde{\sigma}|\mu,\nu)$ as:

$$
\begin{aligned}
&\mathcal{E}_{\Lambda,\mathrm{KL}}(\tilde{\sigma}|\mu,\nu) \\
=&\iint c(x,y)\, d(M\sigma) + \Lambda \int \left(\frac{d\pi_{1\sharp}(M\sigma)}{d\mu} \log \left(\frac{d\pi_{1\sharp}(M\sigma)}{d\mu} \right) - \frac{d\pi_{1\sharp}(M\sigma)}{d\mu} + 1 \right) d\mu \\
&+ \Lambda \int \left(\frac{d\pi_{2\sharp}(M\sigma)}{d\nu} \log \left(\frac{d\pi_{2\sharp}(M\sigma)}{d\nu} \right) - \frac{d\pi_{2\sharp}(M\sigma)}{d\nu} + 1 \right) d\nu \\
=&M\mathcal{E}_{\Lambda,\mathrm{KL}}(\sigma|\mu,\nu) + 2\Lambda(M\log M - M) + 2\Lambda.
\end{aligned}
$$

The optimization problem (Equation 3.16) on $\mathcal{M}(\mathbb{R}^d \times \mathbb{R}^d)$ can now be formulated as:

$$
\inf_{\sigma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \min_{M \geq 0} \ \left\{ M\mathcal{E}_{\Lambda,\mathrm{KL}}(\sigma|\mu,\nu) + 2\Lambda(M\log M - M) + 2\Lambda \right\}.
$$

It is not hard to verify that when σ is fixed, we denote $E(\sigma) = \mathcal{E}_{\Lambda,\mathrm{KL}}(\sigma|\mu,\nu)$ for shorthand. Then the minimum value of $ME(\sigma) + 2\Lambda(M\log M - 1) + 2\Lambda$ ($M \geq 0$) is achieved at $M = e^{-\frac{E(\sigma)}{2\Lambda}}$ and the minimum value is $2\Lambda(1 - e^{-\frac{E(\sigma)}{2\Lambda}})$. Recall definition of $\mathcal{E}_{\min}$ in Theorem 3.3.1, we have:

$$
\begin{aligned}
&\inf_{\sigma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \min_{M \geq 0} \ \left\{ M\mathcal{E}_{\Lambda,\mathrm{KL}}(\sigma|\mu,\nu) + 2\Lambda(M\log M - M) + 2\Lambda \right\} \\
=&\inf_{\sigma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ 2\Lambda(1 - e^{-\frac{E(\sigma)}{2\Lambda}}) \right\} = 2\Lambda(1 - e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}).
\end{aligned}
$$

Since $\tilde{\gamma}$ solves (Equation 3.16), we have:

$$
\begin{aligned}
\mathcal{E}_{\Lambda,\mathrm{KL}}(\tilde{\gamma}|\mu,\nu) =& \inf_{\sigma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \min_{M \geq 0} \ \left\{ M\mathcal{E}_{\Lambda,\mathrm{KL}}(\sigma|\mu,\nu) + 2\Lambda(M\log M - M) + 2\Lambda \right\} \\
=&2\Lambda(1 - e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}).
\end{aligned}
$$

Now we write $\tilde{\gamma} = Z\gamma$, with $Z = \tilde{\gamma}(\mathbb{R}^d \times \mathbb{R}^d)$, $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. We have:

$$
Z\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) + 2\Lambda(Z\log Z - Z) + 2\Lambda = 2\Lambda(1 - e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}})
$$

However, we have:

$$Z\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) + 2\Lambda(Z\log Z - Z) + 2\Lambda \geq 2\Lambda(1 - e^{-\frac{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)}{2\Lambda}}). \tag{3.17}$$

This gives:

$$2\Lambda(1 - e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}) \geq 2\Lambda(1 - e^{-\frac{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)}{2\Lambda}}) \quad \Rightarrow \quad \mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) \leq \mathcal{E}_{\min}.$$

As a result, we have: $\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu) = \mathcal{E}_{\min}$, i.e. γ solves problem (Equation 3.15). And inequality (Equation 3.17) becomes equality, this shows $Z = e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}$. □

The following corollary guarantees the uniqueness of optimal solution to (Equation 3.15):

Corollary 3.3.1.1 (Existence & Uniqueness)**.** *The cET problem* (Equation 3.15) *admits a unique optimal solution.*

Proof. We still assume that $\tilde{\gamma}$ and γ are solutions to (Equation 3.16) and (Equation 3.15) respectively as stated in Theorem 3.3.1. Suppose despite γ, we have another $\gamma' \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ that also solves (Equation 3.15) . Set $Z = e^{-\frac{\mathcal{E}_{\min}}{2\Lambda}}$, we can verify that $\mathcal{E}_{\Lambda,\mathrm{KL}}(Z\gamma|\mu,\nu) = \mathcal{E}_{\Lambda,\mathrm{KL}}(Z\gamma'|\mu,\nu)$. This means that $Z\gamma' \neq Z\gamma$ (i.e. $Z\gamma' \neq \tilde{\gamma}$) is another solution to problem (Equation 3.16). This avoids the uniqueness stated in Theorem B.2.1. □

Despite the discussions on the constrained problem (Equation 3.15) with fixed Λ, we also establish asymptotic results for (Equation 3.15) with quadratic cost $c(x,y) = |x-y|^2$ as $\Lambda \to +\infty$. For the rest of this section, we define:

$$\mathscr{P}_2(\mathbb{R}^d) = \left\{ \gamma \,\middle|\, \gamma \in \mathcal{P}(\mathbb{R}^d),\ \int_{\mathbb{R}^d} |x|^2 d\gamma(x) < +\infty \right\}.$$

Let us now consider $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ and assume it is equipped with the topology of weak convergence. We are able to establish the following Γ-convergence result:

Theorem 3.3.2 (Γ-convergence)**.** *Suppose* $c(x,y) = |x-y|^2$*. Assume that we are given* $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$*,* $\mu \ll \mathscr{L}^d, \nu \ll \mathscr{L}^d$ *and at least one of* μ *and* ν *satisfies the Logarithmic Sobolev inequality with constant* $K > 0$*. Let* $\{\Lambda_n\}$ *be a positive increasing sequence, satisfying* $\lim_{n\to\infty} \Lambda_n = +\infty$*. We consider the sequence of functionals* $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$*. Recall the functional* $\mathcal{E}_\iota(\cdot|\mu,\nu)$ *defined in* (Equation 3.10)*. Then* $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ Γ- *converges to* $\mathcal{E}_\iota(\cdot|\mu,\nu)$ *on* $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$*.*

We can further establish the equi-coercive property for the family of the functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ and we use the Fundamental Theorem of Γ-convergence [81] [82] to establish the following asymptotic results:

Theorem 3.3.3 (Property of Γ-convergence)**.** *Suppose* $c(x,y) = |x-y|^2$*. Assuming* $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$*,* $\mu \ll \mathscr{L}^d, \nu \ll \mathscr{L}^d$*, and both* μ, ν *satisfy the Logarithmic Sobolev inequality with constants* $K_\mu, K_\nu > 0$*. According to Corollary 3.3.1.1, problem* (Equation 3.15) *with functional* $\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)$ *admits a unique optimal solution, let us denote it as* γ_n*. According to Theorem 2.1.2, the Kantorovich problem* (Equation 2.6) *also admits a unique solution, we denote it as* γ_{OT}*. Then* $\lim_{n\to\infty} \gamma_n = \gamma_{OT}$ *in* $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$*.*

The detailed proofs for Theorem 3.3.2 and Theorem 3.3.3 are provided in appendix subsection B.2.2.

3.3.3 Wasserstein Gradient Flow Approach

Wasserstein gradient flow of Entropy Transport functional

As mentioned in subsection 2.2.2, there are already numerous researches [83, 4, 35] regarding Wasserstein gradient flows of different types of functionals defined on the Wasserstein manifold that successfully relate certain kinds of time evolution Partial Differential Equations (PDEs) to the manifold gradient of corresponding functionals.

We now consider our constrained Entropy Transport problem (Equation 3.15). There

are mainly two reasons why we choose to compute the Wasserstein gradient flow of functional $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot|\mu,\nu)$:

- Computing the Wasserstein gradient flow is equivalent to applying gradient descent to determine the minimizer of the entropy transport functional (Equation 3.14);
- In most of the cases, Wasserstein gradient flows can be realized as a time evolution PDE describing the density evolution of a stochastic process. As a result, once we derived the gradient flow, there will be a natural particle version associated to the gradient flow. And this will make the computation of gradient flow tractable since we can evolve the particle system by applying time discretization scheme.

Now let us compute the Wasserstein gradient flow of $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot|\mu,\nu)$:

$$\frac{\partial \gamma_t}{\partial t} = -\mathrm{grad}_W \mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma_t|\mu,\nu), \quad \gamma_t|_{t=0} = \gamma_0 \tag{3.18}$$

To keep our notations concise, we denote $\rho(\cdot,\cdot,t) = \frac{d\gamma_t}{d\mathscr{L}^{2d}}$ $\varrho_1 = \frac{d\mu}{d\mathscr{L}^d}$, $\varrho_2 = \frac{d\nu}{d\mathscr{L}^d}$, we can show that the previous equation (Equation 3.18) can be written as:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \, \nabla(c(x,y) + \Lambda \log(\frac{\rho_1(x,t)}{\varrho_1(x)}) + \Lambda \log(\frac{\rho_2(y,t)}{\varrho_2(y)}))) \tag{3.19}$$

Here $\rho_1(\cdot,t) = \frac{d\pi_{1\sharp}\gamma_t}{d\mathscr{L}^d} = \int \rho(\cdot,y,t)dy$ and $\rho_2(\cdot,t) = \frac{d\pi_{2\sharp}\gamma_t}{d\mathscr{L}^d} = \int \rho(x,\cdot,t)dx$ are density functions of marginals of γ_t. We put the details of our derivation in appendix subsection B.2.3.

Remark 6. *We are currently not clear about the **displacement convexity** [7] of the functional $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot|\mu,\nu)$ on $(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d), g^W)$, which will guarantee its gradient flow to converge to its minimizer. This will be one of our future research directions . In practice, we should rely on the computational results to tell us whether our method works properly.*

3.3.4 Particle formulation

Let us treat (Equation 3.19) as certain kind of continuity equation, i.e. we treat $\rho(\cdot, t)$ as the density of the time-evolving random particles. Then the vector field that drives the random particles at time t should be $-\nabla(c(x,y) + \Lambda \log\left(\frac{\rho_1(x,t)}{\varrho_1(x)}\right) + \Lambda \log\left(\frac{\rho_2(y,t)}{\varrho_2(y)}\right))$. This helps us design the following dynamics $\{(X_t, Y_t)\}_{t\geq 0}$: (here $\dot{X}_t$ denotes the time derivative $\frac{dX_t}{dt}$)

$$\begin{cases} \dot{X}_t = -\nabla_x c(X_t, Y_t) + \Lambda(\nabla \log \varrho_1(X_t) - \nabla \log \rho_1(X_t, t)); \\ \dot{Y}_t = -\nabla_y c(X_t, Y_t) + \Lambda(\nabla \log \varrho_2(Y_t) - \nabla \log \rho_2(Y_t, t)); \end{cases} \tag{3.20}$$

where $\text{Law}(X_0, Y_0) = \gamma_0$. Here $\rho_1(\cdot, t)$ is the density of $\text{Law}(X_t)$ and $\rho_2(\cdot, t)$ is the density of $\text{Law}(Y_t)$. If we assume the process (Equation 3.20) is well-defined, then the probability density $\rho_t(x, y)$ of (X_t, Y_t) should solve the PDE (Equation 3.19).

When we take a closer look at (Equation 3.20), we can verify that the movement of particle (X_t, Y_t) at certain time t depends on the probability density of $\text{Law}((X_t, Y_t))$ at (X_t, Y_t), which can be approximated by the distribution of the surrounding particles near (X_t, Y_t). Such equation (Equation 3.19) can be treated as a limit case of aggregation-diffusion equation [84, 85] with Dirac kernel convolution. Generally speaking, we plan to evolve (Equation 3.20) as a particle aggregation model in order to produce to a sample-wised approximation of the optimal tansport plan γ_{OT} for Kantorovich problem (Equation 2.6).

3.3.5 Proposed algorithm

To simulate the stochastic process (Equation 3.20) with the Euler scheme, we apply the Kernel Density Estimation [76] here to approximate $\nabla \log \rho(x)$ by convolving it with certain kernel function $K(x, \xi)$[2]:

$$\nabla \log \rho(x) \approx \nabla \log(K * \rho)(x) = \frac{(\nabla_x K) * \rho(x)}{K * \rho(x)} \tag{3.21}$$

[2]In this research, we choose the Radial Basis Function (RBF) as the kernel: $K(x, \xi) = \exp(-\frac{|x-\xi|^2}{2\tau^2})$.

Here $K * \rho(x) = \int K(x,\xi)\rho(\xi)d\xi$, $(\nabla_x K) * \rho(x) = \int \nabla_x K(x,\xi)\rho(\xi)d\xi$[3]. Such technique is also known as blobing method [85][86]. With the blobing method, $\nabla \log \rho(x)$ is evaluated based on the locations of the particles:

$$\frac{\mathbb{E}_{\xi\sim\rho}\nabla_x K(x,\xi)}{\mathbb{E}_{\xi\sim\rho}K(x,\xi)} \approx \frac{\sum_{k=1}^N \nabla_x K(x,\xi_k)}{\sum_{k=1}^N K(x,\xi_k)} \quad \xi_1,...,\xi_N, \text{i.i.d.} \sim \rho$$

Now we are able to simulate (Equation 3.20) with the following interacting particle system involving N particles $\{(X_i, Y_i)\}_{i=1,...,N}$. For the i-th particle, we have:

$$\begin{cases} \dot{X}_i(t) = -\nabla_x c(X_i(t), Y_i(t)) - \Lambda\left(\nabla V_1(X_i(t)) + \dfrac{\sum_{k=1}^N \nabla_x K(X_i(t), X_k(t))}{\sum_{k=1}^N K(X_i(t), X_k(t))}\right) \\ \dot{Y}_i(t) = -\nabla_y c(X_i(t), Y_i(t)) - \Lambda\left(\nabla V_2(Y_i(t)) + \dfrac{\sum_{k=1}^N \nabla_x K(Y_i(t), Y_k(t))}{\sum_{k=1}^N K(Y_i(t), Y_k(t))}\right) \end{cases} \tag{3.22}$$

Here we denote $V_1 = -\log \varrho_1$, $V_2 = -\log \varrho_2$. Since we only need the gradients of V_1, V_2, as emphasized in subsection 3.3.1, our algorithm is capable of dealing with unnormalized probability measures. When $t \to \infty$, with sufficient large N and Λ, we can numerically verify that the empirical distribution $\frac{1}{N}\sum_{i=1}^N \delta_{(X_i(t),Y_i(t))}$ will converge to the optimal distribution γ_{cET} of (Equation 3.15). We provide the algorithm and discussion on random batch method in subsection B.2.4.

3.3.6 Experiments

We test our algorithm on several illustrative examples. The experiments are conducted on a computer with 2.4GHz CPU, 15.3GB of memory.

1D Gaussian We consider two 1D Gaussian distributions $\mathcal{N}(-4,1), \mathcal{N}(6,1)$ as the marginal distributions and run our algorithm to compute the sample approximation of the optimal transport plan. In this experiment, we choose $\lambda = 200, \Delta t = 0.001$ and evolve with 1000 particles for 1000 iterations. We initialize our samples $\{(X_i, Y_i)\}$ by sampling $\{X_i\}$

[3]Notice that we always use $\nabla_x K$ to denote the partial derivative of K with respect to the first components.

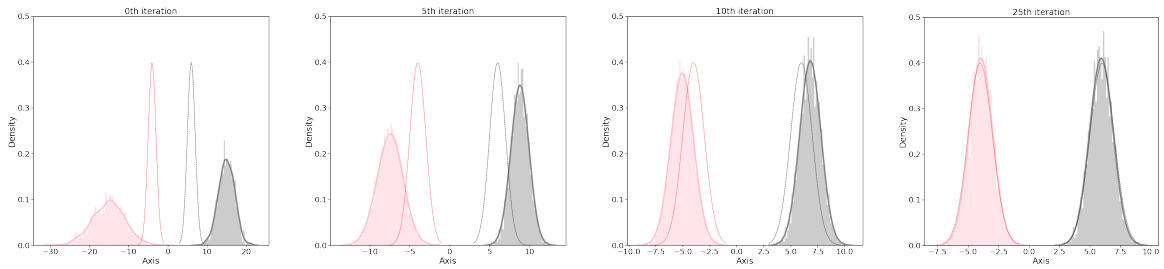


Figure 3.6: Marginal plot for 1D Gaussian example. The red and black dashed curves indicate the two marginal distributions, the solid pink and gray curves are kernel estimated densities of particles at certain iterations. The marginals usually converge fast: after 25 iterations, the marginal samples $\{X_i\}, \{Y_i\}$ already matched with the real marginals very well.

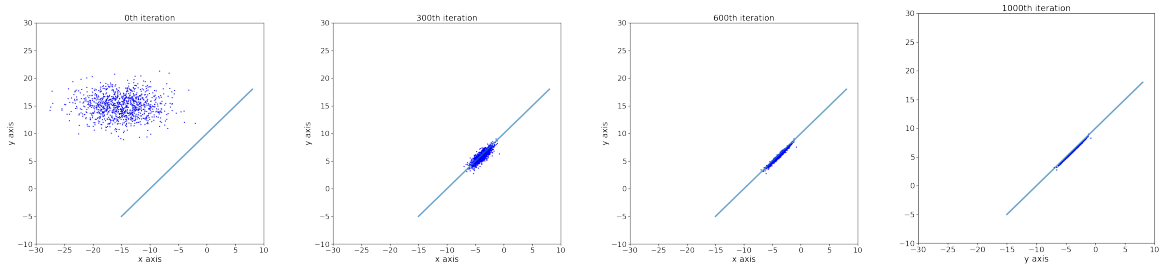


Figure 3.7: The sample approximation for 1D Gaussian example. The blue straight line corresponds to the optimal transport map $T(x) = x + 10$.

from $\mathcal{N}(-15, 4)$, and $\{Y_i\}$ from $\mathcal{N}(15, 2)$. The empirical results are shown in Figure 3.6 and Figure 3.7. We can verify that after 1000 iterations, we obtain a valid approximation of the optimal transport plan.

1D Gaussian Mixture We apply our method to 1D Gaussian mixtures with $\varrho_1 = \frac{1}{2}\mathcal{N}(-1.5, 1) + \frac{1}{2}\mathcal{N}(1.5, 1), \varrho_2 = \frac{1}{2}\mathcal{N}(-4, 2) + \frac{1}{2}\mathcal{N}(4, 2)$. In this experiment, we set $\lambda = 60, \Delta t = 0.0004$ and run with 800 particles (X_i, Y_i)'s for 5000 iterations. We initialize the samples $\{X_i, Y_i\}$ by sampling from $\mathcal{N}((0, 0), 2I_2)$. The well match on the marginal distributions as well as the sample approximation of the optimal transport plan are reflected in Figure 3.8.

For further experiments on synthetic data sets, as well as Wasserstein Barycenter problems, we refer the readers to [22].

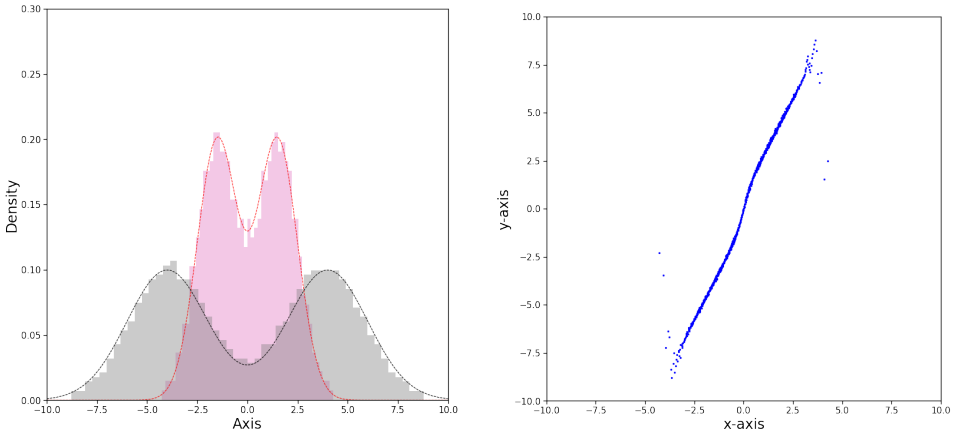


Figure 3.8: 1D Gaussian mixture. Left. Marginal plot. The dash curves are two marginal distributions. The histogram indicates the distribution of the particles after 5000 iterations. Right. Sample approximation of the optimal coupling.

3.3.7 Conclusion

In this research work, we propose the constrained Entropy Transport problem (Equation 3.15) and study some of its theoretical properties. We mainly discover and prove that the optimal distribution of (Equation 3.15) can be treated as an approximation to the optimal transport plan to the original Kantorovich problem (Equation 2.6) in the sense of Γ-convergence. We also derive the Wasserstein gradient flow of the Entropy Transport functional. We propose a novel algorithm that computes for the sample-wised optimal transport plan by evolving an interacting particle system. We test our algorithm on several illustrative examples. Further theoretical analysis on the convergence of Wasserstein gradient flow, as well as numerical experiments on higher dimensional data sets will be considered in our future projects.

3.4 Learning High Dimensional Wasserstein Geodesics

3.4.1 Introduction

As being emphasized in the previous sections, optimal transport distance has been widely used to evaluate the distance between two distributions. Recall subsection 2.1.5, we can recast the classical OT problem (Equation 2.2),(Equation 2.6) as an optimal control problem involving dynamical process [61]. For example, by setting $L(v) = |v|^2$, one obtains

$$W_2^2(\mu,\nu) = \inf\left\{\int_0^1 \int_{\mathbb{R}^d} \rho(x,t)|v(x,t)|^2 dxdt\right\},$$
$$\text{subject to: } \partial_t\rho(x,t) + \nabla\cdot(\rho(x,t)v(x,t)) = 0, \quad \rho(\cdot,0) = \rho_a, \ \ \rho(\cdot,1) = \rho_b. \tag{3.23}$$

Here and for the following part of this section, we denote ρ_a, ρ_b as the densities of distributions μ, ν. Recall Example 2.1.4, the geodesic equation for solving (Equation 3.23) is characterized by the PDE system

$$\partial_t\rho + \nabla\cdot(\rho\nabla\Phi) = 0, \quad \frac{\partial\Phi}{\partial t} + \frac{1}{2}|\nabla\Phi|^2 = 0, \tag{3.24}$$
$$\text{subject to: } \ \rho(\cdot,0) = \rho_a, \quad \rho(\cdot,1) = \rho_b.$$

Knowing the Wasserstein geodesic between μ and ν (We refer the readers to Figure 2.3, trajectory of the particle movement is indicated by the red line.) provides ample information for their Wasserstein distance and optimal transport map. More importantly, since the Wasserstein geodesic is automatically energy-minimizing, it offers a natural sampling mechanism without using additional artificial regularization to generate samples not only for the target distribution ν, but also for all distributions along the Wasserstein geodesic. This is different from several recent OT based models for computing the optimal transport map, such as Jacobian and Kinetic regularized OT [87] and L^2 regularized OT [88].

Wasserstein geodesics also find applications in robotics and optimal control researches.

In [65], the authors apply Brenier-Benamou OT to swarm control and updates the velocity of each agent. In [66], people study the locations of robots by minimizing the Wasserstein distance between original and target distributions. Investigations on diverse OT applications in control theories are provided in [89, 90]. Currently, the research that combines OT with robotics or control is still limited to low dimensions. We believe, having a method to compute the Wasserstein geodesic, especially in high dimensional settings, will be beneficial for developing novel algorithms and applications in robotics and control, such as path planning for multi-agent systems. Furthermore, various OT models bring numerous applications in domain adaptation [13], generative models [8], inverse problems related to stochastic dynamics [67] as well as color transfer [19], which is also one of the experiments in this research.

Last but not least, finding an efficient method to compute the Wasserstein geodesic is important and challenging in applied mathematics as well. It is well-known that directly solving Problem (Equation 3.23) or (Equation 3.24) by the traditional numerical PDE methods, such as finite difference or finite element method which requires spatial discretization, must face the *curse of dimensionality*, meaning that the computational cost grows exponentially as the dimension increases.

In our treatment, we first formulate the OT problem as a saddle point problem without introducing any regularizers. We further reduce the search space for the saddle point problem by leveraging KKT conditions. By parametrizing the drifting vector field as well as the Lagrange multipliers via deep neural networks, we perform our training process alternately. The resulting method is a sample based algorithm that is capable of handling high dimensional Wasserstein geodesic. We summarize our contributions as following:

- We develop a novel saddle point formulation so that high dimensional Wasserstein geodesic, optimal map as well as Wasserstein distance between two given distributions can be computed in one single framework.

- Our scheme is formulated to handle general convex cost functions, including the general

L^p-Wasserstein distance. More importantly, it provides a method without requiring convexity or Lipschitz constraint. Such constraints are usually considered as thorny issues since they can only be roughly enforced in many proposed methods for optimal transport problems.

- We show the effectiveness of our method through extensive numerical experiments with both synthetic and realistic data sets.

We refer the readers to subsection 3.1.1 and subsection 3.1.3 for related references.

It is worth mentioning that in [17], the authors compute high dimensional Wasserstein geodesic by relaxing the terminal constraint and incorporating it in the cost functional, thus they are computing for a slightly biased OT problem, which is different from our propose scheme.

We also note that a similar strategy formulated by [91] derives a saddle point optimization scheme for solving the mean field game equations. We should point out that our problem setting and sampling method are distinct from theirs.

We refer the readers to subsection 2.1.5 and subsection 2.1.6 for related mathematical backgrounds of this section.

3.4.2 Proposed method

Primal-Dual based saddle point scheme

Let us recall that in subsection 2.1.5, we consider the Lagrange multiplier method for solving the dynamical OT problem (Equation 2.20), which is a constrained minimization problem. We consider the functional: (In the following discussion of this section, we denote $\int$ as $\int_{\mathbb{R}^d}$ for simplicity.)

$$\mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b) = \int_0^1 \int \left(L(v) - \frac{\partial \Phi}{\partial t} - \nabla \Phi \cdot v \right) \rho(x,t)\, dxdt + \int (-\Psi_a \rho_a - \Psi_b \rho_b)\, dx$$
$$+ \int (\Psi_a(x) - \Phi(x,0))\rho(x,0)\, dx + \int (\Psi_b(x) + \Phi(x,1))\rho(x,1)\, dx.$$

The dynamical OT problem (Equation 2.20) is equivalent to the saddle point problem

$$\max_{\Phi,\Psi_a,\Psi_b} \min_{\rho,v} \mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b), \tag{Equation 2.25}$$

$$\text{or} \quad \min_{\rho,v} \max_{\Phi,\Psi_a,\Psi_b} \mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b). \tag{Equation 2.26}$$

We want to reduce the number of variables in the saddle point problem (Equation 2.25), (Equation 2.26). This goal can be achieved by incorporating certain equations derived from the KKT conditions (Equation 2.27). Specifically, conditions $\frac{\partial \mathfrak{L}}{\partial \rho} = 0, \frac{\partial \mathfrak{L}}{\partial v} = 0$ yields $\Psi_b(x) = -\Phi(x, 1), \Psi_a(x) = \Phi(x, 0)$, and $v(x, t) = \nabla L^{-1}(\nabla \Phi(x, t))$. Plugging these relations into $\mathfrak{L}(\rho, v, \Phi, \Psi_a, \Psi_b)$, we obtain the following functional of ρ and Φ:

$$\hat{\mathfrak{L}}(\rho, \Phi) = \int_0^1 \int - \left(\frac{\partial \Phi}{\partial t} + H(\nabla \Phi) \right) \rho(x, t)\, dxdt + \int (\Phi(x, 1)\rho_b(x) - \Phi(x, 0)\rho_a(x))\, dx. \tag{3.25}$$

Now instead of solving (Equation 2.25) or (Equation 2.26), we then seek for the saddle points of $\hat{\mathfrak{L}}(\rho, \Phi)$, i.e., we consider

$$\max_{\Phi} \min_{\rho} \hat{\mathfrak{L}}(\rho, \Phi), \tag{3.26}$$

$$\text{or} \min_{\rho} \max_{\Phi} \hat{\mathfrak{L}}(\rho, \Phi). \tag{3.27}$$

Simplification via geodesic pushforward map

In saddle point problems (Equation 3.26) and (Equation 3.27), both variables $\rho(\cdot, t)$ and $\Phi(\cdot, t)$ are time-varying functions. This requires to optimize over rather large space of time-varying functions, which may increase the computational cost as well as the chance of encountering local optima. To mitigate this challenge, recall Theorem 2.1.6 stated in subsection 2.1.5, we can reduce the search space by leveraging the geodesic property of optimal transporting trajectory if L is strictly convex. To be more specific, Theorem 2.1.6 indicates that under the steering of the optimal vector field $v_*(\cdot, t)$, each particle is trans-

porting along geodesic (straight line) with constant speed. This observation further leads to Corollary 2.1.6.1, which asserts

$$\rho_*(\cdot, t) = (\mathrm{Id} + t v_*(\cdot, 0))_\sharp \rho_a. \qquad \text{(Equation 2.46)}$$

Here (ρ_*, v_*) denotes the optimal solution to dynamical OT problem (Equation 2.20).

(Equation 2.46) justifies that the optimal $\rho^*(\cdot, t)$ can be obtained by pushforwarding the initial ρ_a along certain straight lines with initial direction $v^*(\cdot, 0)$. This observation motivates us to restrict the search space of $\{\rho(\cdot, t)\}_{0 \le 0 \le 1}$ on the following set:

$$\{\ \{\rho(\cdot, t)\}_{0 \le t \le 1} \mid \rho(\cdot, t) = (\mathrm{Id} + tF)_\sharp \rho_a \ \text{ for } t \in [0, 1]\ \ \}.$$

Here F is an arbitrary vector field defined on $\mathbb{R}^d$. Under such choice, the time dependent density $\{\rho(\cdot, t)\}$ is now uniquely determined by the vector field F. Combining this with saddle point problems (Equation 3.26), (Equation 3.27), we reformulate our scheme as

$$\min_F \ \max_\Phi \ \mathcal{L}(F, \Phi), \qquad (3.28)$$

$$\text{or } \max_\Phi \ \min_F \ \mathcal{L}(F, \Phi). \qquad (3.29)$$

Here we denote our target functional $\mathcal{L}$ as

$$\mathcal{L}(F, \Phi) = \hat{\mathfrak{L}}((\mathrm{Id} + tF)_\sharp \rho_a, \Phi). \qquad (3.30)$$

In our actual implementation on both schemes (Equation 3.28) and (Equation 3.29), we discovered that the min-max scheme is working much more stable and producing better results than the max-min scheme. Rigorous justification of this phenomena is still under research. But to this stage, our intuition for such phenomena is that the inner optimization over Φ is enforcing the continuity equation constraint (Equation 2.38), which may help

improving the quality of computed $\{\rho(\cdot, t)\}$.

Furthermore, we have the following theoretical guarantee on $\mathcal{L}$:

Theorem 3.4.1. *Denote the optimal solution to* (Equation 2.20) *as* $(\rho^*(x,t), \Phi^*(x,t))$. *Set* $\Phi_0^*(\cdot) = \Phi^*(\cdot, 0)$. *Assume* $\Phi^*(\cdot, t) \in C^2(\mathbb{R}^d)$, *then* $(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*)$ *is the critical point to functional* $\mathcal{L}$, *i.e.*

$$\frac{\partial \mathcal{L}}{\partial F}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0, \quad \frac{\partial \mathcal{L}}{\partial \psi}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0.$$

Furthermore, $\mathcal{L}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = C_{Dym}(\rho_a, \rho_b)$, *where* $C_{Dym}(\rho_a, \rho_b)$ *is denoted as the optimal value of* (Equation 2.20). *By subsection 2.1.6, this is exactly the OT distance between* ρ_a *and* ρ_b *with cost function* $c(x, y) = L(x - y)$.

Theorem 3.4.1 shows that the optimal solution of dynamical OT problem is also the critical point of the functional used in our saddle point optimization scheme (Equation 3.30). If the saddle point of $\mathcal{L}$ is unique, at such saddle point, value of $\mathcal{L}$ is exactly the optimal transport distance between ρ_a, ρ_b. The theorem is proved in the appendix section B.3.

Bidirectional dynamical formulation

To improve the stability and avoid local traps in the training processing, we propose a *bidirectional* scheme by exploiting the symmetric status of ρ_a and ρ_b as in the optimal transport distance Theorem 2.1.3.

Let's consider two OT problems

$$\min_F \max_{\Phi_F} \mathcal{L}^{ab}(F, \Phi_F), \quad \min_G \max_{\Phi_G} \mathcal{L}^{ba}(G, \Phi_G),$$

where $\mathcal{L}^{ab}$ is defined in (Equation 3.28), and $\mathcal{L}^{ba}$ is defined by switching ρ_a and ρ_b in (Equation 3.28). When reaching optima, the vector fields F and G are transport vectors in the opposite directions. At a specific point $x \in \mathbb{R}^d$, moving along straight line in the

direction F ends up at $x + F(x)$. The direction of G at $x + F(x)$ must point to the opposite direction of $F(x)$, which leads to $G(x + F(x)) = -F(x)$. Similarly, we also have $F(x + G(x)) = -G(x)$. Thus we introduce two regularization terms for F and G as

$$\mathcal{R}^{ab}(F, G) = \int |G(x + F(x)) + F(x)|^2 \, \rho_a(x) \, dx,$$
$$\mathcal{R}^{ba}(F, G) = \int |F(x + G(x)) + G(x)|^2 \rho_b(x) \, dx.$$

Our final saddle-point problem becomes

$$\min_{F,G} \max_{\Phi_F, \Phi_G} \mathcal{L}^{ab}(F, \Phi_F) + \mathcal{L}^{ba}(G, \Phi_G) + \lambda(\mathcal{R}^{ab}(F, G) + \mathcal{R}^{ba}(F, G)), \tag{3.31}$$

where λ is a tunable coefficient of our constraint terms.

Overview of the algorithm

We will mainly focus on solving the saddle point problem (Equation 3.31) in our research. we propose an algorithm that is summarized in the following steps. Please check its detailed discussions in appendix section B.3.

- **Preconditioning** We can apply preconditioning techniques to 2-Wasserstein cases in order to make our computation more efficient.
- **Main Algorithm** We set $F_{\theta_1}, G_{\theta_2}$ and $\Phi^F_{\omega_1}, \Phi^G_{\omega_2}$ as fully connected neural networks and optimize over their parameters ω_1, ω_2 and θ_1, θ_2 alternatively via stochastic gradient ascend and descend.
- **Stopping Criteria** When computed F (or G) is close to the optimal solution, the Wasserstein distance $W(\rho_a, \rho_b)$ (or $W(\rho_b, \rho_a)$) can be approximated by

$$\widehat{W}^{ab} = \int L(F(x)) \, \rho_a(x) \, dx, \; \widehat{W}^{ba} = \int L(G(x)) \, \rho_b(x) \, dx.$$

For a chosen threshold $\epsilon > 0$, we treat $|\widehat{W}^{ab} - \widehat{W}^{ba}| < \epsilon$ as the stopping criteria of our algorithm.

3.4.3 Experiments

Experiment Setup: We test our algorithm through a series of synthetic data sets and compare our numerical results with the computational methods introduced in the Python library (Python Optimal Transport (POT)) [92]. We also test our algorithm for realistic data sets including color transfer [93] and transportation between MNIST digits [94]. The detailed information on experimental set up is provided in appendix subsection B.3.4.

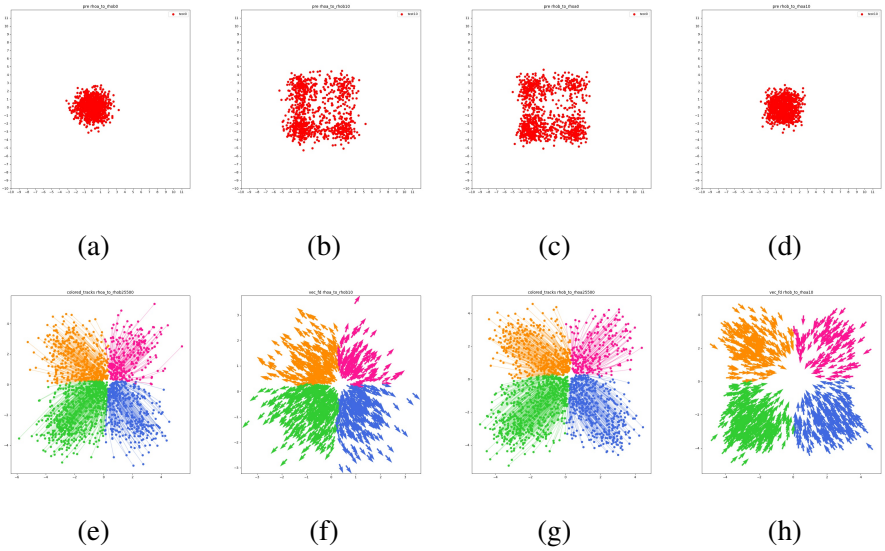
Notice that we are computing Wasserstein geodesic, namely, starting with an initial distribution ρ_{t_0} with $t_0 = 0.0$, in most cases we generate interpolating distributions for next ten time steps, from $t_1 = 0.1$ to $t_{10} = 1$. The cost functions are chosen as $L(v) = |v|^2$. We show the final state of the generated distribution for most of the examples due to space limitation. More experiments can be found in [23].

Synthetic: As a 10-dimensional case, here we set ρ_a as a standard Gaussian distribution and ρ_b as a special distribution where samples are unevenly distributed around four corners. We show the results of two dimensional projection in Figure 3.9.

For this synthetic data set, in the training process we set the batch size $N_t = 2000$ and sample size for demonstration $N_p = 1000$. For Figure 3.9, we can tell that in this 10 dimensional case, the generated samples closely follow the ground-truth distributions. Furthermore, we compare in Figure 3.10 the discrepancy between our numerical results with the results computed by POT package.

Realistic-1: In this case the view of the West Lake in summer and the view of the White Tower in autumn are given, then we aim to do a color transfer and simulate the summer view of White Tower and the autumn view of West Lake, the results are shown in Figure 3.11, the ground-truth and generated palette distributions are also included.

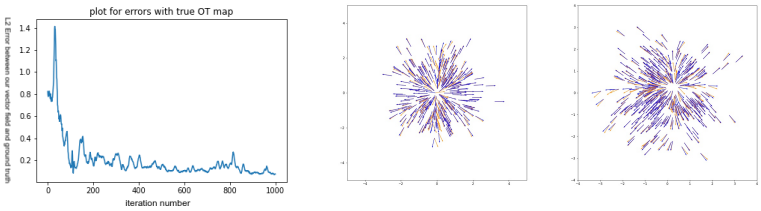
Realistic-2: We choose MNIST as our data set (28×28 dimensional) and study the



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3.9: Syn-3. (a)(b) true ρ_a and generated ρ_b, (c)(d) true ρ_b and generated ρ_a, (e)(g) tracks of sample points from $\rho_a(\rho_b)$ to $\rho_b(\rho_a)$, (f)(h) vector fields from $\rho_a(\rho_b)$ to $\rho_b(\rho_a)$.

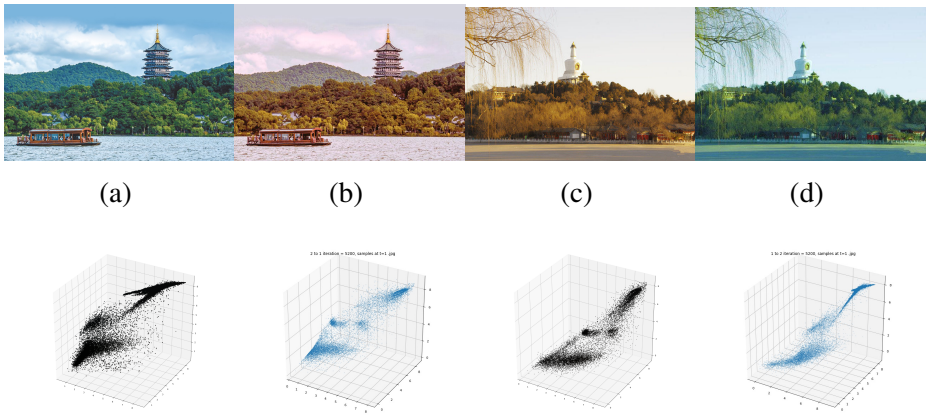


(a) (b) (c)

Figure 3.10: Left: Syn2: $L^2(\rho_a)$ error between our computed F and the real OT map vs iteration number; Middle: Syn1: Plot of our computed F (blue) and the OT vector field computed by POT (orange); Right: Syn3: Plot of our computed F and the OT vector field computed by POT (orange).

Wasserstein mappings as well as geodesic between digit $0(\rho_a)$ and digit $1(\rho_b)$, digit $4(\rho_a)$ and digit $8(\rho_b)$, digit $6(\rho_a)$ and digit $9(\rho_b)$. We present part of the results in Figure 3.12.

For realistic-1, we set the batch size $N_t = 1000$, for realistic-2 in each iteration we take $N_t = 500$ pictures for training. For realistic-2 data, we add small noise to the samples during the training process in order to make our training process more robust.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3.11: Real-2, Color transfer. (a)(b) true summer(generated autumn) view of the West Lake, (c)(d) true autumn(generated summer) view of the White Tower, (e)(f) palette distribution of the true summer West Lake(generated summer White Tower), (g)(h) palette distribution of the true autumn White Tower(generated autumn West Lake).

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3.12: Real-3, Digits transformation. (a)(b) true(generated) digit 4(8), (c)(d) true(generated) digit 8(4), (e)(f) true(generated) digit 6(9), (g)(h)true(generated) digit 9(6).

3.4.4 Conclusion

OT problem has been drawing more attention in machine learning recently. Though many algorithms have been proposed during the past several years for efficient computations, most of them do not consider the Wasserstein geodesics, neither be suitable for estimating optimal transport map with general cost in high dimensions. In this work we present a novel method to compute Wasserstein geodesic between two given distributions with general convex cost. Our method not only computes for the sample based Wasserstein geodesics, but also provides Wasserstein distance and optimal map. We demonstrate the effectiveness of our scheme through a series of experiments on both synthetic and realistic data sets. We are also working on generalizing our numerical scheme to the case of *general Lagrangian*

cost $L(x, v)$. Such generalized method will find its broad applications in optimal control and robotics research, where one needs to steer the distribution of mobile agents to the target distribution by optimizing the running cost.

CHAPTER 4

COMPUTATION OF HIGH DIMENSIONAL FOKKER-PLANCK EQUATIONS VIA PARAMETRIC PUSHFORWARD MAPS

4.1 Introduction

In this chapter, we will mainly focus on the method proposed by us [24],[25] for computing high dimensional Fokker-Planck equations.

The Fokker-Planck equation is a parabolic partial differential equation (PDE) that plays a crucial role in stochastic calculus, statistical physics, biology and many other disciplines [95, 96, 37]. Recently, it has seen many applications in machine learning as well [97, 98, 99]. The Fokker-Planck equation describes the evolution of probability density of a stochastic differential equation (SDE). In this research, we mainly focus on the following linear Fokker-Planck equation

$$\frac{\partial \rho(t,x)}{\partial t} = \nabla \cdot (\rho(t,x)\nabla V(x)) + D\Delta\rho(t,x), \quad \rho(0,x) = p(x), \tag{4.1}$$

where $x \in \mathbb{R}^d$, $V\colon \mathbb{R}^d \to \mathbb{R}$ is a given potential function, $D > 0$ is a diffusion coefficient, and $p(x)$ is the initial (or reference) density function. In numerical algorithms, there exist several classical methods [100] such as finite difference [101] or finite element [102] for solving the Fokker Planck equation. Most of the existing methods are grid based, which may be able to approximate the solution accurately if the grid sizes become small. However, they find limited usage in high dimensional problems, especially for $d > 3$, because the number of unknowns grows exponentially fast as the dimension increases. This is known as the curse of dimensionality. The main goal of this research is providing an alternative strategy, with provable error estimates, to solve high dimensional Fokker-Planck equations.

4.1.1 Neural parametric Fokker-Planck equation

To overcome the challenges imposed by high dimensionality, we leverage the generative models in machine learning [103] and a new interpretation of the Fokker-Planck equation in the theory of optimal transport [7]. We first introduce the KL divergence, also known as relative entropy, defined by

$$\mathcal{D}_{\mathrm{KL}}(\rho||\rho_*) = \int_{\mathbb{R}^d} \rho(x) \log\left(\frac{\rho(x)}{\rho_*(x)}\right) dx \quad \rho_*(x) = \frac{1}{Z_D} e^{-\frac{V(x)}{D}}, \text{ with } Z_D = \int_{\mathbb{R}^d} e^{-\frac{V(x)}{D}}\, dx.$$

Here $\rho_*(x)$ is the Gibbs distribution. A well-known fact is that the Fokker-Planck equation (Equation 4.1) can be viewed as the gradient flow of the functional $D\ \mathcal{D}_{\mathrm{KL}}(\rho||\rho_*)$ on the probability space $\mathcal{P}$ equipped with Wasserstein metric g^W [5, 4]. Recently, this line of research has been extended to parameter space in the field of information geometry [104, 105, 106], leading to an emergent area called transport information geometry [107, 108, 109, 110].

Inspired by aforementioned work, we study the Fokker-Planck equation defined on parameter manifold (space) $\Theta \subset \mathbb{R}^m$ equipped with metric tensor G which is obtained by pulling back the Wasserstein metric g^W to Θ. Here the metric tensor G can be viewed as an $m \times m$ matrix that contains all the metric information on Θ. In this research, we focus on the parameter space from generative models using neural networks. Our line of thoughts can be summarized as following. We start with a given reference distribution p, and consider a suitable family of parametric maps $\{T_\theta\}_{\theta\in\Theta}$. Such $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is also called parametric pushforward map since it generates a family of parametric distributions $\{T_{\theta\sharp}p\}$ by pushing forward p using T_θ (see Definition (Equation 2.1)). Then we consider the map $T_{(\cdot)\sharp} : \Theta \to \mathcal{P}, \theta \mapsto T_{\theta\sharp}p$, which can be treated as an immersion from parameter manifold Θ to probability manifold $\mathcal{P}$. We derive the metric tensor $G(\theta)$ by pulling back the Wasserstein metric via $T_{(\cdot)\sharp}$. Once establishing (Θ, G), we can compute the G-gradient flow of function $H(\theta) = D\ \mathcal{D}_{\mathrm{KL}}(T_{\theta\sharp}p \,||\, \rho_*)$ defined on the parameter manifold. This leads

to an ODE system that can be viewed as a parametric version of Fokker-Planck equation:

$$\dot{\theta}_t = -G(\theta_t)^{-1}\nabla_\theta H(\theta_t). \tag{4.2}$$

Here (and for the rest of this chapter) dot symbol $\dot{\theta}$ stands for time derivative $\frac{d\theta_t}{dt}$. Using the pushforward $\rho_\theta = T_{\theta\sharp}p$, in which θ is the solution of (Equation 4.2), we can approximate the solution ρ_t in (Equation 4.1).

There are many potential applications for the parameteric Fokker Planck equation. For example, the solution of (Equation 4.2) can be immediately used for sampling, which is a crucial task in statistics and machine learning. To be more precise, if the goal is drawing a large number of samples from ρ_t at N different time instances $\{t_1, t_2, ..., t_N\}$ along the solution of (Equation 4.1), we can acquire N sets of parameters $\theta_{t_1}, ..., \theta_{t_N}$ from the solution of (Equation 4.2), which provide N pushforward maps $T_{\theta_{t_1}}, ..., T_{\theta_{t_N}}$. Thus the desired samples at time t_k are $\{T_{\theta_{t_k}}(\boldsymbol{Z}_1), ..., T_{\theta_{t_k}}(\boldsymbol{Z}_M)\}$, in which $\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_M\}$ are samples drawn from the reference distribution p. If needed, the pushforward maps can be conveniently reused to generate more samples with negligible additional cost.

4.1.2 Computational method

For the computation of (Equation 4.2), we want to point out that metric tensor $G(\theta)$ doesn't have an explicit form and thus the direct computation of $G(\theta)^{-1}\nabla_\theta H(\theta)$ is not tractable. To deal with this issue, we design a numerical algorithm based on the semi-implicit Euler scheme of (Equation 4.2) with time step size h. To be more precise, at each time step, the algorithm seeks to solve the following double-minimization problem:

$$\begin{aligned}
&\min_\theta \left\{ \left(\int \left(2\,\nabla\phi(x) \cdot ((T_\theta - T_{\theta_k}) \circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2 \right) \rho_{\theta_k}(x)\, dx \right) + 2hH(\theta) \right\} \\
&\text{with } \phi \text{ solves: } \min_\phi \left\{ \int |\nabla\phi(x) - ((T_\theta - T_{\theta_k}) \circ T_{\theta_k}^{-1}(x))|^2 \rho_{\theta_k}(x)\, dx \right\}.
\end{aligned} \tag{4.3}$$

Here ρ_{θ_k} is the density of the pushforwarded distribution $T_{\theta_k \sharp} p$ (Equation 2.1). And $\phi \colon \mathbb{R}^d \to \mathbb{R}$ is the Kantorovich dual potential variable for constrained probability models in optimal transport theory. Hence (Equation 4.3) is derived following the semi-implicit Euler scheme in the dual variable. The advantage of using this formulation is that it allows us to design an efficient implementation, purely based on sampling techniques which are computational friendly in high dimensional problems, to compute the solution of the parameteric Fokker-Planck equation (Equation 4.2). In our implementation, we endow the pushforward map T_θ with certain kinds of deep neural network known as Normalizing Flow [103], because it is friendly to our scheme evaluations. The dual variable ϕ in the inner maximization is parametrized by the deep Rectified Linear Unit (ReLU) networks [111]. Once the network structures for T_θ and ϕ are chosen, the optimizations are carried out by stochastic gradient descent method [112], in which all terms involved can be computed using samples from the reference distribution p. We stress that this is critical in scaling up the computation in high dimensions. It is worth mentioning that we use neural network as a computational tool without any actual data. Such "data-poor" computation is in significant contrast to the mainstream of deep learning research.

4.1.3 Major innovations of the proposed method

There are two main innovative points regarding our proposed method:

- (Dimension reduction) Reducing the high dimensional evolution PDE to a finite dimensional ODE system on parameter space. Equivalently, we use the dynamics in a finite dimensional to approximate the density evolution of particles that follow the Vlasov-type SDE

  $$\dot{\boldsymbol{X}}_t = -\nabla V(\boldsymbol{X}_t) - D\nabla \log \rho_t(\boldsymbol{X}_t), \quad \rho_t \text{ is the density function of distribution of } \boldsymbol{X}_t.$$

  Here D is the diffusion coefficient as mentioned in (Equation 4.1). The density func-

tion ρ_t corresponds to the Fokker-Planck equation (Equation 4.1).

- (Sampling-friendly) We distill the information of ρ_t into parameters $\{\theta_t\}$ by solving the parametric Fokker-Planck equation (Equation 4.2). By doing so, we are able to obtain an efficient sampling technique to generate samples from ρ_t for any time step t. To be more precise, once we have applied our algorithm to solve (Equation 4.2) for the time-dependent parameters $\{\theta_t\}$, we can then generate samples from ρ_t by pushing forward the samples drawn from a reference distribution p using the pushforward map T_{θ_t} with very little computational cost. Such "implementing once for free future uses" mechanism is one of the significant advantages of our proposed algorithm. It is worth mentioning that in the view of both theoretical derivation and numerical implementation, our method is very different from Langevin Monte Carlo (LMC, MALA) methods [113, 114], which aims at targeting the stationary distribution of the SDE associated to (Equation 4.1); or moment methods [96] , which focuses on keeping track of certain statistical information of the density ρ_t.

4.1.4 Sketch of numerical analysis

In addition to the method proposed for solving (Equation 4.1), we also conducted a mathematical analysis on (Equation 4.2) and our algorithm. We established asymptotic convergence and error estimates for the parametric Fokker-Planck equation (Equation 4.2), which are summarized in the following two theorems:

Theorem 5.1 (Asymptotic convergence)**.** *Consider the Fokker-Planck equation* (Equation 4.1) *with potential V and diffusion coefficient D. Suppose V can be decomposed as $V = U + \phi$ with $U \in \mathcal{C}^2(\mathbb{R}^d)$, $\nabla^2 U \succeq KI$[1] with $K > 0$ and $\phi \in L^\infty(\mathbb{R}^d)$, and $\{\theta_t\}$ solves* (Equation 4.2). *Then the following inequality holds,*

$$\mathcal{D}_{KL}(\rho_{\theta_t}\|\rho_*) \leq \frac{\delta_0}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t},$$

[1]The matrix $\nabla^2 U(x) - KI_{d\times d}$ is non-negative definite for any $x \in \mathbb{R}^d$.

where ρ_* is the Gibbs distribution, $\tilde{\lambda}_D > 0$ is a constant related to the potential function V and D. δ_0 is a constant depending on the approximation power of pushforward map T_θ.

Theorem 5.11 (Approximation error)**.** *Consider the Fokker-Planck equation* (Equation 4.1) *with potential V, diffusion coefficient D and initial density ρ_0. Assume that λ is a lower bound of Hessian of potential V, i.e. $\nabla^2 V \succeq \lambda I$, δ_0 is defined in Theorem 5.1, $E_0 = W_2(\rho_{\theta_0}, \rho_0)$, and $\delta_0, E_0 \ll 1$, then the following uniform bounds for the L^2-Wasserstein error $W_2(\rho_{\theta_t}, \rho_t)$ hold:*

- *When $\lambda > 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq \max\{\sqrt{\delta_0}/\lambda, E_0\} \sim O(\sqrt{\delta_0} + E_0)$,*
- *When $\lambda = 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq \frac{\sqrt{\delta_0}}{\mu_D} \log \frac{B}{\sqrt{\delta_0}+E_0} + E_0 \sim O(\sqrt{\delta_0} \log \frac{1}{\sqrt{\delta_0}+E_0} + E_0)$,*
- *When $\lambda < 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq A\sqrt{\delta_0} + C\left(E_0 + \sqrt{\delta_0}/|\lambda|\right)^\alpha \sim O((E_0 + \sqrt{\delta_0})^\alpha)$.*

Here δ_0 is a constant depending on the approximation power of pushforward map T_θ. $\mu_D, A, B, C > 0$ are constants only depending on V, D, ρ_0, θ_0. And $\alpha = \frac{\mu_D}{|\lambda|+\mu_D}$ is a certain exponent between 0 and 1.

This result reveals that the difference between the solutions of the parametric Fokker-Planck equation (Equation 4.2) and the original equation (Equation 4.1), measured by their Wasserstein distance $W_2(\rho_{\theta_t}, \rho_t)$, has a *uniformly* small upper bound if both the initial error E_0 and δ_0 are small enough. Most of the techniques used in our analysis for establishing such a result rely on the theory of optimal transport and Wasserstein manifold, which are still not commonly used for numerical analysis in relevant literature. Besides error analysis for the continuous version of (Equation 4.2), we are able to provide the order of W_2-error for the numerical scheme when (Equation 4.2) is computed at discrete time by numerical schemes. To be more precise, if we apply forward-Euler scheme to (Equation 4.2) and compute $\{\theta_k\}$ at different time nodes $\{t_k\}$, we can show that error at t_k: $W_2(\rho_{\theta_k}, \rho_{t_k})$ is of order $O(\sqrt{\delta_0}) + O(Ch) + O(E_0)$ for finite time t. This is summarized in the following theorem:

Theorem 5.14 (Error for discrete scheme)**.** *Assume that* $\{\rho_t\}_{t\geq 0}$ *solves* (Equation 4.1) *with potential satisfying* $\lambda I \preceq \nabla^2 V \preceq \Lambda I$*,* $\{\theta_k\}_{k=0}^N$ *is the numerical solution of* (Equation 4.2) *at time nodes* $t_k = kh$ *for* $k = 0, 1, ..., N$ *computed by forward Euler scheme with time step* h*. Recall* δ_0 *as mentioned in Theorem 5.1 and we denote* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$*, then we have:*

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \leq (\sqrt{\delta_0}h + Ch^2)\frac{(1 - e^{-\lambda t_k})}{1 - e^{-\lambda h}} + e^{-\lambda t_k}E_0 \sim O(\sqrt{\delta_0}) + O(Ch) + O(E_0),$$

for all $0 \leq k \leq N$*. Here* C *is a constant depending on* N *and* h*.*

This indicates that the W_2-error is dominated by three different terms: $O(\sqrt{\delta_0})$ is the intrinsic error originated from the approximation mechanism of the parametric Fokker-Planck equation; $O(Ch)$ term is induced by the time discretization; and $O(E_0)$ term is the initial error. We further prove that the difference between the forward Euler scheme and our semi-implicit Euler scheme is of order $O(h^2)$, which implies that the proposed semi-implicit Euler scheme can achieve a similar error bounds as the one presented in Theorem 5.14.

It is worth mentioning that we establish Theorem 5.14 based on totally different techniques than those used for Theorem 5.11. Since the ODE (Equation 4.2) contains the term $G(\theta)^{-1}$, which is hard to handle by the traditional strategies, we interpret it as a particle system governed by a stochastic differential equations (SDEs) of Vlasov type, and obtain the analysis results shown in Theorem 5.14.

4.1.5 Literature review

Numerous works exist for solving the Fokker-Planck equations. A finite difference scheme is proposed in [101] so that it preserves the equilibrium of the original equation. A more general class of equations possessing Wasserstein gradient flow structures is solved in [115]. in which the method is based on a space discretization of a proximal-typed scheme (also known as JKO method [83]). Besides direct solutions, particle simulation techniques

also serve as an efficient way of solving the equation. The so-called "Blob" method is proposed in [85] and solves the equations by evolving a certain interacting particle systems. Related swarming system is also studied in [116, 117, 118, 119, 120]. In [121], the authors propose another type of interacting systems in order to approximate $\nabla \log \rho$, which plays the role of the diffusion term in the Fokker-Planck equation, with higher accuracy and less fluctuation. In [122, 123], the authors mainly focus on exploiting the gradient flow structure, i.e. a particle discretization of the Fokker–Planck equation, to deal with Bayesian inference problems.

In addition to the literature focusing on solving the Fokker-Planck equations, There are existing works on applying neural networks to solve PDE of various types in high dimensional spaces [124, 125, 126, 127, 128, 129]. Among the listed works, algorithms for general types of high dimensional PDEs are provided in [125, 126]; a sampling friendly method is proposed in [129] to deal with the general optimal control problem of diffusion processes. This is equivalent to solving an associated Hamilton-Jacobi-Bellman equation and such technique can also be applied to importance sampling and rare event simulation. Moreover, numerical methods for high dimensional parabolic PDEs, to which the Fokker-Planck equation belongs, are studied in [124] and [127]. Our approach differs from these existing works in many aspects, including motivations, strategies, and the associated numerical analysis.

For example, in [124], the authors propose to use the non-linear Feynmann-Kac formula to re-write certain parabolic PDEs as the Backward Stochastic Differential Equation (BSDE), which is then reformulated as a stochastic control problem (also known as reinforcement learning in machine learning community). By applying deep neural network as the control function and optimizing over network parameters, the solution at any given space-time location can be evaluated. Another example is [127], which mainly focuses on computing the committor function that solves a steady-state (time-independent) Fokker-Planck equation with specific boundary conditions. This committor function can be treated

as the solution to a variational problem associated with an energy functional. A neural network is used to replace the solution in the variational problem. When optimizing over network parameters, the neural network can be used to approximate the committor function.

In this research, we focus on designing a sampling-friendly method for the time dependent Fokker-Planck equation. There are two main reasons that motivate us for this investigation. One, as mentioned before, is to design sample based algorithm to solve PDEs in high dimensions. The other is to provide an alternative sampling strategy that can be potentially faster than LMC. Our approaches are different in terms of how deep networks are leveraged to approximate the solution of the PDE. We use pushforward of a given reference measure by neural networks to create a generative model. This is to approximate the stream of probability distributions, which can be used to generate samples not only at the terminal time, but also any time in between. More importantly, we prove results, obtained by using newly developed techniques based on Wasserstein metric on probability manifold, on the asymptotic convergence and error control of our numerical schemes. To the best of our knowledge, similar results are still lacking in existing studies.

4.1.6 Organization of this chapter

We organize this chapter as follows. In section 4.2, we briefly introduce some background knowledge of the Fokker-Planck equation, including its relation with SDE and its Wasserstein gradient flow structure. In section 4.3, we introduce the Wasserstein statistical manifold (Θ, G) and derive our parametric Fokker-Planck equation as the manifold gradient flow of relative entropy on Θ. We study the geometric property of this equation, including an insightful particle motion based interpretation of the parametric Fokker-Planck equation. In section 4.4, we discuss the straightforward method for 1-D Fokker-Planck equation. In section 4.5, we design a numerical scheme that is tractable for computing our parametric Fokker-Planck equation using deep learning framework. Some important details of imple-

mentation will be discussed. We present asymptotic convergence and error estimates for the parametric Fokker-Planck equation in section 4.6, and provide some numerical examples in section 4.7.

4.2 Background on the Fokker-Planck equation

In this section, we present two different perspectives regarding the Fokker-Planck equations, More discussion can be found in [130].

4.2.1 As the density evolution of stochastic differential equation

The general form of the Fokker-Planck equation is [131, 132]:

$$\begin{aligned}\frac{\partial \rho(x,t)}{\partial t} &= -\nabla \cdot (\rho(x,t)\boldsymbol{\mu}(x,t)) + \frac{1}{2}\nabla^2 : (\boldsymbol{D}(x,t)\rho(x,t)) \\ &= -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}(\rho(x,t)\mu_i(x,t)) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}(D_{ij}(x,t)\rho(x,t)), \quad \rho(x,0) = \rho_0(x).\end{aligned} \tag{4.4}$$

Here $\boldsymbol{\mu} = (\mu_1, ..., \mu_d)^{\mathrm{T}}$ is the drift function and $\boldsymbol{D} = \{D_{ij}\}$ is the $d \times d$ diffusion tensor. Furthermore, $\boldsymbol{D}$ can be written as $\boldsymbol{D} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$, where $\boldsymbol{\sigma}(x,t)$ is a $d \times \tilde{d}$ matrix. One derivation of the Fokker-Planck equation originates from the following stochastic differential equation (SDE) [131, 132],

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)\, dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t)\, d\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim \rho_0,$$

where $\{\boldsymbol{B}_t\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^{\tilde{d}}$, and ρ_0 is the distribution of the initial state. It is well known that the evolution of the density $\rho(x,t)$ of the stochastic process $\{\boldsymbol{X}_t\}_{t\geq 0}$ is described by the above the Fokker-Planck equation.

In this research, we consider a more specific type of (Equation 4.4) by setting $\boldsymbol{\mu}(x,t) = -\nabla V(x)$, $\boldsymbol{\sigma}(x,t) = \sqrt{2D}\, I_{d\times d}$ $(D > 0)$, where $I_{d\times d}$ is the d by d identity matrix, and so

$\boldsymbol{D} = 2D\ I_{d\times d}$. Then (Equation 4.2.1) is,

$$d\boldsymbol{X}_t = -\nabla V(\boldsymbol{X}_t)\ dt + \sqrt{2D}\ d\boldsymbol{B}_t \quad \boldsymbol{X}_0 \sim \rho_0. \tag{4.5}$$

This equation is also called over-damped Langevin dynamics which has broad applications in computational physics, computational biology, Bayesian statistics [113, 133, 134]. The corresponding Fokker-Planck equation is simplified to

$$\frac{\partial \rho(x,t)}{\partial t} = \nabla \cdot (\rho(x,t)\nabla V(x)) + D\Delta\rho(x,t), \quad \rho(x,0) = \rho_0(x). \tag{4.6}$$

In addition, we would like to mention that there is a Vlasov-type SDE corresponding to the Fokker-Planck equation (Equation 4.6):

$$\frac{d\boldsymbol{X}_t}{dt} = -\nabla V(\boldsymbol{X}_t) - D\ \nabla \log \rho(\boldsymbol{X}_t, t), \quad \boldsymbol{X}_0 \sim \rho_0, \tag{4.7}$$

in which $\rho(\cdot, t)$ is the density of $\boldsymbol{X}_t$. This Vlasov-type SDE (Equation 4.7) will be very useful in our proofs for the error estimates of our proposed numerical algorithms.

4.2.2 As the Wasserstein gradient flow of relative entropy

Another useful viewpoint states that (Equation 4.6) is the Wasserstein gradient flow of the following relative entropy functional (also known as Kullbeck-Leibler divergence).

$$\mathcal{H}(\rho) = D\ \mathcal{D}_{\text{KL}}\left(\rho \,\middle|\middle|\, \rho_*\right) = \left(\int V(x)\rho(x) + D\rho(x)\log\rho(x)\ dx\right) + D\ \log Z_D. \tag{4.8}$$

We provide a brief introduction to Wasserstein manifold as well as Wasserstein gradient flow in subsection 2.2.1 and subsection 2.2.2.

4.3 Parametric Fokker-Planck equation

In this section, we provide detailed derivation for our parametric Fokker-Planck equation.

4.3.1 Wasserstein statistical manifold

Consider a parameter space Θ as an open, convex set in $\mathbb{R}^m$, and assume the sample space is $\mathbb{R}^d$. Let T_θ be a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ parametrized by θ. In our discussion, we always assume the invertibility of $T_\theta(x)$, and it is second order differentiable with respect to x and θ, i.e. $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$.

Remark 7. *There are many different choices for T_θ:*

- *We can set $T_\theta(x) = Ux + b$, with $\theta = (U, b)$, U is a $d \times d$ invertible matrix, $b \in \mathbb{R}^d$;*
- *We may also choose T_θ as the linear combination of basis functions i.e., $T_\theta(x) = \sum_{k=1}^m \theta_k \vec{\Phi}_k(x)$, where $\{\vec{\Phi}_k\}_{k=1}^m$ are the basis functions and the parameter θ will be the coefficients: $\theta = (\theta_1, ..., \theta_m)$;*
- *We can also treat T_θ as neural network. Its general structure can be written as the composition of l affine and non-linear activation functions:*

$$T_\theta(x) = \sigma_l(W_l(\sigma_{l-1}(\dots \sigma_1(W_1 x + b_1)\dots)) + b_l).$$

In this case, the parameter θ will be the weight matrices and bias vectors of the neural network, i.e. $\theta = (W_1, b_1, ..., W_l, b_l)$.

Let $p \in \mathcal{P}$ be a reference probability measure with positive density defined on $\mathbb{R}^d$, such as the standard Gaussian. Recall the definition on the pushforward of measure in Equation 2.1, we denote ρ_θ as the density of $T_{\theta\sharp}p$. Such kind of mechanism of producing parametric probability distributions is also known as **generative model**, which has broad

applications in deep learning research [135, 136, 137]. We further assume our T_θ satisfy the following two conditions:

$$\text{Condition 1:} \quad \int |z|^2 \rho_\theta(z)\, dz = \int |T_\theta(x)|^2\, dp(x) < \infty \quad \forall\, \theta \in \Theta. \tag{4.9}$$

This ensures that $\rho_\theta \in \mathcal{P}$ for each $\theta \in \Theta$. In order to introduce Wasserstein metric to the parameter space Θ, we also assume that the Frobenius norm of the operator $\partial_\theta T_\theta(x) : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is locally bounded in the following sense: for any fixed $\theta_* \in \Theta$, there exists $r(\theta_*) > 0$ and two functions $L_1(\cdot|\ \theta_*), L_2(\cdot|\ \theta_*)$ satisfying

Condition 2: (4.10)

$\|\partial_\theta T_\theta(x)\|_F \le L_1(x|\ \theta_*), \|\partial_\theta T_\theta(x)\|_F^2 \le L_2(x|\ \theta_*),\ \forall\, \theta, |\theta - \theta_*| < r(\theta_*)$ and $x \in \mathbb{R}^d$, and $\int L_1(x|\ \theta_*)\, dp(x) < \infty, \quad \int L_2(x|\ \theta_*)\, dp(x) < \infty.$

We define the parametric submanifold $\mathcal{P}_\Theta \subset \mathcal{P}$ as:

$$\mathcal{P}_\Theta = \{\rho_\theta \text{ is density function of } \ T_{\theta\sharp}p \mid \theta \in \Theta\}.$$

Clearly, the connection between $\mathcal{P}$ and Θ is through the pushforward operation $T_{\theta\sharp} : \Theta \to \mathcal{P}_\Theta, \theta \mapsto \rho_\theta$. Hence it is natural to define the Wasserstein metric $G(\theta)$ on parameter space Θ as the pullback of g^W by $T_{\theta\sharp}$. To be specific, we define $G(\theta) = (T_{\theta\sharp})^* g^W$. Using this definition, $T_{\theta\sharp}$ becomes an isometric immersion from Θ to $\mathcal{P}$. For each θ, $G(\theta)$ is a bilinear form defined on $\mathcal{T}_\theta\Theta \simeq \mathbb{R}^m$, which can be identified as an $m \times m$ matrix.

Before computing $G(\theta)$, we introduce a lemma which can help us to better understand $G(\theta)$.

Lemma 4.3.1. *Suppose $\vec{u}, \vec{v}$ are two vector fields defined on $\mathbb{R}^d$, suppose φ, ψ solves $-\nabla\cdot(\rho\nabla\varphi) = -\nabla\cdot(\rho\vec{u})$ and $-\nabla\cdot(\rho\nabla\psi) = -\nabla\cdot(\rho\vec{v})$, or equivalently, $\text{Proj}_\rho[\vec{u}] = \nabla\varphi$ and*

$\text{Proj}_\rho[\vec{v}] = \nabla\psi$ (check Definition 4.5.1). Then:

$$\int \vec{u}(x) \cdot \nabla\psi(x)\rho(x)\, dx = \int \nabla\varphi(x) \cdot \nabla\psi(x)\rho(x)\, dx; \tag{4.11}$$

$$\int |\nabla\psi(x)|^2\rho(x)\, dx \leq \int |\vec{v}(x)|^2\rho(x)\, dx. \tag{4.12}$$

We prove Lemma 4.3.1 in Appendix section C.1. The metric tensor $G(\theta)$ is computed in the following theorem.

Theorem 4.3.2. *Assume Θ satisfies* (Equation 4.9),(Equation 4.10). *T_θ is invertible and $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$. Then Θ can be equipped with the metric tensor $G = (T_{\theta\sharp})^* g^W$, which is an $m \times m$ non-negative definite symmetric matrix of the form:*

$$G(\theta) = \int \nabla\mathbf{\Psi}(T_\theta(x))\nabla\mathbf{\Psi}(T_\theta(x))^T\, dp(x) \tag{4.13}$$

at every $\theta \in \Theta$. More precisely, in entry-wise form,

$$G_{ij}(\theta) = \int \nabla\psi_i(T_\theta(x)) \cdot \nabla\psi_j(T_\theta(x))\, dp(x), \;\; 1 \leq i, j \leq m,$$

in which $\mathbf{\Psi} = (\psi_1, \cdots, \psi_m)^T$ and $\nabla\mathbf{\Psi}$ is an $m \times d$ Jacobian matrix of $\mathbf{\Psi}$. For each $j = 1, 2, \cdots, m$, ψ_j solves the following equation:

$$\nabla \cdot (\rho_\theta \nabla\psi_j(x)) = \nabla \cdot (\rho_\theta \frac{\partial T_\theta}{\partial \theta_j}(T_\theta^{-1}(x))). \tag{4.14}$$

with boundary conditions

$$\lim_{x\to\infty} \rho_\theta(x)\nabla\psi_j(x) = 0.$$

Proof. Suppose $\xi \in \mathcal{T}\Theta$ is a vector field on Θ, for a fixed $\theta \in \Theta$, we first compute the pushforward $(T_{\theta\sharp})_*\xi(\theta)$ of ξ at point θ: We choose any smooth curve $\{\theta_t\}_{t\geq 0}$ on Θ with $\theta_0 = \theta$ and $\dot{\theta}_0 = \xi(\theta)$. If we denote $\rho_{\theta_t} = T_{\theta_t\sharp}p$, we have $(T_{\theta\sharp})_*\xi(\theta) = \frac{\partial\rho_{\theta_t}}{\partial t}\Big|_{t=0}$.

To compute $\frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0}$, we consider an arbitrary $\phi \in C_0^\infty(M)$.

On one hand, $\frac{\rho_{\theta_{\Delta t}}(y)-\rho_{\theta_0}(y)}{\Delta t} = \frac{\partial}{\partial t}\rho(\theta_{\tilde{t}_1}, y)$, where $\tilde{t}_1$ is some point between $0, \Delta t$, since $\phi \in C_0^\infty$ and $\rho(\theta_t, x)$ is at least C^1 with respect to t, y, we can show that the function $\varphi(x) = \sup_{s\in[0,\Delta t]} |\phi(x)\frac{\partial}{\partial t}\rho(\theta_s, y)|$ is continuous on a compact set and thus integrable on $\mathbb{R}^d$. Using dominated convergence theorem, we have:

$$\frac{\partial}{\partial t}\left(\int \phi(y)\rho_{\theta_t}(y)\, dy\right)\Big|_{t=0} = \int \phi(y)\frac{\partial \rho_{\theta_t}(y)}{\partial t}\Big|_{t=0} dy. \tag{4.15}$$

On the other hand, we have:

$$\frac{\phi(T_{\theta_{\Delta t}}(y)) - \phi(T_{\theta_0}(y))}{\Delta t} = \dot{\theta}_{\tilde{t}_2}^{\mathrm{T}}\, \partial_\theta T_{\theta_{\tilde{t}_2}}(x)^{\mathrm{T}}\, \nabla\phi(T_{\theta_{\tilde{t}_2}}(y)), \tag{4.16}$$

in which $\tilde{t}_2$ is also between $0, \Delta t$. For any Δt small enough and $\tilde{t} \in [0, \Delta t]$, we can easily find upper bounds for $\|\dot{\theta}_{\tilde{t}}\| \le A$ and $\|\nabla\phi(\cdot)\|_\infty \le B$. Recall the condition (Equation 4.10), when Δt is small enough, we have $|\theta_{\Delta t} - \theta_0| < r(\theta_0)$, thus we obtain the following upper bound for (Equation 4.16)

$$|\dot{\theta}_{\tilde{t}}^{\mathrm{T}}\, \partial_\theta T_{\theta_{\tilde{t}}}(x)^{\mathrm{T}}\, \nabla\phi(T_{\theta_{\tilde{t}}}(y))| \le AB\|\partial_\theta T_{\theta_{\tilde{t}}}(x)\|_F \le ABL_1(x|\theta_0).$$

By (Equation 4.10), we know $L_1(\cdot|\theta_0) \in L^1(p)$, we can apply dominated convergence theorem to obtain:

$$\frac{\partial}{\partial t}\left(\int \phi(T_{\theta_t}(x))dp\right)\Big|_{t=0} = \int \dot{\theta_t}^{\mathrm{T}}\partial_\theta T_{\theta_t}(x)^{\mathrm{T}}\nabla\phi(T_{\theta_t}(x))|_{t=0}dp. \tag{4.17}$$

Since $\frac{\partial}{\partial t}\int \phi(y)\rho_{\theta_t}(y)\, dy = \frac{\partial}{\partial t}\int \phi(T_{\theta_t}(x))\, dp(x)$, we use (Equation 4.15), (Equation 4.17)

to get:

$$\begin{aligned}\int \phi(y)\frac{\partial \rho_{\theta_t}}{\partial t}(y)\Big|_{t=0}\, dy &= \int \dot{\theta_t}^{\mathrm{T}} \partial_\theta T_{\theta_t}(x)^{\mathrm{T}} \nabla\phi(T_{\theta_t}(x))|_{t=0}\, dp(x) \\ &= \int \dot{\theta}_t^{\mathrm{T}} \left(\frac{\partial T_{\theta_t}}{\partial \theta}(T_{\theta_t}^{-1}(x))\right)^{\mathrm{T}} \nabla\phi(x)\ \rho_{\theta_t}(x)|_{t=0}\, dx \\ &= \int \phi(x) \left(-\nabla\cdot\left(\rho_{\theta_t}(x)\frac{\partial T_{\theta_t}}{\partial \theta}(T_{\theta_t}^{-1}(x))\ \dot{\theta}_t\right)\right)|_{t=0}\, dx.\end{aligned}$$

Because $\phi(x)$ is arbitrary, this weak formulation reveals that

$$(T_{\theta\sharp})_*\xi(\theta) = \frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla\cdot\left(\rho_\theta(x)\ \frac{\partial T_\theta}{\partial \theta}(T_\theta^{-1}(x))\xi(\theta)\right). \tag{4.18}$$

Now let us compute the metric tensor G. Since $T_{\theta\sharp}$ is isometric immersion from Θ to $\mathcal{P}$, the pullback of g^W by $T_{\theta\sharp}$ gives G, i.e. $(T_{\theta\sharp})^* g^W = G(\theta)$. By definition of pullback map, for any $\theta \in \Theta$ and $\xi(\theta) \in \mathcal{T}_\theta\Theta$, we have:

$$G(\theta)(\xi(\theta),\xi(\theta)) = g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta),(T_{\theta\sharp})_*\xi(\theta)). \tag{4.19}$$

To compute the right hand side of (Equation 4.19), recall (Equation 2.58), we need to solve for φ from:

$$\frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla\cdot(\rho_\theta(x)\nabla\varphi(x)). \tag{4.20}$$

By (Equation 4.18), (Equation 4.20) is:

$$\nabla\cdot(\rho_\theta(x)\nabla\varphi(x)) = \nabla\cdot\left(\rho_\theta(x)\frac{\partial_\theta T_\theta}{\partial\theta}(T_\theta^{-1}(\cdot))\xi(\theta)\right). \tag{4.21}$$

We can straightforwardly check that $\varphi(x) = \mathbf{\Psi}^{\mathrm{T}}(x)\xi(\theta)$ is the solution of (Equation 4.21). Now by definition of g^W as mentioned in subsection 2.2.1, we write the right hand side of

(Equation 4.19) as

$$g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta),(T_{\theta\sharp})_*\xi(\theta)) = \int |\nabla\varphi(y)|^2\rho_\theta(y)\,dy \tag{4.22}$$
$$=\xi(\theta)^{\mathrm{T}}\left(\int \nabla\mathbf{\Psi}(y)\nabla\mathbf{\Psi}(y)^{\mathrm{T}}\rho_\theta(y)\,dy\right)\xi(\theta)$$
$$=\sum_{i,j=1}^{m}\left(\int \nabla\psi_i(y)\cdot\nabla\psi_j(y)\rho_\theta(y)\,dy\right)\xi_i(\theta)\xi_j(\theta).$$

Here we assume components of $\xi(\theta)$ as $(\xi_1(\theta),...,\xi_m(\theta))^{\mathrm{T}}$. Before we compute $G(\theta)$, we first verify that the inner product in (Equation 4.22) is finite for any $\xi \in \mathcal{T}\Theta$. To show this, by Cauchy–Schwarz inequality we obtain

$$\int \nabla\psi_i(y)\cdot\nabla\psi_j(y)\rho_\theta(y)\,dy \le \left(\int |\nabla\psi_i(y)|^2\rho_\theta(y)\,dy\right)^{\frac{1}{2}}\left(\int |\nabla\psi_j(y)|^2\rho_\theta(y)\,dy\right)^{\frac{1}{2}}.$$

recall ψ_j defined in (Equation 4.14), then applying (Equation 4.12) of Lemma (Lemma 4.3.1) yields

$$\int |\nabla\psi_j(y)|^2\rho_\theta(y)\,dy \le \int \left|\frac{\partial T_\theta}{\partial\theta_j}(T_\theta^{-1}(y))\right|^2\rho_\theta(y)\,dy \le \int L_2(y|\theta)p(y)\,dy < \infty.$$

The last two inequalities are due to condition (Equation 4.10). As a result, we proved the finiteness of (Equation 4.22).

Finally, let us compute:

$$G(\theta)(\xi(\theta),\xi(\theta)) = g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta),(T_{\theta\sharp})_*\xi(\theta))$$
$$=\xi(\theta)^{\mathrm{T}}\left(\int \nabla\mathbf{\Psi}(T_\theta(x))\nabla\mathbf{\Psi}(T_\theta(x))^{\mathrm{T}}dp(x)\right)\xi(\theta).$$

Thus we can verify that

$$G(\theta) = \int \nabla\mathbf{\Psi}(T_\theta(x))\nabla\mathbf{\Psi}(T_\theta(x))^{\mathrm{T}}\,dp(x),$$

which completes the proof. □

Generally speaking, the metric tensor G does not have an explicit form when $d \geq 2$. It is worth mention that G has an explicit form and can be computed directly when $d = 1$.

Corollary 4.3.2.1. *When dimension d of M equals 1. And we further assume that: $\rho_\theta > 0$ on M and $\lim_{x\to\pm\infty} \rho_\theta(x) = 0$. Then $G(\theta)$ has an explicit form:*

$$G(\theta) = \int \partial_\theta T_\theta(x)^T \partial_\theta T_\theta(x)\, dp(x). \tag{4.23}$$

Remark 8 (Well-posedness of (Equation 4.14))**.** *It is worth commenting on the existence and regularity of* (Equation 4.14)*. Determining what properties or conditions that T_θ should have to guarantee the well-posedness of* (Equation 4.14) *is an interesting and important problem on its own. In references such as [138] and [139], there are sufficient conditions that guarantee the well-posedness of elliptic PDEs defined on $\mathbb{R}^d$. Most of the existing results require uniform lower bound on ρ_θ, i.e. $\rho_\theta(x) > \epsilon > 0$ for all $x \in \mathbb{R}^d$. Such coercive condition is not applicable in our case since $\int \rho_\theta(x)dx = 1$ is finite. On the other hand, section 8.1.2 of [29] provides another sufficient condition on the well-posedness of* (Equation 4.14)*: If there exists $\lambda > 0$ such that the following Poincaré inequality* (Equation 4.24) *holds for any $\varphi \in C^\infty(\mathbb{R}^d)$ with compact support,*

$$\int |\nabla\varphi(x)|^2 \rho_\theta(x)\, dx \geq \lambda \int \left(\varphi(x) - \int \varphi\rho_\theta\, dx\right)^2 \rho_\theta(x)\, dx, \tag{4.24}$$

and $-\nabla\cdot(\rho_\theta \frac{\partial T_\theta}{\partial \theta_j}(T_\theta^{-1}(\cdot))) \in L^2(\rho_\theta)$, Then (Equation 4.14) *admits a unique solution ψ_j with $\nabla\psi_j \in L^2(\rho_\theta)$. To the best of our knowledge, it is still unclear that what kind of properties of T_θ may lead to* (Equation 4.24)*.*

It is worth pointing out that under certain situations discussed in subsection 4.3.4, (Equation 4.14) *does have classical solutions. For example, if we select T_θ as an affine transform and consider the Fokker-Planck equation* (Equation 4.6) *with quadratic poten-*

tial V and Gaussian initial ρ_0, we can prove that (Equation 4.14) *is well-posed along the trajectory of the ODE* (Equation 4.28)*, i.e. the following elliptic equation*

$$-\nabla \cdot (\rho_{\theta_t} \nabla \psi) = -\nabla \cdot (\rho_{\theta_t} \frac{\partial_\theta T_{\theta_t}}{\partial \theta}(T_{\theta_t}^{-1}(x))\dot{\theta}_t), \quad \textit{where } \{\theta_t\} \textit{ solves } \text{(Equation 4.28)},$$

always admits a classical solution $\psi(x) = V(x) + D \log \rho_\theta(x) + $ Const.

In general, The conditions imposed on T_θ to guarantee well-posedness of (Equation 4.14) *is a fundamental and interesting topic subject to further investigation. A good reference related to the topic can be found in [35].*

Following theorem provides several criteria for examining whether G is a Riemannian metric, i.e. whether $G(\theta)$ is positive definite.

Theorem 4.3.3. *For $\theta \in \Theta$, $\{\psi_k\}_{k=1}^m$ satisfies* (Equation 4.14)*, the following four statements are equivalent*

1. *$G(\theta)$ is positive definite;*
2. *For any $\xi \in \mathcal{T}_\theta\Theta$ ($\xi \neq 0$), there exists $z \in M$ such that $\nabla \cdot (\rho_\theta(z) \frac{\partial T_\theta}{\partial \theta}(T_\theta^{-1}(z))\xi) \neq 0$;*
3. *$\{\nabla \psi_k\}_{k=1}^m$, as m functions in the space $L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_{\theta_k})$, are linearly independent;*
4. *$\frac{d}{dt}(T_{\theta + t\xi \sharp} p)|_{t=0} \neq 0$ for any $\xi \in \mathbb{R}^m$.*

Proof. We first verify that 1 and 2 are equivalent. We need the following identity used in Theorem 4.3.2: For any θ, ξ, x, we have

$$\nabla \cdot (\rho_\theta(x) \nabla(\xi^{\mathrm{T}} \mathbf{\Psi}(x))) = \nabla \cdot (\rho_\theta(x) \frac{\partial T_\theta}{\partial \theta}(T_\theta^{-1}(x))\xi). \tag{4.25}$$

($\Leftarrow$): suppose for any $\theta \in \Theta$ and $\xi \in \mathcal{T}_\theta\Theta$, at certain $z \in \mathbb{R}^d$, $\nabla \cdot (\rho_\theta(z) \frac{\partial T_\theta}{\partial \theta}(T_\theta^{-1}(z)\xi) \neq 0$, then $\nabla \cdot (\rho_\theta(z) \nabla(\xi^{\mathrm{T}} \mathbf{\Psi}(z))) \neq 0$, thus $\rho_\theta \nabla(\xi^{\mathrm{T}} \mathbf{\Psi})$ is not identically $\mathbf{0}$. Using continuity of $\rho_\theta \nabla(\xi^{\mathrm{T}} \mathbf{\Psi})$, we know that: $|\nabla(\xi^{\mathrm{T}} \mathbf{\Psi}(x))|^2 \rho_\theta(x) > 0$ in some small neighbourhood of z.

Thus we have:

$$\xi^{\mathrm{T}} G(\theta)\xi = \int |\nabla \mathbf{\Psi}(x)^{\mathrm{T}}\xi|^2 \rho_\theta(x)\, dx > 0, \tag{4.26}$$

holds for any θ and ξ, this leads to the positive definiteness of G.

($\Rightarrow$): Now suppose (Equation 4.26) holds for all θ, ξ, then we have

$$\int -\nabla \cdot (\rho_\theta(x)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(x))) \cdot \xi^{\mathrm{T}}\mathbf{\Psi}(x)\, dx > 0.$$

This leads to the existence of a $z \in \mathbb{R}^d$ such that $-\nabla \cdot (\rho_\theta(z)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(z))) \neq 0$. Combining (Equation 4.25), we have verified the equivalence between 1 and 2.

We recall (Equation 4.18), then $\frac{d}{dt}(T_{\theta+t\xi\sharp}p)|_{t=0} = (T_{\theta\sharp})_*\xi = -\nabla \cdot (\rho_\theta(x)\frac{\partial T_\theta}{\partial \theta}(T_\theta^{-1}x))\xi)$, this verifies the equivalence between 2 and 3.

Finally, as stated before, we can verify $\xi^{\mathrm{T}} G(\theta)\xi = \| \sum_{k=1} \xi_k \nabla \psi_k \|^2_{L^2(\rho_\theta)}$, this formula will directly leads to the equivalence between 1 and 4 and we have proved the equivalence among statements 1,2,3 and 4.

□

To keep our discussion concise in the following sections, we will always assume $G(\theta)$ is positive definite for every $\theta \in \Theta$.

4.3.2 Parametric Fokker-Planck equation

We consider the relative entropy functional on Θ as $H = \mathcal{H} \circ T_{(\cdot)\sharp} : \Theta \to \mathbb{R}$,

$$\begin{aligned} H(\theta) = \mathcal{H}(\rho_\theta) &= \left(\int V(x)\rho_\theta(x) + D\rho_\theta(x) \log \rho_\theta(x)\, dx \right) + D \log Z_D \\ &= \left(\int V(T_\theta(x)) + D \log \rho_\theta(T_\theta(x))\, dp(x) \right) + D \log Z_D. \end{aligned} \tag{4.27}$$

Following the theory in [104], the gradient flow of H on Wasserstein parameter manifold (Θ, G) satisfies

$$\dot{\theta} = -G(\theta)^{-1}\nabla_\theta H(\theta). \tag{4.28}$$

We call (Equation 4.28) *parametric Fokker-Planck equation*. The ODE (Equation 4.28) as the Wasserstein gradient flow on parameter space (Θ, G) is closely related to the Fokker-Planck equation on probability submanifold $\mathcal{P}_\Theta$. We have the following theorem, which is a natural result derived from submanifold geometry.

Theorem 4.3.4. *Suppose $\{\theta_t\}_{t\geq 0}$ solves (Equation 4.28). Then $\{\rho_{\theta_t}\}$ is the gradient flow of $\mathcal{H}$ on probability submanifold $\mathcal{P}_\Theta$. Furthermore, at any time t, $\dot{\rho}_{\theta_t} = \frac{d}{dt}\rho_{\theta_t} \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ is the orthogonal projection of $-\text{grad}_W \mathcal{H}(\rho_{\theta_t}) \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}$ onto the subspace $\mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ with respect to the Wasserstein metric g^W.*

We prove this theorem in the section C.2.

The following theorem is an important new statement closely related to Theorem 4.3.4.

Theorem 4.3.5 (Wasserstein gradient as solution to a least squares problem)**.** *For a fixed $\theta \in \Theta$, $\mathbf{\Psi} \subset \mathbb{R}^m$ as defined in Theorem 4.3.2, then*

$$G(\theta)^{-1}\nabla_\theta H(\theta) = \underset{\eta \in \mathcal{T}_\theta\Theta \cong \mathbb{R}^m}{\arg\min} \left\{ \int |(\nabla \mathbf{\Psi}(T_\theta(x)))^T \eta - \nabla \left(V + D\log\rho_\theta\right) \circ T_\theta(x)|^2 dp(x) \right\}. \tag{4.29}$$

Proof. Direct computation shows that minimizing the function in (Equation 4.29) is equivalent to minimizing:

$$\eta^{\mathrm{T}} \left(\int \nabla \mathbf{\Psi}(T_\theta(x)) \nabla \mathbf{\Psi}(T_\theta(x))^{\mathrm{T}} \, dp \right) \eta - 2\, \eta^{\mathrm{T}} \left(\int \nabla \mathbf{\Psi} \nabla (V + D\log\rho_\theta)\rho_\theta \, dx \right).$$

For each entry in the second term, we have:

$$
\begin{aligned}
&\int \nabla\psi_k(x)\cdot\nabla(V(x)+D\log\rho_\theta(x))\rho_\theta(x)\,dx\\
&=\int -\nabla\cdot(\rho_\theta(x)\nabla\psi_k(x))\cdot(V(x)+D\log\rho_\theta(x))\,dx\\
&=\int -\nabla\cdot(\rho_\theta(x)\partial_{\theta_k}T_\theta(T_\theta^{-1}(x)))\cdot(V(x)+D\log\rho_\theta(x))\,dx\\
&=\int(\nabla V(T_\theta(x))+D\nabla\log\rho_\theta(T_\theta(x)))\cdot\partial_{\theta_k}T_\theta(x)\,dp(x)\\
&=\int\nabla V(T_\theta(x))\cdot\partial_{\theta_k}T_\theta(x)+\partial_{\theta_k}[D\log\rho_\theta(T_\theta(x))]\,dp(x)-\underbrace{\int D\,\partial_{\theta_k}\log\rho_\theta(T_\theta(x))\,dp(x)}_{=D\int\nabla_\theta\rho_\theta(x)dx=0}\\
&=\partial_{\theta_k}\left(\int(V(T_\theta(x))+D\log\rho_\theta(T_\theta(x)))\,dp(x)\right)=\partial_{\theta_k}H(\theta).
\end{aligned}
$$

Recall the definition (Equation 4.13) of $G(\theta)$, the target function to be minimized is $\eta^{\mathrm{T}}G(\theta)\eta-2\eta^{\mathrm{T}}\nabla_\theta H(\theta)$. And the minimizer is clearly $G(\theta)^{-1}\nabla_\theta H(\theta)$. □

In addition to the direct proof, the result in Theorem 4.3.5 can also be understood in a different way. Let us denote $\xi=G(\theta)^{-1}\nabla_\theta H(\theta)$, $\{\theta_t\}$ solves (Equation 4.28) with initial value $\theta_0=\theta$. By Theorem 4.3.4, $\frac{d}{dt}\rho_{\theta_t}\Big|_{t=0}=(T_{\theta\sharp})_*\xi\in\mathcal{T}_{\rho_\theta}\mathcal{P}_\Theta$ is the orthogonal projection of $\mathrm{grad}_W\mathcal{H}(\rho_\theta)$ onto $\mathcal{T}_{\rho_\theta}\mathcal{P}_\Theta$ with respect to the metric g^W. This is equivalent to say that η solves the following least square problem:

$$
\min_\eta g^W(\mathrm{grad}_W\mathcal{H}(\rho_\theta)-(T_{\theta\sharp})_*\eta,\ \mathrm{grad}_W\mathcal{H}(\rho_\theta)-(T_{\theta\sharp})_*\eta). \tag{4.30}
$$

Recall the definition of g^W in subsection 2.2.1 and by (Equation 2.61), we have $\mathrm{grad}_W\mathcal{H}(\rho_\theta)=-\nabla\cdot(\rho_\theta\nabla(V+D\log\rho_\theta))$. Because of (Equation 4.18), $(T_{\theta\sharp})_*\eta=-\nabla\cdot(\rho_\theta\partial_\theta T_\theta(T_\theta^{-1}(\cdot))\eta)$, solving $-\nabla\cdot(\rho_\theta\nabla\varphi)=\mathrm{grad}_W\mathcal{H}(\rho_\theta)-(T_{\theta\sharp})_*\eta$ gives

$$
\varphi=(V+D\log\rho_\theta)-\mathbf{\Psi}^{\mathrm{T}}\eta,
$$

and thus least squares problem (Equation 4.30) can be written as

$$\min_{\eta} \left\{ \int |\nabla \boldsymbol{\Psi}(x)^{\mathrm{T}} \eta - \nabla (V(x) + D \log \rho_\theta(x))|^2 \rho_\theta(x)\, dx \right\},$$

which is exactly (Equation 4.29).

4.3.3 A particle viewpoint of the parametric Fokker Planck Equation

The motion of parameter θ_t solving (Equation 4.28) naturally induce a stochastic dynamics on $\mathbb{R}^d$ whose density evolution is exactly $\{\rho_{\theta_t}\}$. To see this, notice that $\{\theta_t\}$ directly leads to a time dependent map $\{T_{\theta_t}\}$. Let us denote a random variable $\boldsymbol{Z} \sim p$, i.e. $\boldsymbol{Z}$ is distributed according to the reference distribution p. We set $\boldsymbol{Y}_0 = T_{\theta_0}(\boldsymbol{Z}) \sim \rho_{\theta_0}$. At any time t, the map T_{θ_t} sends $\boldsymbol{Y}_0$ to $\boldsymbol{Y}_t = T_{\theta_t}(T_{\theta_0}^{-1}(\boldsymbol{Y}_0)) \sim \rho_{\theta_t}$. Thus, we construct a sequence of random variables $\{\boldsymbol{Y}_t\}$ whose density evolution is exactly $\{\rho_{\theta_t}\}$. We can characterize the dynamical system satisfied by $\{\boldsymbol{Y}_t\}$ by taking time derivative: $\dot{\boldsymbol{Y}}_t = \partial_\theta T_{\theta_t}(\boldsymbol{Z})\dot{\theta}_t = \partial_\theta T_{\theta_t}(T_{\theta_t}^{-1}(\boldsymbol{Y}_t))\dot{\theta}_t$. It is actually more insightful to consider the following dynamic:

$$\dot{\boldsymbol{X}}_t = \nabla \boldsymbol{\Psi}_t(\boldsymbol{X}_t)^{\mathrm{T}}\, \dot{\theta}_t, \quad \boldsymbol{X}_0 = T_{\theta_0}(\boldsymbol{Z}) \sim \rho_{\theta_0}. \tag{4.31}$$

Here $\boldsymbol{\Psi}_t$ is obtained from (Equation 4.14) with parameter θ_t. It is not hard to show that for any time t, $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ has the same distribution. Thus $\boldsymbol{X}_t \sim \rho_{\theta_t}$ for all $t \geq 0$. Recall $\dot{\theta}_t = -G(\theta_t)^{-1}\nabla_\theta H(\theta_t)$, we are able to rewrite (Equation 4.31) as:

$$\dot{\boldsymbol{X}}_t = \nabla \boldsymbol{\Psi}_t(\boldsymbol{X}_t)^{\mathrm{T}} \underbrace{\left(\int \nabla \boldsymbol{\Psi}_t \nabla \boldsymbol{\Psi}_t^{\mathrm{T}} \rho_{\theta_t} dx \right)}_{G(\theta_t)}^{-1} \underbrace{\left(\int \nabla \boldsymbol{\Psi}_t (-\nabla V - D\nabla \log \rho_{\theta_t}) \rho_{\theta_t} d\eta \right)}_{-\nabla_\theta H(\theta_t)}. \tag{4.32}$$

If we define the kernel function $K_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as

$$K_\theta(x, \eta) = \nabla \boldsymbol{\Psi}^{\mathrm{T}}(x) \left(\int \nabla \boldsymbol{\Psi}(x) \nabla \boldsymbol{\Psi}(x)^{\mathrm{T}}\, \rho_\theta(x)\, dx \right)^{-1} \nabla \boldsymbol{\Psi}(\eta).$$

This K_θ induces a linear operator $\mathcal{K}_\theta : L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_\theta) \to L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_\theta)$ by:

$$\mathcal{K}_\theta[\vec{v}] = (\mathcal{K}_\theta * \vec{v})(\cdot) = \int K_\theta(\cdot, \eta)\ \vec{v}(\eta)\ \rho_\theta(\eta)\ d\eta.$$

It can be verified that $\mathcal{K}_\theta$ is an orthogonal projection on the Hilbert space $L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_\theta)$. The range of such projection is the subspace span $\{\nabla\psi_1, ..., \nabla\psi_m\} \subset L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_\theta)$. Here $\psi_1, ..., \psi_m$ are the m components of $\mathbf{\Psi}$ solved from (Equation 4.14). Using the linear operator, we can rewrite (Equation 4.32) as:

$$\dot{\boldsymbol{X}}_t = -\mathcal{K}_{\theta_t}[\nabla V + D\nabla \log \rho_{\theta_t}](\boldsymbol{X}_t), \quad \rho_{\theta_t} \text{ is the probability density of } \boldsymbol{X}_t,\ \boldsymbol{X}_0 \sim \rho_{\theta_0}. \tag{4.33}$$

We can compare (Equation 4.33) with the following dynamic without projection:

$$\dot{\tilde{\boldsymbol{X}}}_t = -(\nabla V + D\nabla \log \rho_t)(\tilde{\boldsymbol{X}}_t), \quad \rho_t \text{ is the probability density of } \tilde{\boldsymbol{X}}_t,\ \tilde{\boldsymbol{X}}_0 \sim \rho_0. \tag{4.34}$$

As discussed in subsection 4.2.1, (Equation 4.34) is the Vlasov-type SDE that involves the density of random particle. If assuming (Equation 4.34) admits a regular solution, we have $\rho(x,t) = \rho_t(x)$, which solves the original Fokker Planck equation (Equation 4.6). From orthogonal projection viewpoint, we can treat that the approximate solution ρ_{θ_t} of (Equation 4.6) is actually originated from the projection of vector field that drives the SDE (Equation 4.34).

We would like to mention that the expectation of ℓ^2 discrepancy between $\nabla V + D\nabla \log \rho$ and its $\mathcal{K}_\theta$ projection is:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{X}\sim\rho_\theta} |\mathcal{K}_\theta[\nabla V + D\nabla \log \rho_\theta](\boldsymbol{X}) - (\nabla V + D\nabla \log \rho_\theta)(\boldsymbol{X})|^2 \\ =& \int |\nabla \mathbf{\Psi}(x)^{\mathrm{T}} \xi - (-\nabla V - D\nabla \log \rho_\theta)(x)|^2 \rho_\theta(x)\ dx, \end{aligned} \tag{4.35}$$

in which $\xi = -G(\theta)^{-1}\nabla_\theta H(\theta)$. This is an essential term appeared in our error analysis

part.

Remark 9. *We should mention the relationship between our kernel K_{θ_t} and the Neural Tangent Kernel (NTK) introduced in [140]. Using our notation, Neural Tangent Kernel can be written as $K_\theta^{NTK} = \partial_\theta T_\theta(x)\partial_\theta T_\theta(\xi)^T$. If we consider the flat gradient flow $\dot{\theta} = -\nabla_\theta H(\theta)$ of relative entropy on Θ, its corresponding particle dynamic is*

$$\dot{\boldsymbol{X}}_t = \int K_{\theta_t}^{NTK}(T_{\theta_t}^{-1}(\boldsymbol{X}_t), T_{\theta_t}^{-1}(\eta))(-\nabla V(\eta) - D\nabla \log \rho_{\theta_t}(\eta))\rho_{\theta_t}(\eta)\, d\eta$$

Different from our K_θ, which introduces an orthogonal projection, Neural Tangent Kernel introduces an non-negative definite transform to the vector field $-\nabla V - D\nabla \log \rho_{\theta_t}$.

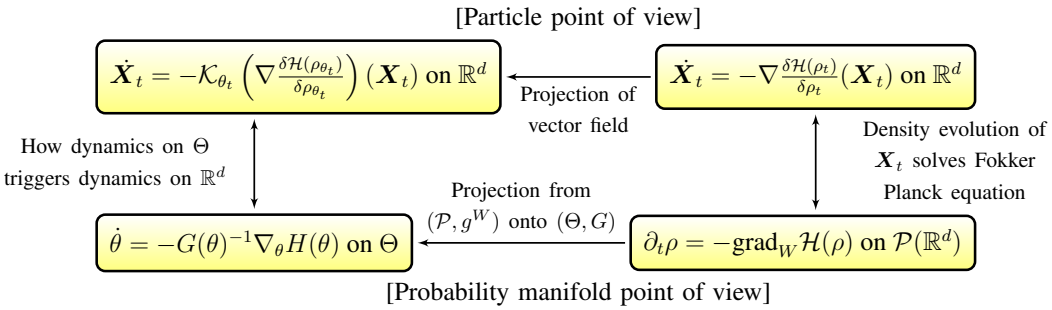


Figure 4.1: Illustrative diagram

Remark 10. *Figure 4.1 illustrates the relation between* (Equation 4.6)*,* (Equation 4.28)*,* (Equation 4.34) *and* (Equation 4.33)*. It is worth mentioning that the probability manifold point of view discussed in Theorem 4.3.4 is useful for our analysis of the continuous dynamics* (Equation 4.28)*, while particle point of view helps us on establishing the numerical analysis for the time discrete scheme (i.e. forward-Euler) of* (Equation 4.28)*.*

4.3.4 An example of the parametric Fokker-Planck equation with quadratic potential

The solution of the parametric Fokker-Planck equation (Equation 4.28) can serve as an approximation to the solution of the original equation (Equation 4.6). In some special cases, ρ_{θ_t} exactly solves (Equation 4.6). In this section, we provide such examples.

Let us consider the Fokker-Planck equations with quadratic potentials whose initial conditions are Gaussian:

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \quad \text{and} \quad \rho_0 \sim \mathcal{N}(\mu_0, \Sigma_0). \tag{4.36}$$

Here $\mathcal{N}(\mu, \Sigma)$ denotes Gaussian distribution with mean μ and covariance Σ. We consider parameter space $\Theta = (\Gamma, b) \subset \mathbb{R}^m$ ($m = \frac{1}{2}d(d+1) + d$), where Γ is a $d \times d$ symmetric positive definite matrix and $b \in \mathbb{R}^d$. We define the parametric map as $T_\theta(x) = \Gamma x + b$, and choose the reference measure $p = \mathcal{N}(0, I)$.

Lemma 4.3.6. *Let $\mathcal{H}$ be the relative entropy defined in (Equation 4.8) and H defined in (Equation 4.27). For $\theta \in \Theta$, if the vector function $\nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right) \circ T_\theta$ can be written as the linear combination of $\{\frac{\partial T_\theta}{\partial\theta_1}, ..., \frac{\partial T_\theta}{\partial\theta_m}\}$, i.e. there exists $\zeta \in \mathbb{R}^m$, such that $\nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$. Then:*
(1) $\zeta = G(\theta)^{-1}\nabla_\theta H(\theta)$, which is the Wasserstein gradient of H at θ.
(2) $\mathcal{P}_\Theta$ as $\mathrm{grad}_W\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta}$, then $\mathrm{grad}_W\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}_W\mathcal{H}(\rho_\theta)$, where $\mathrm{grad}_W\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta}$ is the gradient of $\mathcal{H}$ on the submanifold $\mathcal{P}_\Theta$.

Proof. Suppose that $\zeta \in \mathbb{R}^m$ satisfies $\nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$, then we have

$$\int |\partial_\theta T_\theta(x)\zeta - \nabla(\frac{\delta\mathcal{H}}{\delta\rho}) \circ T_\theta(x)|^2 \, dp(x) = 0.$$

By definition of $\mathbf{\Psi}$ in Theorem 4.3.2, one can verify

$$-\nabla \cdot \left(\rho_\theta \left((\nabla\mathbf{\Psi})^{\mathrm{T}}\zeta - \nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right)\right)\right) = -\nabla \cdot \left(\rho_\theta \left(\partial_\theta T_\theta \circ T_\theta^{-1}\zeta - \nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right)\right)\right)$$

Now we apply (Lemma 4.3.1) of Lemma 4.3.1 to obtain:

$$\int |(\nabla\mathbf{\Psi}(T_\theta(x)))^{\mathrm{T}}\zeta - \nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right) \circ T_\theta(x)|^2 \, dp(x) \leq 0.$$

This implies,

$$\begin{aligned}&\inf_{\eta}\int |(\nabla\mathbf{\Psi}(T_\theta(x)))^{\mathrm{T}}\eta - \nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right)\circ T_\theta(x)|^2\, dp(x)\\ =&\int |(\nabla\mathbf{\Psi}(T_\theta(x)))^{\mathrm{T}}\zeta - \nabla\left(\frac{\delta\mathcal{H}}{\delta\rho}\right)\circ T_\theta(x)|^2\, dp(x) = 0.\end{aligned}$$

By Theorem 4.3.5, we get $\zeta = G(\theta)^{-1}\nabla_\theta H(\theta)$ and $\|(T_{\theta\sharp})_*\zeta - \mathrm{grad}_W\mathcal{H}(\rho_\theta)\|_{g^W(\rho_\theta)} = 0$. The latter leads to $(T_{\theta\sharp})_*\zeta = \mathrm{grad}_W\mathcal{H}(\rho_\theta)$. According to Theorem 4.3.4, $(T_{\theta\sharp})_*\zeta = \mathrm{grad}_W\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta}$. As a result, we have $\mathrm{grad}_W\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}_W\mathcal{H}(\rho_\theta)$. □

Back to our example with quadratic potential (Equation 4.36) and $T_\theta(x) = \Gamma x + b$, we can compute

$$\rho_\theta(x) = T_{\theta\sharp}p(x) = \frac{f(T_\theta^{-1}(x))}{|\det(\Gamma)|} = \frac{f(\Gamma^{-1}(x-b))}{|\det(\Gamma)|},\ f(x) = \frac{\exp(-\frac{1}{2}|x|^2)}{(2\pi)^{\frac{d}{2}}}.$$

Then we have,

$$\nabla\left(\frac{\delta\mathcal{H}(\rho_\theta)}{\delta\rho}\right)\circ T_\theta(x) = \nabla(V + D\log\rho_\theta)\circ T_\theta(x) = \Sigma^{-1}(\Gamma x + b - \mu) - D\Gamma^{-T}x,$$

which is affine with respect to x.

Notice that

$$\partial_{\Gamma_{ij}}T_\theta(x) = (\dots, 0, \dots, \underset{i-\mathrm{th}}{x_j}, \dots, 0, \dots)^{\mathrm{T}},\ \partial_{b_i}T_\theta = (\dots, 0, \dots, \underset{i-\mathrm{th}}{1}, \dots, 0, \dots)^{\mathrm{T}}.$$

We can verify that $\zeta = (\Sigma^{-1}\Gamma - D\Gamma^{-T}, \Sigma^{-1}(b-\mu))$ solves $\nabla\left(\frac{\delta\mathcal{H}(\rho_\theta)}{\delta\rho}\right)\circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$. By (1) of Lemma 4.3.6, $\zeta = G(\theta)^{-1}\nabla_\theta H(\theta)$. Thus the ODE (Equation 4.28) for our

example can be written as

$$\dot{\Gamma} = -\Sigma^{-1}\Gamma + D\Gamma^{-T} \quad \Gamma_0 = \sqrt{\Sigma_0}, \tag{4.37}$$

$$\dot{b} = \Sigma^{-1}(\mu - b) \quad b_0 = \mu_0. \tag{4.38}$$

By (2) of Lemma 4.3.6, we know $\mathrm{grad}_{\mathrm{W}}\mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}_{\mathrm{W}}\mathcal{H}(\rho_\theta)$ for all $\theta \in \Theta$, which indicates that there is no error between our parametric Fokker-Planck and the original equations.

Following (Equation 4.37) and (Equation 4.38), we have the following corollary,

Corollary 4.3.6.1. *The solution of the Fokker-Planck equation (Equation 4.6) with condition(Equation 4.36) is a Gaussian distribution for all $t > 0$.*

Proof. If we denote $\{\Gamma_t, b_t\}$ as the solutions to (Equation 4.37),(Equation 4.38), set $\theta_t = (\Gamma_t, b_t)$, then $\rho_t = T_{\theta_t \sharp} p$ solves the Fokker Planck Equation (Equation 4.6) with conditions (Equation 4.36). Since the pushforward of Gaussian distribution p by an affine transform T_θ is still a Gaussian, we conclude that for any $t > 0$, the solution $\rho_t = T_{\theta_t \sharp} p$ is always Gaussian distribution. □

Remark 11. *This is a well known property for Ornstein–Uhlenbeck process [141]. We provide an alternative proof under our framework.*

4.4 Numerical method for 1D Fokker-Planck equation

Since the Wasserstein metric tensor G has an explicit solution when dimension $d = 1$, it is convenient to numerically compute ODE (Equation 4.28).

For example, we can choose a series of basis functions $\{\varphi_k\}_{k=1}^n$. Each φ_k can be chosen as a sinusoidal function or a piece-wise linear function defined on a certain interval $[-l, l]$. It is also beneficial to choose orthogonal or near-orthogonal basis functions because they

will keep the metric tensor G far away from ill-posedness. We set $T_\theta(x) = \sum_{k=1}^m \theta_k \varphi_k(x)^2$. Then according to (Equation 4.23), we can compute G as

$$G_{ij}(\theta) = \mathbb{E}_{\mathbf{X}\sim p}\Big[\varphi_i(\mathbf{X})\varphi_j(\mathbf{X})\Big] \quad 1 \le i, j \le m$$

Recall that $F(\theta) = \int V(x)\rho_\theta(x)dx + \beta \int \rho_\theta(x)\log\rho_\theta(x)dx$. The second part of F is the entropy of ρ_θ. For general T_θ, $\rho_\theta = T_{\theta\sharp}p$ cannot be directly computed efficiently. However, we can compute the entropy term by solving the following variational problem [142]:

$$\int \rho_\theta(x)\log\rho_\theta(x)\,dx = \sup_h \Big\{ \int h(x)\rho_\theta(x)\,dx - \int e^{h(x)}dx \Big\} + 1 \tag{4.39}$$

We can solve (Equation 4.39) by parametrizing h. Suppose the optimal solution is h^*. Then by envelope theorem [143], we can compute $\nabla_\theta F(\theta)$ as

$$\begin{aligned}\nabla_\theta F(\theta) &= \partial_\theta \left(\int V(x)\rho_\theta(x)\,dx + \beta \int h^*(x)\rho_\theta(x)\,dx \right) \\ &= \mathbb{E}_{\mathbf{x}\sim p}\Big[\partial_\theta T_\theta(\mathbf{X})^T \nabla_y (V(y) + \beta h^*(y))|_{y=T_\theta(\mathbf{X})}\Big]\end{aligned} \tag{4.40}$$

Notice that both the metric tensor G and $\nabla_\theta F(\theta)$ are written in forms of expectations, thus we can compute them by Monte Carlo simulations. And finally, (Equation 4.28) can be computed by forward Euler method.

Our numerical results are always demonstrated by sample points: For each time node t, we sample points $\{\mathbf{X}_1, ..., \mathbf{X}_N\}$ from p, then $\{T_{\theta_t}(\mathbf{X}_1), ..., T_{\theta_t}(\mathbf{X}_N)\}$ are our numerical samples from distribution ρ_t which solves the Fokker-Planck equation.

Here are two illustrative numerical results based on our method. We exhibit them in the form of histograms. Consider the potential $V(x) = (x+1)^2(x-1)^2$. Suppose the initial distribution is $\rho_0 = \mathcal{N}(0, I)$. Figure 4.2 contains histograms of ρ_t which solves

[2] In application, suitably choosing T_θ which is not necessarily invertibile or smooth can still provide valid results.

$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V)$ at different time nodes; we know ρ_t converges to $\frac{\delta_{-1}+\delta_{+1}}{2}$ as $t \to \infty$. Here δ_a is the Dirac distribution concentrated on point a. Figure 4.3 contains histograms of ρ_t which solves $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \frac{1}{4}\Delta \rho$ at different time nodes, we know ρ_t will converge to Gibbs distribution $\rho_* = \frac{1}{Z}\exp(-4(x+1)^2(x-1)^2)$, with Z being a normalizing constant, as $t \to \infty$. The density function of ρ_* is exhibited in Figure 4.3.

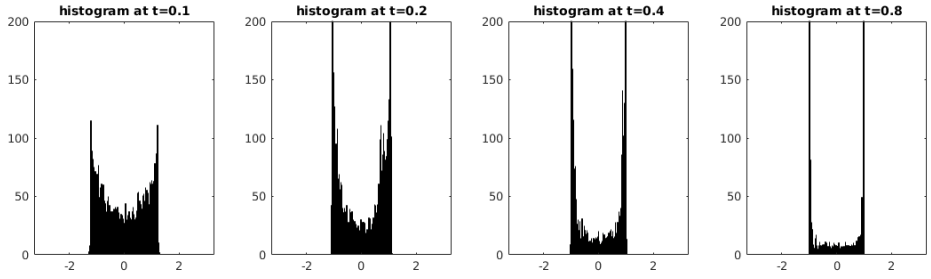


Figure 4.2: Histograms of ρ_t solving $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V)$

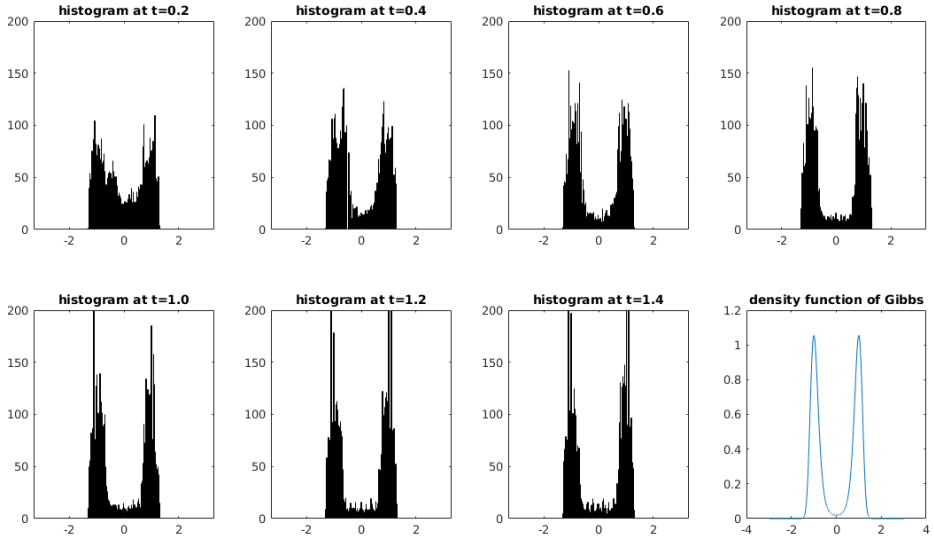


Figure 4.3: Histograms of ρ_t solving $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \frac{1}{4}\Delta \rho$

4.5 Numerical methods for high dimensional Fokker-Planck equations

In this section, we introduce our sampling efficient numerical method to compute the proposed parametric Fokker-Planck equations.

Before we start, we want to mention that as stated in [130], when dimension $d = 1$, $G(\theta)$ has explicit solution. Thus the push-forward approximation of 1D Fokker-Planck equation can be directly computed by solving the ODE system (Equation 4.28) with numerical methods, such as forward-Euler scheme. In this section, our focus is on numerical methods for (Equation 4.28) with dimension $d \geq 2$. It turns out to be very challenging to compute (Equation 4.28) by the forward-Euler scheme directly. There are two reasons. One is that there is no known explicit formula for $G(\theta)$, and direct computation based on (Equation 4.13) can be expansive because it requires to solve multiple differential equations. The other is incurred by the high dimensionality, which is the main goal of this research. To overcome the challenge of dimensionality, we choose to use deep neural networks to construct our $T(\theta)$. However, directly evaluating $G(\theta)^{-1}\nabla_\theta H(\theta)$ is difficult, alternative strategies must be sought.

There are a few papers investigating numerical methods for gradient flows on Riemannian manifolds, such as Fisher natural gradient [144] and Wasserstein gradient [115]. The well known JKO scheme [83] calculates the time discrete approximation of the Wasserstein gradient flow using an optimization formulation,

$$\partial_t \rho_t = -\text{grad}_W \mathcal{F}(\rho_t), \qquad \rho_{k+1} = \operatorname*{argmin}_{\rho \in \mathcal{P}} \left\{ \frac{W_2^2(\rho, \rho_k)}{2h} + \mathcal{F}(\rho) \right\}, \tag{4.41}$$

where h is the time step size, $\mathcal{F}$ could be a suitable functional defined on $\mathcal{P}$. Along the line of JKO scheme, there are further developments in machine learning recently [145].

In our approach, we design schemes that computes the exact Wasserstein gradient flow directly with provable accuracy guarantee. Our algorithms are completely sample based so that they can be run efficiently under deep learning framework, and can scale up to high dimensional cases.

4.5.1 Normalizing Flow as push forward maps

We choose T_θ as the so-called normalizing flow [103]. Here is a brief sketch of its structure: T_θ is written as the composition of K invertible nonlinear transforms:

$$T_\theta = f_K \circ f_{K-1} \circ ... \circ f_2 \circ f_1,$$

where each f_k $(1 \leq k \leq K)$ takes the form

$$f_k(x) = x + \sigma(w_k^{\mathrm{T}} x + b_k) u_k.$$

Here $w_k, u_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, and σ is a nonlinear function, which can be chosen as tanh for example. In [103], it has been shown that f_k is invertible iff $w_k^{\mathrm{T}} u_k \geq -1$. Figure 4.4 shows several snapshots of how a normalizing flow T_θ with length equal to 10 pushes forward standard Gaussian distribution to a target distribution.

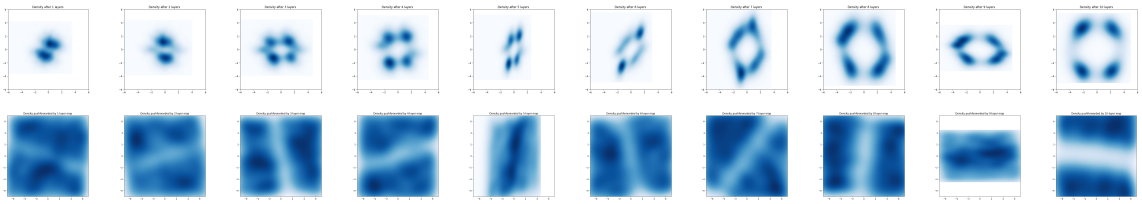


Figure 4.4: Top row from left to right are the probability densities of distributions $f_{1\sharp}p$, $(f_2 \circ f_1)_\sharp p$, ..., $(f_{10} \circ f_9 \circ ... \circ f_1)_\sharp p$. The last image displays our target distribution. Bottom row displays the push-forward effect of each single-layer transformation f_k $(1 \leq k \leq 10)$.

In a normalizing flow, the parameters are: $\theta = (w_1, u_1, b_1, ..., w_K, u_K, b_K)$. The determinant of the Jacobi matrix of T_θ, an important quantity for our schemes, can be explicitly computed by

$$\mathbf{det}\left(\frac{\partial T_\theta(x)}{\partial x}\right) = \prod_{k=1}^{K} (1 + \sigma'(w_k^{\mathrm{T}} x_k + b_k) w_k^{\mathrm{T}} u_k),$$

where $x_k = f_k \circ f_{k-1} \circ ... \circ f_1(x)$. Using the structure of normalizing flow, the logarithm

of the density $\rho_\theta = T_{\theta\sharp}p$ can be written as

$$\log \rho_\theta(x) = \log p \circ T_\theta^{-1}(x) - \sum_{k=1}^{K} \log(1 + \sigma'(w_k^{\mathrm{T}} \tilde{x}_k) w_k^{\mathrm{T}} u_k), \tag{4.42}$$

$$\tilde{x}_k = f_k \circ ... \circ f_1(T_\theta^{-1}(x)) = f_{k+1}^{-1} \circ ... \circ f_K^{-1}(x).$$

Then we can explicitly write the relative entropy functional $H(\theta)$ defined in (Equation 4.27) as,

$$H(\theta) = \mathbb{E}_{\mathbf{X}\sim p}[V(T_\theta(\mathbf{X})) + \mathcal{L}_\theta(\mathbf{X})], \tag{4.43}$$

where $\mathcal{L}_\theta$ is defined by,

$$\mathcal{L}_\theta(\cdot) = \log p(\cdot) - \sum_{k=1}^{K} \log(1 + \sigma'(w_k^{\mathrm{T}} F_k(\cdot)) w_k^{\mathrm{T}} u_k) \quad F_k(\cdot) = f_k \circ f_{k-1} \circ ... \circ f_1(\cdot).$$

Once $H(\theta)$ is computed explicitly, so does the gradient $\nabla_\theta H(\theta)$.

In summary, we choose the normalizing flow because it has sufficient expression power to approximate complicated distributions on $\mathbb{R}^d$ [103], and the relative entropy $H(\theta)$ has a very concise form (Equation 4.43), and its gradient can be conveniently computed.

Remark 12. *We want to emphasize here that the normalizing flow is not the only choice for T_θ. One may choose other network structures [146, 147] as long as they have sufficient approximation power and can compute the gradient of relative entropy efficiently.*

4.5.2 Numerical scheme

For the convenience of our presentation, at the beginning of this section, we first introduce the following definition.

Definition 4.5.1 (Orthogonal projection onto space of gradient fields)**.** *Consider vector field $\vec{v} \in L^2(\mathbb{R}^d; \mathbb{R}^d, \rho)$. Define $\mathrm{Proj}_\rho[\vec{v}] = \nabla\psi$ as the $L^2(\rho)$-orthogonal projection of $\vec{v}$*

onto the subspace of gradient fields. Where ψ solves:

$$\min_{\psi} \left\{ \int |\vec{v}(x) - \nabla\psi(x)|^2 \rho(x)\, dx \right\}. \tag{4.44}$$

Or equivalently ψ solves $-\nabla \cdot (\rho(x)\nabla\psi(x)) = -\nabla \cdot (\rho(x)\vec{v}(x))$.

Proposed Double-Minimization Scheme

Our numerical scheme is inspired by the following semi-implicit scheme of (Equation 4.28),

$$\frac{\theta_{k+1} - \theta_k}{h} = -G^{-1}(\theta_k)\nabla_\theta H(\theta_{k+1}).$$

Equivalently, we can write it as a proximal algorithm,

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{2}\langle \theta - \theta_k, G(\theta_k)(\theta - \theta_k)\rangle + hH(\theta) \right\}. \tag{4.45}$$

Recall $\mathbf{\Psi}$ as defined in Theorem 4.3.2, if we denote $\psi = \mathbf{\Psi}^{\mathrm{T}}(\theta - \theta_k)$, we have $\langle (\theta - \theta_k), G(\theta)(\theta - \theta_k)\rangle = \int |\nabla\psi|^2 \rho_{\theta_k}\, dx$ with ψ solves the equation

$$-\nabla \cdot (\rho_{\theta_k}\nabla\psi(x)) = -\nabla \cdot (\rho_{\theta_k}\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta - \theta_k)). \tag{4.46}$$

By Definition 4.5.1, $\nabla\psi$ is the orthogonal projection of vector field $\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(\cdot))(\theta - \theta_k)$. Equivalently, ψ can also be obtained by solving the least square problem (Equation 4.44).

Based on the observation that $\nabla\psi$ is obtained via orthogonal projection after replacing $\partial_\theta T_{\theta_k}(\theta - \theta_k)$ by finite difference $T_\theta - T_{\theta_k}$, we end up with the following double-

minimization scheme for solving (Equation 4.45)

$$\begin{aligned}&\min_{\theta}\left\{\left(\int\left(2\,\nabla\phi(x)\cdot((T_\theta-T_{\theta_k})\circ T_{\theta_k}^{-1}(x))-|\nabla\phi(x)|^2\right)\rho_{\theta_k}(x)\,dx\right)+2hH(\theta)\right\}\\&\text{with }\phi\text{ solves: }\ \min_{\phi}\left\{\int|\nabla\phi(x)-((T_\theta-T_{\theta_k})\circ T_{\theta_k}^{-1}(x))|^2\rho_{\theta_k}(x)\,dx\right\}.\end{aligned}\tag{4.47}$$

Scheme (Equation 4.47) has an equivalent saddle point optimization formulation

$$\min_{\theta}\max_{\phi}\left\{\left(\int(2\nabla\phi(x)\cdot((T_\theta-T_{\theta_k})\circ T_{\theta_k}^{-1}(x))-|\nabla\phi(x)|^2)\rho_{\theta_k}(x)\,dx\right)+2hH(\theta)\right\},\tag{4.48}$$

which can be directly derived from (Equation 4.45) via adjoint method. Their equivalence is explained in the next remark.

Remark 13. *We briefly demonstrate the equivalence among three schemes* (Equation 4.45)*,* (Equation 4.47) *and* (Equation 4.48)*. Our target function* $\frac{1}{2}\langle\theta-\theta_k,G(\theta_k)(\theta-\theta_k)\rangle+hH(\theta)$ *can be formulated as*

$$\int\frac{1}{2}|\nabla\psi(x)|^2\rho_{\theta_k}(x)\,dx+hH(\theta)\quad\textit{with the constraint: }\psi\textit{ solves}\text{ (Equation 4.46).}$$

By introducing the dual variable ϕ*, and applying the adjoint method, we obtain*

$$\begin{aligned}&\frac{1}{2}\langle\theta-\theta_k,G(\theta_k)(\theta-\theta_k)\rangle+hH(\theta)\\&=\max_{\phi}\min_{\psi}\left\{\int\frac{1}{2}|\nabla\psi|^2\rho_{\theta_k}dx+hH(\theta)\right.\\&\qquad\left.+\int\phi(\nabla\cdot(\rho_{\theta_k}\nabla\psi)-\nabla\cdot(\rho_{\theta_k}\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)))\,dx\right\}\\&=\max_{\phi}\min_{\psi}\left\{\int\left(\frac{1}{2}|\nabla\psi|^2-\nabla\phi\cdot\nabla\psi+\nabla\phi\cdot\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)\right)\rho_{\theta_k}\,dx+hH(\theta)\right\}\\&=\max_{\phi}\left\{\int\left(-\frac{1}{2}|\nabla\phi|^2+\nabla\phi\cdot\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)\right)\rho_{\theta_k}\,dx+hH(\theta)\right\}\end{aligned}\tag{4.49}$$

In implementation, we substitute $\partial_\theta T_{\theta_k}(\theta-\theta_k)$ *by* $T_\theta - T_{\theta_k}$ *since the latter is more tractable in computation. As a consequence, by substituting* (Equation 4.49) *into* (Equation 4.45) *we obtain (by multiplying the entire function by* 2*) the saddle scheme* (Equation 4.48)*. To verify the equivalence between* (Equation 4.48) *and* (Equation 4.47)*, we check the identity*

$$\begin{aligned}&\int (2\nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\,dx\\ =&-\int |\nabla\phi(x) - (T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)|^2\rho_{\theta_k}(x)\,dx + \underbrace{\int |(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)|^2\rho_{\theta_k}(x)\,dx}_{\textit{Constant w.r.t. }\phi}\end{aligned}$$

Thus the ϕ*-minimization process of* (Equation 4.47) *is equivalent to the* ϕ*-maximization process of* (Equation 4.48)*. This leads to the equivalence between* (Equation 4.47) *and* (Equation 4.48)*.*

Remark 14. *Our proposed schemes* (Equation 4.47)*,* (Equation 4.48) *can be viewed as an approximation to the JKO scheme* (Equation 4.41) *with* $\mathcal{F}$ *being the relative entropy* $H(\theta)$*. To see this, we denote*

$$\mathcal{E}(\phi) = \int (2\nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\,dx,$$

and set $\hat{\psi} = \underset{\phi}{\text{argmax}}\,\mathcal{E}(\phi)$*. We let* $\vec{v}_h(x) = \frac{1}{h}(T_\theta\circ T_{\theta_k}^{-1}(x) - x)$*. Under mild conditions, one can show*

$$W_2^2(\rho_\theta,\rho_{\theta_k}) = W_2^2((Id + h\vec{v}_h)_\sharp\rho_{\theta_k},\rho_{\theta_k}) = \int |\nabla\hat{\psi}|^2\rho_{\theta_k}\,dx + o(h^2) = \max_\phi \mathcal{E}(\phi) + o(h^2). \tag{4.50}$$

By replacing $W_2^2(\rho_\theta,\rho_{\theta_k})$ *in* (Equation 4.41) *by its approximation* $\max_\phi \mathcal{E}(\phi)$*, we obtain scheme* (Equation 4.47)*,* (Equation 4.48)*.*

Although (Equation 4.47) and (Equation 4.48) are mathematically equivalent, we use them for different purposes. The saddle scheme (Equation 4.48) is our main tool to investigate the theoretical properties of our proposed method in subsubsection 4.5.2, because

it better reflects the nature of our approximation method. In our implementation, as discussed in subsubsection 4.5.2, we prefer the double-minimization scheme (Equation 4.47). Our experience indicates that (Equation 4.47) makes our code run more efficiently and behaves more stably than (Equation 4.48).

Local error of the proposed scheme

We now analyze the local error of scheme (Equation 4.48) as well as (Equation 4.47) compared with the semi-implicit scheme (Equation 4.45). Let us denote $\max_\phi \mathcal{E}(\phi)$ as $\widehat{W}_2^2(\theta, \theta_k)$ (Here $\widehat{W}_2$ is treated as an approximation of L^2-Wasserstein distance (remark Remark 14)). It is straightforward to verify $\widehat{W}_2(\theta, \theta') \geq 0$ and $\widehat{W}_2(\theta, \theta) = 0$. Consider the following assumption,

$$\widehat{W}_2^2(\theta, \theta') \geq l(|\theta - \theta'|) \qquad \text{for any} \quad \theta, \theta' \in \Theta. \tag{4.51}$$

Here $l : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfies $l(0) = 0$. $l(r)$ is continuous, strictly increasing when $r \leq r_0$ for a positive r_0 and is bounded below by $\lambda_0 > 0$ when $r > r_0$. Notice that this assumption generally guarantees positive definiteness of $\widehat{W}_2$. Clearly, (Equation 4.51) only depends on the structure of T_θ, and we expect that (Equation 4.51) holds for the neural networks used as pushforward maps, including the ones we used in this research.

Theorem 4.5.1. *Suppose assumption* (Equation 4.51) *holds true for the class of pushforward maps* $\{T_\theta\}$. *Then the local error of scheme* (Equation 4.48) *is of order* h^2*, i.e., assume that* θ_{k+1} *is the optimal solution to* (Equation 4.48)*, then*

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})| \sim O(h^2). \tag{4.52}$$

or equivalently: $\limsup_{h\to 0^+} \frac{|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})|}{h^2} < +\infty$.

Before proving Theorem 4.5.1, we introduce a few additional notations. We define ϵ

ball in parameter space as $B_\epsilon(\theta_k) = \{\theta \mid |\theta - \theta_k| \leq \epsilon\}$, let $T_\theta^{(i)}$ be the i-th component ($1 \leq i \leq d$) of map T_θ. For fixed θ_k and $\epsilon > 0$ small enough, we assume the following two quantities are finite

$$L(\theta_k, \epsilon) = \sum_{i=1}^{d} \mathbb{E}_{x\sim p} \sup_{\theta \in B_\epsilon(\theta_k)} \left\{ |\partial_\theta T_\theta^{(i)}(x)|^2 \right\}, \; H(\theta_k, \epsilon) = \sum_{i=1}^{d} \mathbb{E}_{x\sim p} \sup_{\theta \in B_\epsilon(\theta_k)} \left\{ \|\partial_{\theta\theta}^2 T_\theta^{(i)}(x)\|_2^2 \right\}. \tag{4.53}$$

To prove Theorem 4.5.1, we need the following three lemmas:

Lemma 4.5.2. *Suppose we fix $\theta_0 \in \Theta$, for arbitrary $\theta \in \Theta$ and $\nabla\phi \in L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_{\theta_0})$ we consider*

$$F(\theta, \nabla\phi \mid \theta_0) = \left(\int (2\nabla\phi(x) \cdot (T_\theta - T_{\theta_0}) \circ T_{\theta_0}^{-1}(x) - |\nabla\phi(x)|^2) \, \rho_{\theta_0}(x) \, dx \right) + 2hH(\theta). \tag{4.54}$$

Then $F(\theta, \nabla\phi \mid \theta_0) < \infty$, furthermore, $F(\cdot, \nabla\phi \mid \theta_0) \in C^1(\Theta)$. We can compute

$$\partial_\theta F(\theta, \nabla\phi \mid \theta_0) = 2 \left(\int \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T \, \nabla\phi(x) \, \rho_{\theta_0}(x) \, dx + h \, \nabla_\theta H(\theta) \right). \tag{4.55}$$

Lemma 4.5.3. *Suppose we fix $\theta_0 \in \Theta$ and define $J(\theta) = \sup_{\nabla\phi \in L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_{\theta_0})} F(\theta, \nabla\phi \mid \theta_0)$. Then J is differentiable. If we denote $\hat{\psi}_\theta = \operatorname{argmax}_\phi \{F(\theta, \nabla\phi \mid \theta_0)\}$, then*

$$\nabla_\theta J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) = 2 \left(\int \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T \, \nabla\hat{\psi}_\theta(x) \, \rho_{\theta_0}(x) \, dx + h \, \nabla_\theta H(\theta) \right).$$

This lemma is an anology of the envelope theorem [143] under our problem setting.

Lemma 4.5.4. *Under assumption*(Equation 4.51)*, the optimal solution of* (Equation 4.48) *θ_{k+1} satisfies,*

$$|\theta_{k+1} - \theta_k| \sim o(1) \quad \textit{i.e.} \quad \lim_{h \to 0^+} |\theta_{k+1} - \theta_k| = 0.$$

This lemma provides *a prior* estimation of $|\theta_{k+1} - \theta_k|$.

We prove Lemma 4.5.2, Lemma 4.5.3 and Lemma 4.5.4 in Appendix section C.3.

Proof of Theorem 4.5.1. Let us consider $F(\theta, \nabla\phi \mid \theta_k)$, we denote

$$\nabla\hat{\psi}_\theta = \underset{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_k})}{\operatorname{argmax}} \{F(\theta, \nabla\phi \mid \theta_k)\}.$$

Then we can set

$$\nabla\hat{\psi}_\theta = \text{Proj}_{\rho_{\theta_k}}[(T_\theta - T_{\theta_k}) \circ T_{\theta_k}^{-1}], \quad \text{and} \quad J(\theta) = \sup_{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_k})} F(\theta, \nabla\phi \mid \theta_k)$$

Apply Lemma 4.5.3, we obtain:

$$\nabla_\theta J(\theta) = 2\left(\int \partial_\theta T_\theta(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\, \nabla\hat{\psi}_\theta(x)\, \rho_{\theta_k}(x)\, dx + h\, \nabla_\theta H(\theta)\right).$$

Due to the differentiability of $J(\theta)$, at the optimizer θ_{k+1}, the gradient must vanish, i.e.

$$\left(\int \partial_\theta T_{\theta_{k+1}}(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\, \nabla\hat{\psi}_{\theta_{k+1}}(x)\, \rho_{\theta_k}(x)\, dx\right) + h\nabla_\theta H(\theta_{k+1}) = 0. \tag{4.56}$$

We use Taylor expansion at θ_{k+1} to get $T_{\theta_{k+1}} - T_{\theta_k} = \partial_\theta T_{\theta_k}(\theta_{k+1} - \theta_k) + R(\theta_{k+1}, \theta_k)$, in which $R(\theta, \theta')(\cdot) \in L^2(\mathbb{R}^d; \mathbb{R}^m, \rho_{\theta_k})$, the i-th entry of $R(\theta, \theta')$ is $R_i(\theta, \theta')(x) = \frac{1}{2}(\theta - \theta')^{\mathrm{T}}\partial^2_{\theta\theta} T^{(i)}_{\tilde{\theta}_i(x)}(x)(\theta - \theta'),\ 1 \le i \le m$, where each $\tilde{\theta}_i(x) = \lambda_i(x)\theta + (1 - \lambda_i(x))\theta'$ for some $\lambda_i(x) \in [0, 1]$. Then we can write:

$$\begin{aligned}\nabla\hat{\psi}_{\theta_{k+1}} =& \text{Proj}_{\rho_{\theta_k}}[(T_{\theta_{k+1}} - T_{\theta_k}) \circ T_{\theta_k}^{-1}] \\ =& \text{Proj}_{\rho_{\theta_k}}[\partial_\theta T_{\theta_k} \circ T_{\theta_k}^{-1}(\theta_{k+1} - \theta_k)] + \text{Proj}_{\rho_{\theta_k}}[R(\theta_{k+1}, \theta_k) \circ T_{\theta_k}^{-1}].\end{aligned} \tag{4.57}$$

On the other hand,

$$\partial_\theta T_{\theta_{k+1}} = \partial_\theta T_{\theta_k} + r(\theta_{k+1}, \theta_k). \tag{4.58}$$

Here $r(\theta, \theta') \in L^2(\mathbb{R}^d; \mathbb{R}^{d\times m}, \rho_{\theta_k})$, the (i, j) entry of $r(\theta, \theta')$ is $(\theta_{k+1} - \theta_k)^{\mathrm{T}}\partial_\theta(\partial_{\theta_j} T^{(i)}_{\tilde{\theta}_{ij}(x)}(x))$, $1 \le i \le d,\ 1 \le j \le m$, where each $\tilde{\theta}_{ij}(x) = \mu_{ij}(x)\theta_{k+1} + (1 - \mu_{ij}(x))\theta_k$, for some

$\mu_{ij}(x) \in (0,1)$. Applying (Equation 4.58), (Equation 4.57) to (Equation 4.56), we obtain

$$\begin{aligned}
&\int \partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))^{\mathrm{T}} \mathrm{Proj}_{\rho_{\theta_k}}[\partial_\theta T_{\theta_k} \circ T_{\theta_k}^{-1}(x)(\theta_{k+1}-\theta_k)]\ \rho_{\theta_k}(x)\ dx \\
&+ \int \partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))^{\mathrm{T}} \mathrm{Proj}_{\rho_{\theta_k}}[R(\theta_{k+1},\theta_k) \circ T_{\theta_k}^{-1}](x)\ \rho_{\theta_k}(x)\ dx \\
&+ \int r(\theta_{k+1},\theta_k)(T_{\theta_k}^{-1}(x))^{\mathrm{T}} \mathrm{Proj}_{\rho_{\theta_k}}[(T_{\theta_{k+1}} - T_{\theta_k}) \circ T_{\theta_k}^{-1}](x)\ \rho_{\theta_k}(x)\ dx \quad = -h\nabla_\theta H(\theta_{k+1}).
\end{aligned} \tag{4.59}$$

Recall definition of $\mathbf{\Psi}$ in Theorem 4.3.2, use (Equation 4.11) in Lemma 4.3.1, we know that the first term on the left hand side of (Equation 4.59) equals

$$\int \nabla \mathbf{\Psi}(x) \nabla \mathbf{\Psi}(x)^{\mathrm{T}} (\theta_{k+1}-\theta_k)\ \rho_{\theta_k}(x)\ dx = G(\theta_k)(\theta_{k+1}-\theta_k).$$

By applying Cauchy–Schwarz inequality and (Equation 4.12) in Lemma 4.3.1, we bound the i-th entry of the second term in (Equation 4.59) by:

$$\begin{aligned}
&\left(\int |\partial_\theta T_{\theta_k}^{(i)}(x)|^2\ dp(x) \cdot \int \sum_{i=1}^{d} |(\theta_{k+1}-\theta_k)\partial_{\theta\theta}^2 T_{\tilde{\theta}_i(x)}^{(i)}(x)(\theta_{k+1}-\theta_k)|^2\ dp(x) \right)^{\frac{1}{2}} \\
&\leq \left(\mathbb{E}_p |\partial_\theta T_{\theta_k}^{(i)}(x)|^2 \cdot \mathbb{E}_p \left[\sum_{i=1}^{d} \|\partial_{\theta\theta}^2 T_{\tilde{\theta}_i(x)}^{(i)}(x)\|_2 \right] \right)^{\frac{1}{2}} |\theta_{k+1}-\theta_k|^2 \overset{\text{denote as}}{=} A^{(i)} |\theta_{k+1}-\theta_k|^2.
\end{aligned}$$

To bound the third term in (Equation 4.59), we consider $T_{\theta_{k+1}}(x) - T_{\theta_k}(x)$, the i-th entry can be written as

$$T_{\theta_{k+1}}^{(i)}(x) - T_{\theta_k}^{(i)}(x) = \partial_\theta T_{\bar{\theta}_i(x)}(x)(\theta_{k+1}-\theta_k),$$

here $\bar{\theta}_i(x) = \zeta_i(x)\theta_{k+1} + (1-\zeta_i(x))\theta_k$ for some $\zeta_i(x) \in (0,1)$. The i-th entry of the third

term of (Equation 4.59) can be bounded by:

$$\left(\int \sum_{i=1}^{d} |(\theta_{k+1}-\theta_k)^{\mathrm{T}} \partial_{\theta\theta} T^{(i)}_{\tilde{\theta}_{ij}(x)}(x)|^2 \, dp(x) \cdot \int |T^{(i)}_{\theta_{k+1}}(x) - T^{(i)}_{\theta_k}(x)|^2 \, dp(x)\right)^{\frac{1}{2}}$$
$$\leq \left(\mathbb{E}_p\left[\sum_{i=1}^{d} \|\partial^2_{\theta\theta} T_{\tilde{\theta}_{ij}(x)}(x)\|_2^2\right] \cdot \mathbb{E}_p |\partial_\theta T^{(i)}_{\tilde{\theta}_i(x)}(x)|^2\right)^{\frac{1}{2}} |\theta_{k+1}-\theta_k|^2 \overset{\text{denote as}}{=} B^{(i)} |\theta_{k+1}-\theta_k|^2.$$

We denote $A \in \mathbb{R}^m$ with entries $A^{(i)}$, $1 \leq i \leq m$ and similarly $B \in \mathbb{R}^m$ with entries $B^{(i)}$, $1 \leq i \leq m$. (Equation 4.59) leads to the following inequality,

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})| \leq \|G(\theta_k)^{-1}\|_2 (|A| + |B|) \, |\theta_{k+1}-\theta_k|^2.$$

As we have shown in Lemma 4.5.4 that $|\theta_{k+1} - \theta_k| \sim o(1)$ for any $\epsilon > 0$ when step size h is small enough, we always have $\theta_{k+1} \in B_\epsilon(\theta_k)$. Recall the notations in (Equation 4.53), we have $|A|, |B| \leq \sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)}$. Thus we have

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})| \leq 2\sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)} \|G(\theta_k)^{-1}\|_2 |\theta_{k+1}-\theta_k|^2.$$

Denote $\theta_{k+1} - \theta_k = \eta$, $G(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1}) = \xi$ and $C = 2\sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)} \|G(\theta_k)^{-1}\|_2$, the previous inequality is

$$|\eta \; - \; h \, \xi| \leq C|\eta|^2. \tag{4.60}$$

Since $|\eta - h\xi| \geq |\eta| - h|\xi|$, we have

$$C|\eta|^2 \geq |\eta| - h|\xi|. \tag{4.61}$$

Solving (Equation 4.61) gives

$$|\eta| \leq \frac{2|\xi|h}{1 + \sqrt{1 - 4C|\xi|h}} \quad \text{or} \quad |\eta| > \frac{1 + \sqrt{1 - 4Ch|\xi|}}{2C}.$$

The second inequality leads to $|\theta_{k+1} - \theta_k| > \frac{1}{2C}$ for any $h > 0$, which avoids $|\theta_{k+1} - \theta_k| \sim o(1)$. Thus, when h is sufficiently small, we have

$$|\eta| \leq \frac{2|\xi|h}{1 + \sqrt{1 - 4C|\xi|h}}. \tag{4.62}$$

Combining (Equation 4.62) and (Equation 4.60), we have:

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})| \leq \frac{4\,C\,|\xi|^2}{(1 + \sqrt{1 - 4C|\xi|h})^2}\,h^2 \leq 4C|\xi|^2h^2. \tag{4.63}$$

This proves the result. □

Remark 15. *One may be aware of the relation between the positive definite condition* (Equation 4.51) *and the positive definiteness of the metric tensor $G(\theta_k)$. A positive definite $G(\theta)$ guarantees the inequality $\widehat{W}_2^2(\theta, \theta') \geq C|\theta - \theta'|^2$ for $\theta' \in B_{r_0}(\theta)$ (r_0 depends on θ is small enough). However, we are not able to bound $\widehat{W}_2^2(\theta, \theta')$ from below when $|\theta - \theta'| > r_0$. On the other hand,* (Equation 4.51) *is a locally weaker condition than positive definiteness of $G(\theta)$.*

Implementation

As mentioned before in subsubsection 4.5.2, we prefer the double-minimization scheme (Equation 4.47) than the saddle scheme (Equation 4.48). We will thus implement scheme (Equation 4.47). Let us denote

$$J(\theta) = \left(\int \left(2\,\nabla\hat{\psi}(T_{\theta_k}(x)) \cdot ((T_\theta(x) - T_{\theta_k}(x))) - |\nabla\hat{\psi}(T_{\theta_k}(x))|^2 \right) dp(x) \right) + 2hH(\theta) \tag{4.64}$$

$$\text{with } \hat{\psi} = \underset{\phi}{\text{argmin}} \left\{ \int |\nabla\phi(T_{\theta_k}) - (T_\theta(x) - T_{\theta_k}(x))|^2 dp(x) \right\} \tag{4.65}$$

We then solve ODE (Equation 4.28) at t_k by solving

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} J(\theta), \tag{4.66}$$

Here we provide some detailed discussion on our implementation.

- In our numerical computation, we approximate ϕ by $\psi_\nu : M \to \mathbb{R}$, which is a ReLU neural network [148]. Here ν denotes the parameter vector of the network ψ_ν. We know that in this case, ψ_ν is a piece-wise affine function and its gradient $\nabla\psi_\nu(\cdot)$ forms a piece-wise constant vector field.

- The entire procedure of solving (Equation 4.66) can be formulated as nested loops:

  – (inner loop) Every inner loop aims at solving (Equation 4.65) on ReLU functions ψ_ν, i.e. solving:

  $$\min_{\nu} \left\{ \mathbb{E}_{\boldsymbol{X}\sim p} |\nabla\psi_\nu(T_{\theta_k}(\boldsymbol{X})) - (T_\theta(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X}))|^2 \right\}. \tag{4.67}$$

  One can use Stochastic Gradient Descent (SGD) methods like RMSProp [112] or ADAM [149] with learning rate α_{in} to deal with this inner loop optimization. In our implementation, we will stop after M_{in} iterations. Let us denote the optimal ν in each inner loop as $\hat{\nu}$;

  – (outer loop) We apply similar SGD method to $J(\theta)$: using Lemma 4.5.3, we are able to compute $\nabla_\theta J(\theta)$ as:

  $$\nabla_\theta J(\theta) = \partial_\theta \left(\left(\int 2\nabla\hat{\psi}(x) \cdot (T_\theta \circ T_{\theta_k}^{-1}(x)) \rho_{\theta_k}(x)\, dx \right) + 2hH(\theta) \right).$$

  If we treat optimal $\hat{\psi}$ as $\psi_{\hat{\nu}}$, what we need to do in each outer loop is to consider:

$$\tilde{J}(\theta) = \mathbb{E}_{\boldsymbol{X}\sim p}\, 2[\nabla\psi_{\hat{\nu}}(T_{\theta_k}(\boldsymbol{X}))\cdot T_\theta(\boldsymbol{X})] + 2h[V(T_\theta(\boldsymbol{X})) + \mathcal{L}_\theta(\boldsymbol{X})] \quad (4.68)$$

and update θ for one step by our chosen SGD method with learning rate α_{out} applied to optimize $\tilde{J}(\theta)$. In our actual computation, we will stop the outer loop after M_{out} iterations.

- We now present the entire algorithm for computing (Equation 4.28) based on the scheme (Equation 4.47) in Algorithm Algorithm 2. This algorithm contains the following parameters: $T, N; M_{\text{out}}, K_{\text{out}}, \alpha_{\text{out}}; M_{\text{in}}, K_{\text{in}}, \alpha_{\text{in}}$. Recall we set reference distribution p as standard Gaussian on $M = \mathbb{R}^d$.

Remark 16 (Rescaling)**.** *In our implementation, $T_\theta(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X})$ is usually of order $O(\alpha_{out})$, which is a small quantity. We can rescale it so that each inner loop can be solved in a more stable way with larger stepsize (learning rate). That is to say, we choose some small $\epsilon \sim O(\alpha_{out})$ and consider*

$$\min_\theta \max_\phi \left\{ \underbrace{\left(\int (2\nabla\phi(x)\cdot\left(\frac{1}{\epsilon}(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)\right) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\, dx \right)}_{\mathcal{E}_\epsilon(\phi)} + \frac{2h}{\epsilon^2}H(\theta) \right\}. \quad (4.69)$$

We can also check

$$\textit{argmax}\, \mathcal{E}_\epsilon(\phi) = \textit{Proj}_{\rho_{\theta_k}}[\frac{1}{\epsilon}(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}] = \frac{1}{\epsilon}\textit{Proj}_{\rho_{\theta_k}}[(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}] = \frac{1}{\epsilon}\textit{argmax}\, \mathcal{E}(\phi).$$

Using this, we are able to verify $\max_\phi\ \mathcal{E}_\epsilon(\phi) = \frac{1}{\epsilon^2}\max_\phi\ \mathcal{E}(\phi)$. *Thus the optimal solution of* (Equation 4.69) *is*

$$\textit{argmin}_\theta \left\{ \frac{1}{\epsilon^2}\max_\phi\ \mathcal{E}(\phi) + \frac{2h}{\epsilon^2}H(\theta) \right\} = \textit{argmin}_\theta \left\{ \max_\phi\ \mathcal{E}(\phi) + 2hH(\theta) \right\}$$

Algorithm 2 Computing (Equation 4.28) by scheme (Equation 4.48) on the time interval $[0, T]$

Initialize θ
for $i = 1, ..., N$ **do**
    Save current parameter value to θ_0: $\theta_0 = \theta$
    for $j = 1, ...M_{\text{out}}$ **do**
        for $p = 1, ..., M_{\text{in}}$ **do**
            Sample $\{\mathbf{X}_1, ..., \mathbf{X}_{K_{\text{in}}}\}$ from p
            Apply one SGD (ADAM) step with learning rate α_{in} to loss function of variable λ.

$$\frac{1}{K_{\text{in}}} \left(\sum_{k=1}^{K_{\text{in}}} |\nabla \psi_\nu(T_{\theta_0}(\mathbf{X}_k)) - (T_\theta(\mathbf{X}_k) - T_{\theta_0}(\mathbf{Y}_k))|^2 \right)$$

        end for
        Sample $\{\mathbf{X}_1, ..., \mathbf{X}_{K_{\text{out}}}\}$ from p
        Apply one SGD (ADAM) step with learning rate α_{out} to loss function of variable θ.

$$\frac{1}{K_{\text{out}}} \left(\sum_{k=1}^{K_{\text{out}}} 2[\nabla \psi_\nu(T_{\theta_0}(\boldsymbol{X}_k)) \cdot T_\theta(\boldsymbol{X}_k)] + 2h[V(T_\theta(\boldsymbol{X}_k)) + \mathcal{L}_\theta(\boldsymbol{X}_k)] \right)$$

    end for
    Set $\theta_i = \theta$
end for
The sequence of probability densities $\{T_{\theta_0 \sharp} p, T_{\theta_1 \sharp} p, ..., T_{\theta_N \sharp} p\}$ will be the numerical solution of $\{\rho_{t_0}, \rho_{t_1}, ..., \rho_{t_N}\}$, where $t_i = i\frac{T}{N}$ $(i = 0, 1, ..., N-1, N)$. Here ρ_t solves the original Fokker-Planck equation (Equation 4.6).

This shows that the equivalence between the modified scheme (Equation 4.69) *and the original scheme* (Equation 4.48).

In our actual implementation, we still prefer double-minimization scheme. We solve

$$\min_{\nu} \left\{ \mathbb{E}_{\boldsymbol{X}\sim p} \left| \nabla \psi_{\nu}(T_{\theta_k}(\boldsymbol{X})) - \left(\frac{T_{\theta}(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X})}{\epsilon} \right) \right|^2 \right\}, \tag{4.70}$$

instead of (Equation 4.67) *in each inner loop and set:*

$$\tilde{J}(\theta) = \mathbb{E}_{\boldsymbol{X}\sim p}\, 2[\nabla \psi_{\hat{\nu}}(T_{\theta_k}(\boldsymbol{X})) \cdot T_{\theta}(\boldsymbol{X})] + \frac{2h}{\epsilon}[V(T_{\theta}(\boldsymbol{X})) + \mathcal{L}_{\theta}(\boldsymbol{X})] \tag{4.71}$$

in each outer loop. In actual experiments, we set $\epsilon = \alpha_{out}$.

Remark 17 (Sufficiently large sample size)**.** *It is worth mentioning that the sample size* K_{in}, K_{out} *in each SGD step (especially* K_{in}*) should be chosen reasonably large so that the inner optimization problem can be solved with enough accuracy. In our practice, we usually choose* $K_{in} = K_{out} = \max\{1000, 300d\}$*. Here* d *is the dimension of sample space. This is very different from the small batch technique applied to training neural network in deep learning [150].*

Remark 18 (Using fixed samples)**.** *Our numerical experiments indicate that the same samples can be used for both the inner and outer iterations, which may reduce the computational cost of our original algorithm.*

4.6 Asymptotic properties and error estimations

In this section, we establish numerical analysis for the parametric Fokker-Planck equation (Equation 4.28).

4.6.1 An important quantity

Before our analysis, we introduce an important quantity that plays an essential role in our numerical analysis. Let us recall the optimal value of least square problem (Equation 4.29) in Theorem 4.3.5 of subsection 4.3.2, or equivalently (Equation 4.30) of subsection 4.3.2, (Equation 4.35) of subsection 4.3.3. If we denote the upper bound of all possible values to be δ_0, i.e.

$$\delta_0 = \sup_{\theta\in\Theta} \min_{\xi\in\mathbb{R}^m} \left\{ \int \left| \sum_{k=1}^{M} \xi_k \nabla\psi_k(x) - \nabla\left(V(x) + D\log\rho_\theta(x)\right) \right|^2 \rho_\theta(x)dx \right\}, \tag{4.72}$$

where ψ_k are solutions to (Equation 4.14) in Theorem 4.3.2. This quantity provides crucial error bound between our parametric equation and original equation in the forthcoming analysis. Ideally, we hope δ_0 to be sufficiently small. And this can be guaranteed if the neural network we select has universal approximation power. δ_0 can be bounded by another constant with more approachable form

$$\hat{\delta}_0 = \sup_{\theta\in\Theta} \min_{\xi\in\mathbb{R}^m} \left\{ \int \left| \sum_{k=1}^{M} \xi_k \frac{\partial T_\theta(x)}{\partial\theta_k} - \nabla\left(V(x) + D\log\rho_\theta(x)\right) \right|^2 \rho_\theta(x)dx \right\}. \tag{4.73}$$

By (Equation 4.12) of Lemma 4.3.1, one can verify $\delta_0 \leq \hat{\delta}_0$. From (Equation 4.73), we observe that $\hat{\delta}_0$ is determined by the optimal linear combination of $\{\frac{\partial T_\theta}{\partial\theta_k}\}_{k=1}^M$ to approximate the vector field $\nabla(V + D\log\rho_\theta)$. One may understand this approximation from three different aspects.

- If T_θ is chosen as a linear combination of basis functions, i.e. $T_\theta(x) = \sum_{k=1}^M \theta_k\Phi_k(x)$, we can give an explicit estimate on $\hat{\delta}_0$. For example, if $\Phi_k(x)$ is picked as the Fourier basis and $\nabla(V + D\log\rho_\theta) \in H^s$ ($s > 1$), the classical spectral method theory can be applied to obtain an estimate $\hat{\delta}_0 = O(M^{-s})$ [151, 152]. If Radial Basis Function is selected, an related approximation bounded can be obtained too [153].

- Having a small value for $\hat{\delta}_0$ as well as δ_0 is equivalent to find a suitable T_θ such that a specific vector field $\nabla(V + D\log\rho_\theta)$ can be accurately approximated in our estimate. In other words, when neural networks are used for T_θ, one needs to pick a neural network structure such that it can approximate $\nabla(V + D\log\rho_\theta)$ well. This seems to be an easier question than the task for the so-called universal approximation theory for neural networks, which requires T_θ to approximate an arbitrary function in a space.

- In our implementation, we use Normalizing Flows, a special type of deep neural networks. Our numerical examples seem to show promising performance. In the existing literature, although there are several references providing the universal approximation power of neural networks [154, 155], the results are mainly focused on general ReLU networks and on the approximation power of function value, which is different from our case. To the best of our knowledge, there is no existing study discussing explicit bounds for vector field approximation by deep neural networks. We believe that the question of how δ_0 or $\hat{\delta}_0$ explicitly depends on the structure of T_θ is a fundamental research problem that deserves careful investigations.

It is also worth mentioning that δ_0 is used for *a priori* estimate in this section, because we don't know the exact trajectory of $\{\theta_t\}$ when solving ODE (Equation 4.28), and we take supremum over Θ to obtain δ_0. Once solved for $\{\theta_t\}$, denote $\mathcal{C}$ as the set covering its trajectory, i.e.

$$\mathcal{C} = \{\theta \mid \exists\, t \geq 0,\ \text{s.t. } \theta = \theta_t\} \tag{4.74}$$

We define another quantity δ_1:

$$\delta_1 = \sup_{\theta\in\mathcal{C}} \min_{\xi\in\mathcal{T}_\theta\Theta} \left\{ \int \left| (\nabla\boldsymbol{\Psi}(T_\theta(x))^{\mathrm{T}}\xi - \nabla\left(V + D\log\rho_\theta\right)\circ T_\theta(x) \right|^2 dp(x) \right\}. \tag{4.75}$$

Clearly, we have $\delta_1 \leq \delta_0$. We can obtain corresponding *posterior* estimates for the asymptotic convergence and error analysis by replacing δ_0 with δ_1.

4.6.2 Asymptotic Convergence Analysis

In this section, we consider the solution $\{\theta_t\}_{t\geq 0}$ of our parametric Fokker-Planck equation (Equation 4.28). We define:

$$\mathcal{V} = \left\{ V \;\middle|\; \begin{array}{c} V \in \mathcal{C}^2(\mathbb{R}^d),\, V \text{ can be decomposed as: } V = U + \phi, \text{ with } U, \phi \in \mathcal{C}^2(\mathbb{R}^d); \\ \nabla^2 U \succeq KI \text{ with } K > 0 \text{ and } \phi \in L^\infty(\mathbb{R}^d) \end{array} \right\}$$

As we know, for the Fokker-Planck equation (Equation 4.6), when the potential $V \in \mathcal{V}$, $\{\rho_t\}$ will converge to the Gibbs distribution $\rho_* = \frac{1}{Z_D} e^{-V(x)/D}$ as $t \to \infty$ under the measure of KL divergence [156]. For (Equation 4.28), we wish to study its asymptotic convergence property. We come up with the following result:

Theorem 4.6.1 (*a priori* estimation on asymptotic convergence)**.** *Consider the Fokker-Planck equation* (Equation 4.6) *with the potential $V \in \mathcal{V}$. Suppose $\{\theta_t\}$ solves the parametric Fokker-Planck equation* (Equation 4.28)*, denote δ_0 as in* (Equation 4.72)*. Let $\rho_*(x) = \frac{1}{Z_D} e^{-V(x)/D}$ be the Gibbs distribution of original equation* (Equation 4.6)*. Then we have the inequality:*

$$\mathcal{D}_{KL}(\rho_{\theta_t} \| \rho_*) \leq \frac{\delta_0}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0} \| \rho_*) e^{-D\tilde{\lambda}_D t}. \tag{4.76}$$

Here $\tilde{\lambda}_D > 0$ is the constant associated to the Logarithmic Sobolev inequality discussed in Lemma 4.6.2 with potential function $\frac{1}{D}V$.

To prove Theorem 4.6.1, we need the following two lemmas:

Lemma 4.6.2. *[Holley-Stroock Perturbation] Suppose the potential $V \in \mathcal{V}$ is decomposed as $V = U + \phi$ where $\nabla^2 U \succeq KI$ and $\phi \in L^\infty$. Let $\tilde{\lambda} = Ke^{-osc(\phi)}$, where $osc(\phi) = \sup \phi - \inf \phi$. Then the following Logarithmic Sobolev inequality holds for any probability density ρ:*

$$\mathcal{D}_{KL}(\rho \| \rho_*) \leq \frac{1}{\tilde{\lambda}} \mathcal{I}(\rho | \rho_*). \tag{4.77}$$

Here $\rho_* = \frac{1}{Z} e^{-V}$ and $\mathcal{I}(\rho|\rho_*)$ is the Fisher information functional defined as:

$$\mathcal{I}(\rho|\rho_*) = \int \left| \nabla \log \left(\frac{\rho(x)}{\rho_*(x)} \right) \right|^2 \rho(x)\, dx.$$

Lemma 4.6.2 is first proved in [156].

Lemma 4.6.3. *For any $\theta \in \Theta$, we have:*

$$D^2\, \mathcal{I}(\rho_\theta|\rho_*) \leq \delta_0 + \nabla_\theta H(\theta) \cdot G(\theta)^{-1} \nabla_\theta H(\theta), \tag{4.78}$$

where δ_0 is defined in (Equation 4.72).

Proof of Lemma 4.6.3. Let us denote $\xi = G(\theta)^{-1} \nabla_\theta H(\theta)$ for convenience. Suppose $\{\theta_t\}$ solves (Equation 4.28) with $\theta_0 = \theta$. By Theorem 4.3.4, $\frac{d}{dt} \rho_{\theta_t} \Big|_{t=0} = -(T_{\theta\sharp})_* \xi$ is orthogonal projection of $-\text{grad}_W \mathcal{H}(\rho_\theta)$ onto $\mathcal{T}_{\rho_\theta} \mathcal{P}$ with respect to metric g^W. Thus the orthogonal relation gives:

$$\begin{aligned} g^W(-\text{grad}_W \mathcal{H}(\rho_\theta), -\text{grad}_W \mathcal{H}(\rho_\theta)) = g^W(\text{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_* \xi, \text{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_* \xi) \\ + g^W((T_{\theta\sharp})_* \xi, (T_{\theta\sharp})_* \xi). \end{aligned} \tag{4.79}$$

One can verify that the left hand side of (Equation 4.79) is:

$$g^W(-\text{grad}_W \mathcal{H}(\rho_\theta), -\text{grad}_W \mathcal{H}(\rho_\theta)) = \int |\nabla(V(x) + D \log \rho_\theta(x))|^2 \rho(x)\, dx = D^2\, \mathcal{I}(\rho_\theta|\rho_*). \tag{4.80}$$

Recall the equivalence between (Equation 4.29) and (Equation 4.30) and the definition of δ_0 in (Equation 4.72), we know that the first term on the right hand side of (Equation 4.79) has an upper bound

$$g^W(\text{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_* \xi, \text{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_* \xi) \leq \delta_0. \tag{4.81}$$

The second term on the right hand side of (Equation 4.79) is:

$$g^W((T_{\theta\sharp})_*\xi, (T_{\theta\sharp})_*\xi) = (T_{\theta\sharp})^* g^W(\xi,\xi) = G(\theta)(G(\theta)^{-1}\nabla_\theta H(\theta),\ G(\theta)^{-1}\nabla_\theta H(\theta))$$
$$= \nabla_\theta H(\theta)\cdot G(\theta)^{-1}\nabla_\theta H(\theta) \tag{4.82}$$

Combining (Equation 4.79), (Equation 4.80),(Equation 4.81) and (Equation 4.82) yields to (Equation 4.78). □

Proof of Theorem 4.6.1. Let us recall the relationship between KL divergence and relative entropy,

$$\mathcal{D}_{\mathrm{KL}}(\rho\|\rho_*) = \frac{1}{D}\mathcal{H}(\rho) + \log(Z_D).$$

Actually, we can treat $\mathcal{D}_{\mathrm{KL}}(\rho_\theta\|\rho_*)$ as a Lyapunov function for our ODE (Equation 4.28), because by taking time derivative of $\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*)$, we obtain

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) = \frac{1}{D}\frac{d}{dt}\mathcal{H}(\rho_{\theta_t}) = \frac{1}{D}\dot{\theta}_t\cdot\nabla H(\theta_t) = -\frac{1}{D}\nabla H(\theta_t)\cdot G^{-1}(\theta_t)\nabla H(\theta_t).$$

Using the inequality in Lemma 4.6.3, we are able to show:

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{D} - D\,\mathcal{I}(\rho_{\theta_t}|\rho_*).$$

By Lemma 4.6.2, we have:

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{D} - D\,\tilde{\lambda}_D\,\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*).$$

Therefore we obtain, by Grownwall's inequality, the following estimate,

$$\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{\tilde{\lambda}_D D^2}(1-e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{\mathrm{KL}}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t}.$$

□

Remark 19. *Following the previous proof, we can show a similar convergence estimation for the solution* $\{\rho_t\}_{t\geq 0}$ *of* (Equation 4.6)*. Such result was first discovered in [157].*

$$\mathcal{D}_{KL}(\rho_t\|\rho_*) \leq \mathcal{D}_{KL}(\rho_0\|\rho_*)\, e^{-D\tilde{\lambda}_D t} \quad \forall\, t > 0. \tag{4.83}$$

A nominal modification of our proof for Theorem 4.6.1 leads to a *posterior* version of our asymptotic convergence analysis, which is stated in the following theorem.

Theorem 4.6.4 (*Posterior* estimation on asymptotic convergence)**.**

$$\mathcal{D}_{KL}(\rho_{\theta_t}\|\rho_*) \leq \frac{\delta_1}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t},$$

where δ_1 *is defined in* (Equation 4.75).

4.6.3 Wasserstein error estimations

In this subsection, we establish our error bounds for both continuous and discrete version of the parametric Fokker-Planck equation (Equation 4.28) as approximations to the original equation (Equation 4.6).

Wasserstein error for the parametric Fokker-Planck equation

The following theorem provides an upper bound between the solutions of (Equation 4.6) and (Equation 4.28).

Theorem 4.6.5. *Assume that* $\{\theta_t\}_{t\geq 0}$ *solves* (Equation 4.28) *and* $\{\rho_t\}_{t\geq 0}$ *solves* (Equation 4.6)*. If the Hessian of the potential function* V *in* (Equation 4.6) *is bounded below by a constant* λ*, i.e.* $\nabla^2 V \succeq \lambda\, I$*. the* 2*-Wasserstein difference between* ρ_t *and* ρ_{θ_t} *can be bounded as*

$$W_2(\rho_{\theta_t}, \rho_t) \leq \Omega_\lambda(t) = \begin{cases} \frac{\sqrt{\delta_0}}{\lambda}(1 - e^{-\lambda t}) + e^{-\lambda t} W_2(\rho_{\theta_0}, \rho_0), & \text{if } \lambda \neq 0, \\ \sqrt{\delta_0} t + W_2(\rho_{\theta_0}, \rho_0), & \text{if } \lambda = 0. \end{cases} \tag{4.84}$$

To prove this inequality, we need the following lemmas.

Lemma 4.6.6 (Constant speed of geodesic)**.** *The geodesic connecting $\rho_0, \rho_1 \in \mathcal{P}(M)$ is described by,*

$$\begin{cases} \frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t \nabla \psi_t) = 0 \\ \frac{\partial \psi_t}{\partial t} + \frac{1}{2}|\nabla \psi_t|^2 = 0 \end{cases} \qquad \rho_t|_{t=0} = \rho_0,\ \rho_t|_{t=1} = \rho_1. \tag{4.85}$$

Using the notation $\dot{\rho}_t = \partial_t \rho_t = -\nabla \cdot (\rho_t \nabla \psi_t) \in \mathcal{T}_{\rho_t}\mathcal{P}(M)$, $g^W(\dot{\rho}_t, \dot{\rho}_t)$ is constant for $0 \le t \le 1$ and $g^W(\dot{\rho}_t, \dot{\rho}_t) = W_2^2(\rho_0, \rho_1)$ for $0 \le t \le 1$.

Lemma 4.6.7 (Displacement convexity of relative entropy)**.** *Suppose $\{\rho_t\}$ solves the geodesic equation* (Equation 4.85)*, the relative entropy $\mathcal{H}$ in* (Equation 4.8) *has potential V satisfying $\nabla^2 V \succeq \lambda I$, then we have $\frac{d}{dt} g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) \ge \lambda W_2^2(\rho_0, \rho_1)$. Or equivalently, $\frac{d^2}{dt^2}\mathcal{H}(\rho_t) \ge \lambda W_2^2(\rho_0, \rho_1)$.*

Lemma 4.6.6 originates from section 7.2 of [35]. A generalization of it has been proved in Lemma 5 of [158]. A more general version on the displacement convexity related to Lemma 4.6.7 has been discussed in chapter 16 and 17 of [7]. To be self-contained, we provide direct proofs to both Lemma 4.6.6 and Lemma 4.6.7 in Appendix section C.4.

Proof of Theorem 4.6.5. Figure 4.5 provides a sketch of our proof: For a given time t, the geodesic $\{\bar{\rho}_\tau\}_{0 \le \tau \le 1}$ on Wasserstein manifold $\mathcal{P}(M)$ that connects ρ_{θ_t} and ρ_t satisfies the geodesic equations (Equation 4.85). If differentiating $W_2^2(\rho_{\theta_t}, \rho_t)$ with respect to time t according to Theorem 23.9 of [7], we are able to deduce

$$\frac{d}{dt} W_2^2(\rho_{\theta_t}, \rho_t) = 2g^W(\dot{\rho}_{\theta_t}, -\dot{\bar{\rho}}_0) + 2g^W(\dot{\rho}_t, \dot{\bar{\rho}}_1), \tag{4.86}$$

in which $\dot{\bar{\rho}}_0 = \partial_\tau \bar{\rho}_\tau|_{\tau=0} = -\nabla \cdot (\bar{\rho}_0 \nabla \psi_0)$, $\dot{\bar{\rho}}_1 = \partial_\tau \bar{\rho}_\tau|_{\tau=1} = -\nabla \cdot (\bar{\rho}_1 \nabla \psi_1)$. Notice that

$$\dot{\rho}_{\theta_t} = (T_{\theta_t^\sharp})_* \dot{\theta}_t \quad \dot{\rho}_t = -\mathrm{grad}_W \mathcal{H}(\rho_t) = \nabla \cdot (\rho_t \nabla (V + D \log \rho_t)).$$

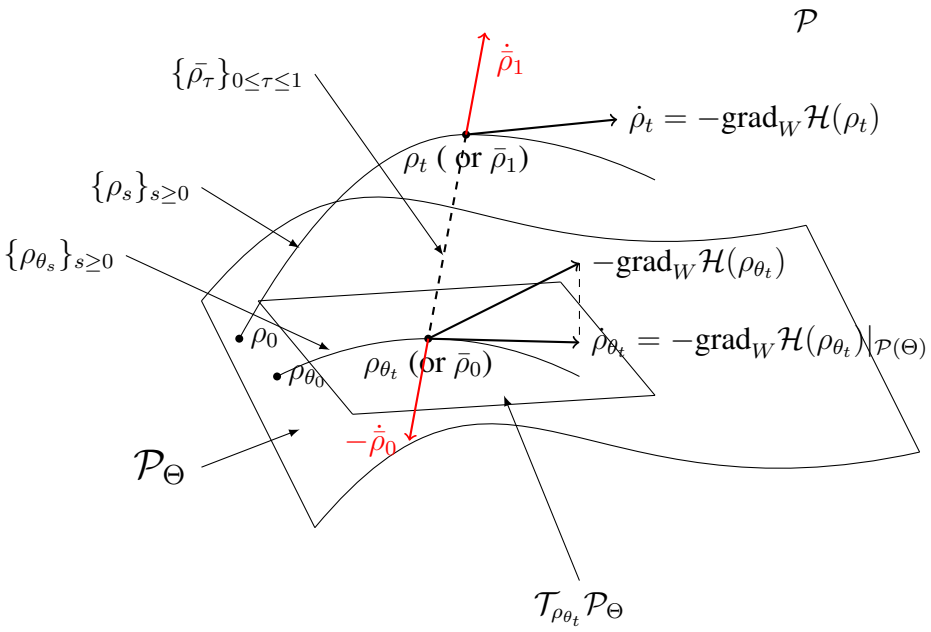


Figure 4.5: An illustrative diagram for the proof of Theorem 4.6.5

Using the definition (Equation 2.58) of Wasserstein metric, we can compute (recall that $\rho_{\theta_t} = \bar{\rho}_0$, $\rho_t = \bar{\rho}_1$):

$$g^W(\dot{\rho}_{\theta_t}, \dot{\bar{\rho}}_0) = \int \nabla(V + D\log\bar{\rho}_0)\cdot\psi_0\,\bar{\rho}_0\,dx \quad g^W(\dot{\rho}_t, \dot{\bar{\rho}}_1) = \int \nabla(V + D\log\bar{\rho}_1)\cdot\psi_1\,\bar{\rho}_1\,dx.$$

Now we can write (Equation 4.86) as,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}W_2^2(\rho_{\theta_t}, \rho_t)\\
=&g^W((T_{\theta_t\sharp})_*\dot{\theta}_t + \mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}), -\dot{\bar{\rho}}_0) + g^W(-\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}), -\dot{\bar{\rho}}_0) + g^W(-\mathrm{grad}_W\mathcal{H}(\rho_t), \dot{\bar{\rho}}_1)\\
=&g^W(\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi,\ -\dot{\bar{\rho}}_0) - (g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_1),\ \dot{\bar{\rho}}_1) - g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_0),\ \dot{\bar{\rho}}_0)),
\end{aligned} \tag{4.87}$$

here we set $\xi = -\dot{\theta}_t$.

For the first term in the last line of (Equation 4.87), we use Cauchy–Schwarz inequality,

(Equation 4.72), and Lemma 4.6.6, which implies $g(\dot{\bar{\rho}}_0, \dot{\bar{\rho}}_0) = W_2^2(\rho_{\theta_t}, \rho_t)$, to obtain

$$\begin{aligned}
&g^W(\text{grad}_W \mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi, -\dot{\bar{\rho}}_0) \\
\leq &\sqrt{g^W(\text{grad}_W \mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi,\ \text{grad}_W \mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi)}\sqrt{g^W(\dot{\bar{\rho}}_0, \dot{\bar{\rho}}_0)} \\
\leq &\sqrt{\delta_0} W(\rho_{\theta_t}, \rho_t). \qquad (4.88)
\end{aligned}$$

For the second term in (Equation 4.87) , we write it as:

$$g^W(\text{grad}_W \mathcal{H}(\bar{\rho}_1), \dot{\bar{\rho}}_1) - g^W(\text{grad}_W \mathcal{H}(\bar{\rho}_0), \dot{\bar{\rho}}_0) = \int_0^1 \frac{d}{d\tau} g^W(\text{grad}_W \mathcal{H}(\bar{\rho}_\tau), \dot{\bar{\rho}}_\tau)\, d\tau. \quad (4.89)$$

By Lemma 4.6.7, we have:

$$g^W(\text{grad}_W \mathcal{H}(\bar{\rho}_1), \dot{\bar{\rho}}_1) - g^W(\text{grad}_W \mathcal{H}(\bar{\rho}_0), \dot{\bar{\rho}}_0) \geq \lambda\, W_2^2(\rho_{\theta_t}, \rho_t). \quad (4.90)$$

Combining inequalities (Equation 4.88), (Equation 4.90) and (Equation 4.87), we get

$$\frac{1}{2}\frac{d}{dt} W_2^2(\rho_{\theta_t}, \rho_t) \leq -\lambda W_2^2(\rho_{\theta_t}, \rho_t) + \sqrt{\delta_0}\, W_2(\rho_{\theta_t}, \rho_t).$$

This is:

$$\frac{d}{dt} W_2(\rho_{\theta_t}, \rho_t) \leq -\lambda W_2(\rho_{\theta_t}, \rho_t) + \sqrt{\delta_0}.$$

When $\lambda \neq 0$, the Grownwall's inequality gives

$$W_2(\rho_{\theta_t}, \rho_t) \leq \frac{\sqrt{\delta_0}}{\lambda}(1 - e^{-\lambda t}) + e^{-\lambda t} W_2(\rho_{\theta_0}, \rho_0).$$

When $\lambda = 0$, the inequality is $\frac{d}{dt} W_2(\rho_{\theta_t}, \rho_t) \leq \sqrt{\delta_0}$, direct integration yields

$$W_2(\rho_{\theta_t}, \rho_t) \leq \sqrt{\delta_0} t + W_2(\rho_{\theta_0}, \rho_0)\ .$$

□

When the potential V is strictly convex, i.e. $\lambda > 0$. (Equation 4.84) in Theorem 4.6.5 provides a nice estimation of the error term $W_2(\rho_{\theta_t}, \rho_t)$ at any time t that is always upper bounded by $\max\{\frac{\sqrt{\delta_0}}{\lambda}, W_2(\rho_{\theta_0}, \rho_0)\}$.

In case that the potential V is not strictly convex, i.e. λ could be 0 or negative, the right hand side in (Equation 4.84) may increase to infinity when time $t \to \infty$. However, (Equation 4.76) and (Equation 4.83) reveals that both ρ_{θ_t} and ρ_t stay in a small neighbourhood of the Gibbs ρ_* when t is large. When taking this into account, we are able to show that the error term $W_2(\rho_{\theta_t}, \rho_t)$ doesn't get arbitrarily large. In the following theorem, we provide a uniform bound for the error depending on t.

Theorem 4.6.8. *Suppose* $\{\rho_t\}_{t\geq 0}$ *solves* (Equation 4.6) *and* $\{\rho_{\theta_t}\}_{t\geq 0}$ *solves* (Equation 4.28)*, the Hessian of the potential* $V \in \mathcal{V}$ *is bounded from below by* λ*, i.e.* $\nabla^2 V \succeq \lambda I$*, then*

$$W_2(\rho_{\theta_t}, \rho_t) \leq \min\left\{\Omega_\lambda(t),\ \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}} + \left(\sqrt{\left|2K_1 - \frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}\right|} + \sqrt{\frac{2K_2}{\tilde{\lambda}_D}}\right) e^{-\frac{\tilde{\lambda}_D}{2}Dt}\right\}, \tag{4.91}$$

where function $\Omega_\lambda(t)$ *is defined in* (Equation 4.84)*,* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$*,* $K_1 = \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)$*, and* $K_2 = \mathcal{D}_{KL}(\rho_0\|\rho_*)$.

Lemma 4.6.9 (Talagrand inequality [159, 7])**.** *If the Gibbs distribution* ρ_* *satisfies the Logarithmic Sobolev inequality* (Equation 4.77) *with constant* $\tilde{\lambda} > 0$*,* ρ_* *also satisfies the Talagrand inequality:*

$$\sqrt{2\frac{\mathcal{D}_{KL}(\rho\|\rho_*)}{\tilde{\lambda}}} \geq W_2(\rho, \rho_*). \quad \textit{for any } \rho \in \mathcal{P}. \tag{4.92}$$

Proof of Theorem 4.6.8. The first term has been proved in Theorem 4.6.5, the second term is just a quick result of Theorem 4.6.1 and Talagrand inequality: for t fixed, (Equation 4.76)

together with Talagrand inequality (Equation 4.92) gives:

$$W_2(\rho_{\theta_t},\rho_*) \le \sqrt{2\frac{\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*)}{\tilde{\lambda}_D}} \le \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}(1-e^{-\tilde{\lambda}_D Dt}) + 2K_1 e^{-\tilde{\lambda}_D Dt}}$$
$$\le \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}} + \sqrt{\left|2K_1 - \frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}\right|} e^{-\frac{\tilde{\lambda}_D}{2}Dt}.$$

Similarly, (Equation 4.83) and (Equation 4.92) gives

$$W_2(\rho_t,\rho_*) \le \sqrt{2\frac{\mathcal{D}_{\mathrm{KL}}(\rho_t\|\rho_*)}{\tilde{\lambda}_D}} \le \sqrt{\frac{2K_2}{\tilde{\lambda}_D}} e^{-\frac{\tilde{\lambda}_D}{2}Dt}.$$

Applying triangle inequality of Wasserstein distance $W_2(\rho_{\theta_t},\rho_t) \le W_2(\rho_{\theta_t},\rho_*)+W_2(\rho_t,\rho_*)$, we get (Equation 4.91). □

Based on Theorem 4.6.8, we can obtain a uniform *a priori* error estimate.

Theorem 4.6.10 (Main Theorem on *a priori* error analysis of the parametric Fokker-Planck equation)**.** *Assume* $E_0 = W_2(\rho_{\theta_0},\rho_0)$ *and* δ_0 *defined in* (Equation 4.72) *are sufficiently small in the sense that*

$$E_0 < A\sqrt{\delta_0} + B,\ \sqrt{\delta_0} + E_0 \le Be^{-\mu_D(A+1)}. \tag{4.93}$$

Then the approximation error $W_2(\rho_{\theta_t},\rho_t)$ *at any time* $t > 0$ *can be uniformly bounded by* E_0 *and* δ_0.

- *When* $\lambda > 0$, $W_2(\rho_{\theta_t},\rho_t) \le \max\{\sqrt{\delta_0}/\lambda, E_0\} \sim O(\sqrt{\delta_0}+E_0)$,
- *When* $\lambda = 0$, $W_2(\rho_{\theta_t},\rho_t) \le \frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0} + E_0 \sim O(\sqrt{\delta_0}\log\frac{1}{\sqrt{\delta_0}+E_0} + E_0)$,
- *When* $\lambda < 0$, $W_2(\rho_{\theta_t},\rho_t) \le A\sqrt{\delta_0}+B^{\frac{|\lambda|}{|\lambda|+\mu_D}}\left(E_0+\sqrt{\delta_0}/|\lambda|\right)^{\frac{\mu_D}{|\lambda|+\mu_D}} \sim O((E_0+\sqrt{\delta_0})^{\frac{\tilde{\lambda}_D D}{2|\lambda|+\tilde{\lambda}_D D}})$.

Here A, B, μ_D *are* $O(1)$ *constants depending on* V, D, ρ_0, θ_0*. Their values are given in* (Equation 4.95).

Proof of Theorem 4.6.10 . When $\lambda > 0$, by (Equation 4.91), we have $E(t) \leq \frac{\sqrt{\delta_0}}{\lambda} + \left(E_0 - \frac{\sqrt{\delta_0}}{\lambda}\right) e^{-\lambda t}$, the right hand side can be bounded by $\max\{E_0, \frac{\sqrt{\delta_0}}{\lambda}\}$.

When $\lambda < 0$, we denote the right hand side of (Equation 4.91) as

$$E(t) = \min\left\{-\frac{1}{|\lambda|}\sqrt{\delta_0} + \left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right) e^{|\lambda| t}, A\sqrt{\delta_0} + Be^{-\mu_D t}\right\}, \tag{4.94}$$

where

$$A = \frac{\sqrt{2}}{\tilde{\lambda}_D D}, \quad B = \sqrt{\left|2K_1 - \frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}\right|} + \sqrt{\frac{2K_2}{\tilde{\lambda}_D}}, \quad \text{and} \quad \mu_D = \frac{\tilde{\lambda}_D D}{2} \tag{4.95}$$

are all positive numbers. The first term in (Equation 4.94) is increasing as a function of time t while the second term is decreasing, combining $E_0 < A\sqrt{\delta_0} + B$, we know $t_0 = \text{argmax}_{t\geq 0} E(t)$ is unique and satisfies

$$-\frac{1}{|\lambda|}\sqrt{\delta_0} + \left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right) e^{|\lambda| t_0} = A\sqrt{\delta_0} + Be^{-\mu_D t_0}, \tag{4.96}$$

as indicated in Figure 4.6.

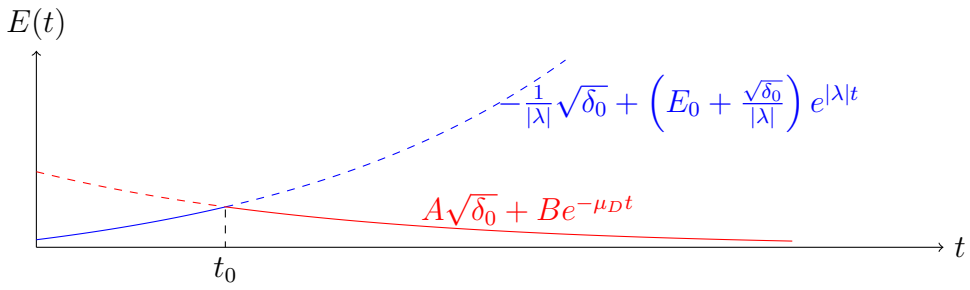


Figure 4.6: An illustrative diagram for the proof of Theorem 4.6.10

Since $A > 0$, (Equation 4.96) leads to $\left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right) e^{|\lambda| t_0} > Be^{-\mu_D t_0}$, thus

$$t_0 > \frac{\log B - \log\left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right)}{|\lambda| + \mu_D}. \tag{4.97}$$

Using (Equation 4.97), we show

$$\max_{t\geq 0} E(t) = E(t_0) = A\sqrt{\delta_0}+B\, e^{-\mu_D t_0} < A\sqrt{\delta_0}+B^{\frac{|\lambda|}{|\lambda|+\mu_D}} \left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right)^{\frac{\mu_D}{|\lambda|+\mu_D}}. \quad (4.98)$$

As a result, $W_2(\rho_{\theta_t}, \rho_t)$ can be uniformly bounded by the right hand side of (Equation 4.98). Since A, B are $O(1)$ coefficients, this uniform bound is dominated by the term

$$O(\left(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\right)^{\frac{\mu_D}{|\lambda|+\mu_D}}) = O((E_0 + \sqrt{\delta_0})^{\frac{\tilde{\lambda}_D D}{2|\lambda|+\tilde{\lambda}_D D}}).$$

At last, when $\lambda = 0$, by (Equation 4.91)

$$E(t) = \min\left\{\sqrt{\delta_0}t + E_0, A\sqrt{\delta_0} + Be^{-\mu_D t}\right\},$$

Let us denote $f(t) = A\sqrt{\delta_0} + Be^{-\mu_D t} - \sqrt{\delta_0}t - E_0$. Similar to the analysis for the case $\lambda < 0$, we denote $t_0 = \text{argmax}_{t\geq 0} E(t)$, then t_0 is unique and solves $f(t_0) = 0$. Since $f(t)$ is decreasing with $f(A+1) > 0$, $t_0 > A + 1$. Then we have

$$\max_{t\geq 0} E(t) = E(t_0) = A\sqrt{\delta_0} + Be^{-\mu_D t_0} = \sqrt{\delta_0}t_0 + E_0 > \sqrt{\delta_0}(A+1) + E_0$$

This leads to $Be^{-\mu_D t_0} > \sqrt{\delta_0} + E_0$, i.e. $t_0 < \frac{1}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0}$. Thus we have

$$\max_{t\geq 0} E(t) = E(t_0) = \sqrt{\delta_0}t_0 + E_0 < \frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0} + E_0.$$

Therefore $W_2(\rho_{\theta_t}, \rho_t)$ can be uniformly bounded by the term $\frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0} + E_0 \sim O(\sqrt{\delta_0}\log\frac{1}{\sqrt{\delta_0}+E_0} + E_0)$.

□

Remark 20. *In the case that $V \in \mathcal{V}$ is not convex, we can decompose V by $V = U + \phi$ with $\nabla^2 U \succeq KI$ ($K > 0$) and $\nabla^2\phi \succeq K_\phi I$. We can still assume $\nabla^2 V \succeq \lambda I$, but λ may be*

negative. One can verify that $K_\phi < 0$ *and* $|K_\phi| - K \geq |\lambda|$*. On the other hand, one can compute* $\tilde{\lambda}_D = \frac{K}{D} e^{-\frac{osc(\phi)}{D}}$*. Combining them together, we provide a lower bound for* α*:*

$$\alpha \geq \gamma(D, U, \phi) = \frac{1}{1 + 2\left(\frac{|K_\phi|}{K} - 1\right) e^{\frac{osc(\phi)}{D}}}$$

One can verify that increasing the diffusion coefficient D *or convexity* K*, or decreasing the oscillation* $osc(\phi)$ *and convexity* K_ϕ *can improve the lower bound* $\gamma(D, U, \phi)$ *for order* α*.*

Similarly, one can establish the corresponding *posterior* error estimate for $W_2(\rho_{\theta_t}, \rho_t)$:

Theorem 4.6.11 (*Posterior* error analysis of the parametric Fokker-Planck equation)**.** *Suppose* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$ *and* δ_1*defined in* (Equation 4.75) *satisfy the condition* (Equation 4.93) *with* δ_0 *replaced by* δ_1*. Then*

1. *When* $\lambda \geq 0$*,* $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O(E_0 + \sqrt{\delta_1})$*;*
2. *When* $\lambda = 0$*,* $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O(\sqrt{\delta_1} \log \frac{1}{\sqrt{\delta_1} + E_0} + E_0)$*;*
3. *When* $\lambda < 0$*,* $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O((E_0 + \sqrt{\delta_1})^{\frac{\tilde{\lambda}_D D}{2|\lambda| + \tilde{\lambda}_D D}})$*.*

Wasserstein error for the time discrete schemes

To solve (Equation 4.28) numerically, we need time discrete schemes, such as the one proposed in (Equation 4.48). In this subsection, we present the error estimate in Wasserstein distance for our scheme. We begin our analysis by focusing on the forward Euler scheme, meaning that we apply forward Euler scheme to solve (Equation 4.28) and compute θ_k at each time step. We denote $\rho_{\theta_k} = T_{\theta_k \sharp} p$. We estimate the W_2-error between ρ_{θ_k} and the real solution ρ_{t_k}. Then we analyze the W_2 distance between the solutions obtained by forward Euler scheme and our scheme (Equation 4.48) respectively, which in turn give us the W_2 error estimate for our scheme.

Theorem 4.6.12 (*a priori* error analysis of forward Euler scheme)**.** *Let* θ_k *(*$k = 0, 1, \ldots, N$*) be the solution of forward Euler scheme applied to* (Equation 4.28) *at time* $t_k = kh$ *on* $[0, T]$

with time step size $h = \frac{T}{N}$, $\rho_{\theta_k} = T_{\theta_k \sharp} p$, *and* $\{\rho_t\}_{t\geq 0}$ *solves the Fokker-Planck Equation* (Equation 4.6) *exactly. Assume that the Hessian of the potential function* $V \in \mathcal{C}^2(\mathbb{R}^d)$ *can be bounded from above and below, i.e.* $\lambda I \preceq \nabla^2 V \preceq \Lambda I$. *Then*

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \leq (\sqrt{\delta_0}h + Ch^2)\frac{1 - e^{-\lambda t_k}}{1 - e^{-\lambda h}} + e^{-\lambda t_k} W_2(\rho_{\theta_0}, \rho_0) \quad \textit{for any} \quad t_k = kh, \tag{4.99}$$

for all $0 \leq k \leq N$. *Here* C *is a constant whose formula is provided in* (Equation 4.116).

In order to estimate $W_2(\rho_{\theta_k}, \rho_{t_k})$, we use the triangle inequality of W_2 distance [7] to separate it into three parts:

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \leq W_2(\rho_{\theta_k}, \tilde{\rho}^\star_{t_k}) + W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) + W_2(\tilde{\rho}_{t_k}, \rho_{t_k}). \tag{4.100}$$

Here $\{\tilde{\rho}_t\}_{t_{k-1}\leq t\leq t_k}$ satisfies:

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \nabla \cdot (\tilde{\rho}_t \nabla V) + D\Delta \tilde{\rho}_t \ , \quad \tilde{\rho}_{t_{k-1}} = \rho_{\theta_{k-1}}, \tag{4.101}$$

and $\{\rho^\star_t\}_{t\geq t_{k-1}}$ satisfies:

$$\frac{\partial \rho^\star_t}{\partial t} = \nabla \cdot (\rho^\star_t \nabla (V + D \log \rho_{\theta_{k-1}})) \ , \quad \rho^\star_{t_{k-1}} = \rho_{\theta_{k-1}}. \tag{4.102}$$

Figure 4.7 shows the relations of different items used in our proof. We present three lemmas that estimate three terms in (Equation 4.100) respectively.

Lemma 4.6.13. $W_2(\rho_{\theta_k}, \rho^\star_{t_k})$ *in (Equation 4.100) can be upper bounded by* $\sqrt{\delta_0}h + O(h^2)$.

An explicit formula for the coefficient of h^2 is included in the following proof.

Proof. We establish the desired estimation by introducing several different pushforward maps as shown in Figure 4.8 and then applying triangle inequality.

(a) We know $\rho_{\theta_{k-1}} = T_{\theta_{k-1}\sharp} p$ and $\rho_{\theta_k} = T_{\theta_k \sharp} p$, let us denote $T_{t_{k-1}\to t_k} = T_{\theta_k} \circ T^{-1}_{\theta_{k-1}}$. Then $\rho_{\theta_k} = T_{t_{k-1}\to t_k \sharp} \rho_{\theta_{k-1}}$.

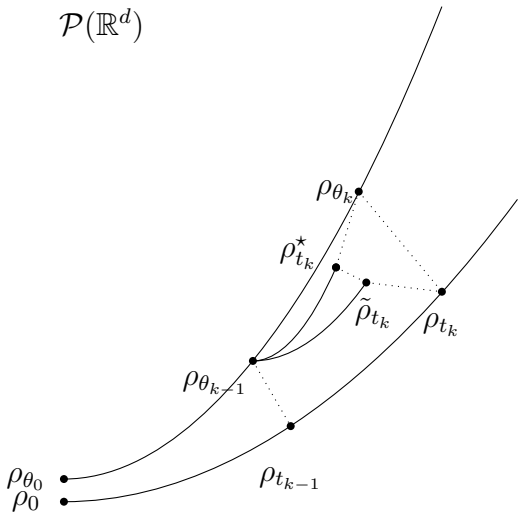


Figure 4.7: Trajectory of $\{\rho_{\theta_k}\}_{k=0,\dots,N}$ is our numerical solution; trajectory of $\{\rho_t\}_{t\geq 0}$ is the real solution of the Fokker-Planck Equation; $\{\tilde{\rho}_t\}_{t\geq t_{k-1}}$ solves (Equation 4.101); $\{\rho^\star_t\}_{t\geq t_{k-1}}$ solves (Equation 4.102).

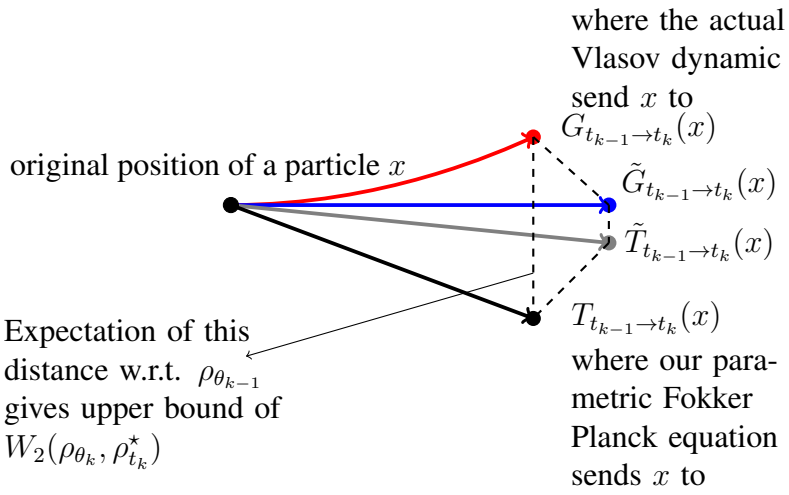


Figure 4.8: Illustration of proof strategy for Lemma 4.6.13

(b) Let $\xi_{k-1} = \dot{\theta}_{k-1} = -G(\theta_{k-1})^{-1}\nabla_\theta H(\theta_{k-1})$ and by convention, we denote $\mathbf{\Psi}$ as solution of (Equation 4.14). We consider the map $\tilde{T}_{t_{k-1}\to t_k}(\cdot) = \mathrm{Id} + h\nabla\mathbf{\Psi}(\cdot)^{\mathrm{T}}\xi_{k-1}$.

(c) We denote $\zeta_\theta(\cdot) = V(\cdot) + D\log\rho_\theta(\cdot)$. The particle version (recall (Equation 4.7)) of (Equation 4.102) is:

$$\dot{z}_t = -\nabla\zeta_{\theta_{k-1}}(z_t) \quad 0 \le t \le h \quad \text{with initial condition } z_0 = x \sim \rho_{\theta_{k-1}}. \tag{4.103}$$

we denote the solution map of (Equation 4.103) by $G_{t_{k-1}\to t_k}(x) = z_{t_k}$. Then $\rho^\star_{t_k} = G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$.

(d) The map $G_{t_{k-1}\to t_k}$ is obtained by solving an ODE, in order to compare the difference with $T_{t_{k-1}\to t_k}$, we consider the ODE with fixed initial vector field:

$$\dot{\tilde{z}}_t = -\nabla\zeta_{\theta_{k-1}}(x) \quad 0 \le t \le h \quad \tilde{z}_0 = x \sim \rho_{\theta_{k-1}}. \tag{4.104}$$

This ODE will induce the solution map $\tilde{G}_{t_{k-1}\to t_k}(\cdot) = \mathrm{Id} - h\nabla\zeta_{\theta_{k-1}}(\cdot)$.

With the maps defined in (a),(b),(c),(d), and using the triangle inequality of W_2 distance, we have,

$$\begin{aligned} W_2(\rho_{\theta_k}, \tilde{\rho}^\star_{t_k}) &= W_2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}) \\ &\le \underbrace{W_2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(A)}} + \underbrace{W_2(\tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(B)}} \\ &\quad + \underbrace{W_2(\tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(C)}}. \end{aligned}$$

In the rest of the proof, We give upper bounds for distances (A),(B) and (C) respectively.

(A) Let us define $\xi(\theta) = -G(\theta)^{-1}\nabla H(\theta)$. Now we set $\theta(\tau) = \theta_{k-1} + \frac{\tau}{h}(\theta_k - \theta_{k-1}) = \theta_{k-1} + \tau\xi(\theta_{k-1})$. For any x, consider $x_\tau = T_{\theta(\tau)}(T^{-1}_{\theta_{k-1}}(x))$ with $0 \le \tau \le h$, then

$\{x_\tau\}_{0\le\tau\le h}$ satisfies

$$\dot{x}_\tau = \partial_\theta T_{\theta(\tau)}(T^{-1}_{\theta(\tau)}(x_\tau))\xi(\theta_{k-1}) \quad 0 \le \tau \le h. \tag{4.105}$$

If $x_0 \sim \rho_{\theta_{k-1}}$ in (Equation 4.105), it is clear that $x_h \sim T_{t_{k-1}\to t_k \sharp}\rho_{\theta_{k-1}}$. Furthermore, we denote the distribution of x_τ as ρ_τ and $\{\psi_\tau\}$ satisfying

$$-\nabla\cdot(\rho_\tau(x)\partial_\theta T_{\theta(\tau)}(T^{-1}_{\theta(\tau)}(x))\xi_{k-1}) = -\nabla\cdot(\rho_\tau(x)\nabla\psi_\tau(x)) \quad 0\le\tau\le h. \tag{4.106}$$

If we consider

$$\dot{y}_\tau = \nabla\psi_\tau(y_\tau) \quad 0\le\tau\le h \quad \text{with } y_0 \sim \rho_{\theta_{k-1}},$$

and denote ϱ_τ as the distribution of y_τ, by continuity equation and (Equation 4.106), we know $\rho_\tau = \varrho_\tau$ for $0\le\tau\le h$, thus $y_h \sim T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$. On the other hand, when $\tau = 0$, (Equation 4.106) shows $\nabla\psi_0(x) = \nabla\mathbf{\Psi}(x)^{\mathrm{T}}\xi_{k-1}$. Combining them together, we bound term (A) as

$$\begin{aligned}
&W_2^2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}) \\
\le&\mathbb{E}_{y_0\sim\rho_{\theta_{k-1}}}|y_h - (y_0 + h\nabla\psi_0(y_0))|^2 = \mathbb{E}_{y_0\sim\rho_{\theta_{k-1}}}\left|\int_0^h(\nabla\psi_\tau(y_\tau) - \nabla\psi_0(y_0))\,d\tau\right|^2 \\
=&\mathbb{E}_{y_0}\left|\int_0^h\int_0^\tau\frac{d}{ds}(\nabla\psi_s(y_s))\,ds\,d\tau\right|^2 = \mathbb{E}_{y_0}\left|\int_0^h\int_s^h\frac{d}{ds}(\nabla\psi_s(y_s))\,d\tau\,ds\right|^2 \\
=&\mathbb{E}_{y_0}\left|\int_0^h(h-s)\frac{d}{ds}(\nabla\psi_s(y_s))\,ds\right|^2 \le \mathbb{E}_{y_0}\int_0^h(h-s)^2\,ds\int_0^h\left|\frac{d}{ds}(\nabla\psi_s(y_s))\right|^2\,ds \\
=&\frac{h^3}{3}\int_0^h\mathbb{E}_{y_0}\left|\frac{d}{ds}(\nabla\psi_s(y_s))\right|^2\,ds \\
=&\frac{h^4}{3}\left(\frac{1}{h}\int_0^h\mathbb{E}_{y_s}\left|\frac{\partial\nabla\psi_s(y_s)}{\partial t} + \nabla^2\psi_s(y_s)\nabla\psi_s(y_s)\right|^2 ds\right).
\end{aligned}$$

Notice that y_s follows the distribution

$$\rho_s = (T_{\theta_{k-1}+s\xi(\theta_{k-1})} \circ T_{\theta_{k-1}}^{-1})_\sharp \rho_{\theta_{k-1}} = T_{\theta_{k-1}+s\xi(\theta_{k-1})\sharp} p.$$

If we define

$$\mathfrak{M}(\theta, s) = \int \left| \frac{\partial}{\partial t} \nabla \psi_s(T_{\theta(s)}(z)) + \nabla^2 \psi_s(T_{\theta(s)}(z)) \nabla \psi_s(T_{\theta(s)}(z)) \right|^2 p(z)\, dz \quad (4.107)$$

with $-\nabla \cdot (\rho_s \nabla \psi_s) = -\nabla \cdot (\rho_s\, \partial_\theta T_{\theta(s)} \circ T_{\theta(s)}^{-1}\, \xi(\theta)),\ \rho_s = T_{\theta+s\xi(\theta)\sharp} p;$

and $\theta(s) = \theta + s\xi(\theta)$.

we are able to derive

$$W_2^2(T_{t_{k-1}\to t_k \sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k \sharp}\rho_{\theta_{k-1}}) \le \frac{1}{3} \sup_{0\le s\le h} \mathfrak{M}(\theta_{k-1}, s) h^4. \quad (4.108)$$

(B) We have

$$\begin{aligned} &W_2^2(\tilde{T}_{t_{k-1}\to t_k \sharp}\rho_{\theta_{k-1}}, \tilde{G}_{t_{k-1}\to t_k \sharp}\rho_{\theta_{k-1}}) \le \int |\tilde{T}_{t_{k-1}\to t_k}(x) - \tilde{G}_{t_{k-1}\to t_k}(x)|^2 \rho_{\theta_{k-1}}(x)\, dx \\ &= h^2 \left(\int |\nabla \mathbf{\Psi}(x)^{\mathrm{T}} \xi(\theta_{k-1}) - (-\nabla \zeta_{\theta_{k-1}}(x))|^2 \rho_{\theta_{k-1}}(x)\, dx \right) \\ &= h^2 \left(\int |\nabla \mathbf{\Psi}(T_{\theta_{k-1}}(x))^{\mathrm{T}} \xi(\theta_{k-1}) - (-\nabla (V + D \log \rho_{\theta_{k-1}}) \circ T_{\theta_{k-1}}(x))|^2\, dp(x) \right) \\ &\le \delta_0\, h^2. \end{aligned}$$

The last inequality is due to Theorem 4.3.5 and definition (Equation 4.72).

(C) Recall that $\{z_t\}$ and $\{\tilde{z}_t\}$ solve (Equation 4.103) and (Equation 4.104) with initial condition $z_0 = \tilde{z}_0 = x$ respectively, similar to the analysis in (A), we can estimate

term (C) as

$$
\begin{aligned}
&W_2^2(\tilde{G}_{t_{k-1}\to t_k\sharp\rho_{\theta_{k-1}}}, G_{t_{k-1}\to t_k\sharp\rho_{\theta_{k-1}}}) \\
\leq&\mathbb{E}_{x\sim\rho_{\theta_{k-1}}}|z_h-\tilde{z}_h|^2 = \mathbb{E}_{x\sim\rho_{\theta_{k-1}}}\left|\int_0^h \nabla\zeta_{k-1}(x)-\nabla\zeta_{k-1}(z_\tau)\,d\tau\right|^2 \\
=&\mathbb{E}_x\left|\int_0^h\int_0^\tau \frac{d}{ds}\nabla\zeta_{\theta_{k-1}}(z_s)\,ds\,d\tau\right|^2 = \mathbb{E}_x\left|\int_0^h (h-s)\frac{d}{ds}\nabla\zeta_{\theta_{k-1}}(z_s)\,ds\right|^2 \\
\leq&\mathbb{E}_x\frac{h^3}{3}\int_0^h\left|\frac{d}{ds}\nabla\zeta_{\theta_{k-1}}(z_s)\right|^2 ds = \frac{h^4}{3}\left(\frac{1}{h}\int_0^h \mathbb{E}_{z_s}\left|\nabla^2\zeta_{\theta_{k-1}}(z_s)\zeta_{\theta_{k-1}}(z_s)\right|^2\,ds\right).
\end{aligned}
$$

We define

$$
\begin{aligned}
&\mathfrak{N}(\theta,s) = \mathbb{E}_{z_s}\left|\nabla^2\zeta_\theta(z_s)\zeta_\theta(z_s)\right|^2, \quad \text{with} \quad \zeta_\theta(\cdot) = V(\cdot)+D\log\rho_\theta(\cdot), \\
&\dot{z}_t = -\nabla\zeta_\theta(z_t), \;\; z_0\sim\rho_\theta.
\end{aligned}
$$

Similar to (A), we have:

$$
W_2^2(\tilde{G}_{t_{k-1}\to t_k\sharp\rho_{\theta_{k-1}}}, G_{t_{k-1}\to t_k\sharp\rho_{\theta_{k-1}}}) \leq \frac{1}{3}\sup_{0\leq s\leq h}\mathfrak{N}(\theta_{k-1},h)h^4
$$

Combining the estimates for terms (A),(B) and (C), and defining

$$
M(\theta,h) = \sup_{0\leq s\leq h}\mathfrak{M}(\theta_{k-1},s), \quad N(\theta,h) = \sup_{0\leq s\leq h}\mathfrak{N}(\theta_{k-1},s), \tag{4.109}
$$

we obtain

$$
W_2(\rho_{\theta_k},\tilde{\rho}^\star_{t_k}) \leq \sqrt{\delta_0}h + \frac{M(\theta_{k-1},h)+N(\theta_{k-1},h)}{\sqrt{3}}\,h^2.
$$

□

Lemma 4.6.14. *The second term in (Equation 4.100) can be upper bounded by $O(h^2)$.*

Proof. Recall that $\tilde{\rho}_t$ is defined by (Equation 4.101) and ρ_t^* is defined by (Equation 4.102).

We can rewrite (Equation 4.102) as:

$$\frac{\partial \rho_t^\star}{\partial t} = \nabla \cdot (\rho_t^\star (\nabla V + D\nabla \log \rho_{\theta_{k-1}} - D\nabla \log \rho_t^\star)) + D\Delta \rho_t^\star, \quad t_{k-1} \le t \le t_k.$$

We consider the following Stochastic Differential Equations (SDEs) sharing the same trajectory of Brownian motion $\{\boldsymbol{B}_\tau\}_{0\le\tau\le h}$ and initial condition:

$$dx_\tau = -\nabla V(x_\tau)d\tau + \sqrt{2D}\, d\boldsymbol{B}_\tau \tag{4.110}$$

$$dx_\tau^\star = -\nabla V(x_\tau^\star)d\tau + (D\nabla \log \rho^\star_{t_{k-1}+\tau}(x_\tau^\star) - D\nabla \log \rho_{\theta_{k-1}}(x_\tau^\star))d\tau + \sqrt{2D}\, d\boldsymbol{B}_\tau \tag{4.111}$$

with initial condition: $x_0 = x_0^\star \sim \rho_{\theta_{k-1}}$ and $0 \le \tau \le h$.

Subtracting (Equation 4.110) from (Equation 4.111), we get:

$$x_\tau^\star - x_\tau = \int_0^\tau \nabla V(x_s) - \nabla V(x_s^\star) + \vec{r}(x_s^\star, s)\, ds,$$

in which we denote $\vec{r}(x,\tau) = D\nabla \log \rho^\star_{t_{k-1}+\tau}(x) - D\nabla \log \rho_{\theta_{k-1}}(x)$ for convenience. Hence,

$$\begin{aligned}
\mathbb{E}|x_\tau^\star - x_\tau|^2 &= \mathbb{E}\left|\int_0^\tau \nabla V(x_s) - \nabla V(x_s^\star) + \vec{r}(x_s^\star, s)\, ds\right|^2 \\
&\le 2\,\mathbb{E}\left|\int_0^\tau \nabla V(x_s) - \nabla V(x_s^\star)\, ds\right|^2 + 2\,\mathbb{E}\left|\int_0^\tau \vec{r}(x_s^\star, s)\, ds\right|^2 \\
&\le 2\,\mathbb{E}\left[\tau \int_0^\tau |\nabla V(x_s) - \nabla V(x_s^\star)|^2\, ds\right] + 2\,\mathbb{E}\left[\tau \int_0^\tau |\vec{r}(x_s^\star, s)|^2\, ds\right] \\
&= 2\tau\left(\int_0^\tau \mathbb{E}|\nabla V(x_s) - \nabla V(x_s^\star)|^2 + \mathbb{E}|\vec{r}(x_s^\star, s)|^2\, ds\right)
\end{aligned}$$

Since Hessian of V is bounded from above by Λ, $|\nabla V(x) - \nabla V(y)| \le \Lambda |x - y|$ for any

$x, y \in \mathbb{R}^d$, we have the inequality:

$$\mathbb{E}|x_\tau^\star - x_\tau|^2 \leq 2\tau\Lambda^2 \int_0^\tau \mathbb{E}|x_s^\star - x_s|^2 \, ds + 2\tau \int_0^\tau \mathbb{E}|\vec{r}(x_s^\star, s)|^2 \, ds \tag{4.112}$$

If we define $U_\tau = \int_0^\tau \mathbb{E}|x_s^\star - x_s|^2 \, ds$ and $R_\tau = \int_0^\tau \mathbb{E}|\vec{r}(x_s^\star, s)|^2 \, ds$, (Equation 4.112) becomes:

$$U_\tau' \leq 2\Lambda^2 \tau U_\tau + 2\tau R_\tau$$

By integrating this inequality, we have

$$\begin{aligned} U_\tau &\leq \int_0^\tau 2e^{\Lambda(\tau^2 - s^2)} s R_s \, ds, \\ \text{and} \quad U_\tau' &\leq 4\Lambda^2 \tau \int_0^\tau e^{\Lambda(\tau^2 - s^2)} s R_s ds + 2\tau R_\tau. \end{aligned}$$

Therefore

$$W_2(\rho_{t_k}^\star, \tilde{\rho}_{t_k}) \leq \sqrt{\mathbb{E}|x_h^\star - x_h|^2} = \sqrt{U_h'} \leq \sqrt{4\Lambda^2 h \int_0^h e^{\Lambda(h^2 - s^2)} s R_s \, ds + 2h R_h}$$

Since R_τ is increasing with respect to τ, we are able to estimate

$$W_2(\rho_{t_k}^\star, \tilde{\rho}_{t_k}) \leq \sqrt{4\Lambda^2 h^2 \int_0^h e^{\Lambda(h^2 - s^2)} s ds + 2h} \, \sqrt{R_h} = \sqrt{2\Lambda(e^{\Lambda h^2} - 1)h + 2h} \sqrt{R_h}. \tag{4.113}$$

Next we estimate R_h. Recall $\rho_{t_{k-1}}^* = \rho_{\theta_{k-1}}$ as in (Equation 4.102), we have

$$\begin{aligned} R_h =& \int_0^h \mathbb{E}_{x_s^\star} |D \log \rho_{t_{k-1}+s}^\star(x_s^\star) - D \log \rho_{t_{k-1}}^\star(x_s^\star)|^2 ds \\ =& D^2 \int_0^h \mathbb{E}_{x_s^\star} \left| \int_0^s \frac{\partial}{\partial t} \nabla \log \rho_{t_{k-1}+t}^\star(x_s^\star) dt \right|^2 ds \\ \leq& D^2 \int_0^h \mathbb{E}_{x_s^\star} \left[s \int_0^s \left| \frac{\partial}{\partial t} \nabla \log \rho_{t_{k-1}+t}^\star(x_s^\star) \right|^2 dt \right] ds \\ =& D^2 \int_0^h \int_0^s s \int \left| \frac{\partial}{\partial t} \nabla \log \rho_{t_{k-1}+t}^\star \right|^2 \rho_{t_{k-1}+s}^\star dx dt \, ds. \end{aligned}$$

By (Equation 4.102), one can further compute

$$\frac{\partial}{\partial t}\log \rho^{\star}_{t_{k-1}+t} = -\nabla \log \rho^{\star}_{t_{k-1}+t}\cdot \nabla \zeta_{\theta_{k-1}} - \Delta \zeta_{\theta_{k-1}}.$$

Let us define

$$\mathfrak{L}(\theta,t,s) = \int |\nabla(\nabla \log \rho_t \cdot \nabla\zeta_\theta + \Delta\zeta_\theta)|^2 \rho_s \, dx \quad \text{with } \zeta_\theta = V + D\log\rho_\theta$$
$$\text{and } \frac{\partial \rho_s}{\partial s} + \nabla\cdot(\rho_s \nabla\zeta_\theta) = 0 \quad \rho_0 = \rho_\theta$$

Then we have the estimation

$$R_h \le D^2 \int_0^h \int_0^s s\cdot \left(\sup_{0\le t\le s\le h} \mathfrak{L}(\theta_{k-1},t,s)\right) dt\, ds = \frac{D^2}{3}\sup_{0\le t\le s\le h}\mathfrak{L}(\theta_{k-1},t,s)\, h^3.$$

Let us also define

$$L(\theta,h) = \left(\sup_{0\le t\le s\le h}\mathfrak{L}(\theta,t,s)\right)^{\frac{1}{2}} \tag{4.114}$$

Thus (Equation 4.113) becomes $W_2(\rho^{\star}_{t_k},\tilde{\rho}_{t_k}) \le \sqrt{\frac{2D^2}{3}(\Lambda(e^{\Lambda h^2}-1)+2)}L(\theta_{k-1},h)\ h^2$. When the stepsize h is small enough, we have $e^{\Lambda h^2} < 2$. Let us denote $K(D,\Lambda) = \sqrt{\frac{2D^2}{3}(\Lambda+2)}$. Thus we have $W_2(\rho^{\star}_{t_k},\tilde{\rho}_{t_k}) \le K(D,\Lambda)L(\theta_{k-1},h)\ h^2$. □

Remark 21. *Analyzing the discrepancy of stochastic particles under different movements provides a natural upper bound for W_2 distance. Both Lemma 4.6.13 and Lemma 4.6.14 are derived by making use of the particle version of their corresponding density evolution. Such proving strategy was motivated from subsection 4.3.3.*

Lemma 4.6.15. *The third term in (Equation 4.100) satisfies*

$$W_2(\rho_{t_k},\tilde{\rho}_{t_k}) \le e^{-\lambda h} W_2(\rho_{t_{k-1}},\rho_{\theta_{k-1}}).$$

Here we recall that λ satisfies $\nabla^2 V \succeq \lambda I$.

This lemma is a direct corollary of the following theorem:

Theorem 4.6.16. *Suppose the potential $V \in C^2(\mathbb{R}^d)$ satisfying $\nabla^2 V \succeq \lambda I$ for a finite real number λ, i.e. the matrix $\nabla^2 V(x) - \lambda I$ is semi-positive definite for any $x \in \mathbb{R}^d$. Given $\rho_1, \rho_2 \in \mathcal{P}$, and denote $\rho_t^{(1)}$ and $\rho_t^{(2)}$ the solutions of the Fokker-Planck equation with different initial distributions ρ_1 and ρ_2 respectively, i.e.*

$$\begin{aligned}
\frac{\partial \rho_t^{(1)}}{\partial t} &= \nabla \cdot (\rho_t^{(1)} \nabla V) + D\Delta \rho_t^{(1)} \quad \rho_0^{(1)} = \rho_1, \\
\frac{\partial \rho_t^{(2)}}{\partial t} &= \nabla \cdot (\rho_t^{(2)} \nabla V) + D\Delta \rho_t^{(2)} \quad \rho_0^{(2)} = \rho_2.
\end{aligned}$$

Then

$$W_2(\rho_t^{(1)}, \rho_t^{(2)}) \leq e^{-\lambda t} W_2(\rho_1, \rho_2) \tag{4.115}$$

This is a known stability result on Wasserstein gradient flows. One can find its proof in [35] or [7]. With the results in Lemma 4.6.13, Lemma 4.6.14, Lemma 4.6.15, we are ready to prove Theorem 4.6.12.

Proof. (Proof of Theorem 4.6.12) For convenience, we write

$$\text{Err}_k = W_2(\rho_{\theta_k}, \rho_{t_k}) \quad k = 0, 1, ..., N.$$

Combining Lemma 4.6.13, Lemma 4.6.14 and Lemma 4.6.15, the triangle inequality in (Equation 4.100) becomes

$$\text{Err}_k \leq \sqrt{\delta_0}\, h + \left(\frac{1}{\sqrt{3}} M(\theta_{k-1}, h) + \frac{1}{\sqrt{3}} N(\theta_{k-1}, h) + K(D, \Lambda) L(\theta_{k-1}, h) \right) h^2 + e^{-\lambda h}\, \text{Err}_{k-1}.$$

Let us denote the constant C depending on initial parameter θ_0, time stepsize h and time

steps N:

$$C(\theta_0, h, N) = \max_{0 \le k \le N-1} \left\{ \frac{1}{\sqrt{3}} M(\theta_{k-1}, h) + \frac{1}{\sqrt{3}} N(\theta_{k-1}, h) + K(D, \Lambda) L(\theta_{k-1}, h) \right\}. \tag{4.116}$$

In the following discussion, we denote $C = C(\theta_0, h, N)$. By (Equation 4.116), We have:

$$\text{Err}_k \le \sqrt{\delta_0} h + Ch^2 + e^{-\lambda h} \text{Err}_{k-1} \tag{4.117}$$

Multiplying $e^{\lambda k h}$ to both sides of (Equation 4.117), we get:

$$e^{\lambda k h} \text{Err}_k \le (\sqrt{\delta_0}\, h + Ch^2) e^{\lambda k h} + e^{\lambda (k-1) h} \text{Err}_{k-1}. \tag{4.118}$$

For any n, $1 \le n \le N$, summing (Equation 4.118) from 1 to n, we reach

$$e^{\lambda n h} \text{Err}_n \le (\sqrt{\delta_0} h + Ch^2) \left(\sum_{k=1}^{n} e^{\lambda k h} \right) + \text{Err}_0 = (\sqrt{\delta_0} h + Ch^2) \frac{e^{\lambda (n+1) h} - e^{\lambda h}}{e^{\lambda h} - 1} + \text{Err}_0.$$

Recall that $t_n = nh$ for $1 \le n \le N$, it leads to:

$$\text{Err}_n \le (\sqrt{\delta_0} h + Ch^2) \frac{1 - e^{-\lambda t_n}}{1 - e^{-\lambda h}} + e^{-\lambda t_n} \text{Err}_0 \quad n = 1, ..., N.$$

□

Theorem 4.6.12 indicates that the error $W_2(\rho_{\theta_k}, \rho_{t_k})$ is upper bounded by $O(\sqrt{\delta_0}) + O(Ch) + O(W_2(\rho_{\theta_0}, \rho_0))$. Here $O(\sqrt{\delta_0})$ is the essential error term that originates from the approximation mechanism of our parametric Fokker-Planck equation. The $O(Ch)$ error term is induced by the finite difference scheme. And the $O(W_2(\rho_{\theta_0}, \rho_0))$ term is the initial error.

It is worth mentioning that the error bound for forward Euler scheme in (Equation 4.99) matches the error bound for the continuous scheme (Equation 4.84) as we reduce the effects

introduced by finite difference. More precisely, under assumption $\lim_{h\to 0} C(\theta_0, h, N)h = 0$, we have:

$$\lim_{h\to 0}(\sqrt{\delta_0}h + Ch^2)\frac{1-e^{-\lambda t}}{1-e^{-\lambda h}} + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0) = \frac{\sqrt{\delta_0}}{\lambda}(1-e^{-\lambda t}) + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0),$$

this indicates that the error bounds (Equation 4.99) and (Equation 4.84) are compatible when $h \to 0$.

Remark 22 ($O(h)$ error order)**.** *Under further assumptions that $\Theta = \mathbb{R}^m$, $T_\theta(x) \in C^3(\Theta \times \mathbb{R}^d)$ and*

$$\lim_{\theta\to\infty} H(\theta) = +\infty \tag{4.119}$$

we can show the finite difference error term $O(Ch)$ is of order $O(h)$. In fact, the solution obtained from forward Euler scheme is always restricted in a fixed bounded region of Θ. To be more precise, suppose the initial value is θ_0, we consider $\Theta_0 = \{\theta | H(\theta) \leq H(\theta_0)\}$. By (Equation 4.119)*, one can verify Θ_0 is bounded and closed set and thus is compact. We set $l = \max_{\theta\in\Theta_0} |G(\theta)^{-1}\nabla_\theta H(\theta)|$. Then we consider a slightly larger set $\Theta_0^l = \{\theta \mid$ there exists $\theta' \in \Theta_0$, s.t. $|\theta - \theta'| \leq l\}$. Notice that Θ_0^l is also bounded. We define*

$$\sigma_{min}^G = \min_{\theta\in\Theta_0^l} \sigma_{min}(G(\theta)) \quad \sigma_{max}^H = \max_{\theta\in\Theta_0^l} \sigma_{max}(\nabla_{\theta\theta}^2 H(\theta)).$$

Here $\sigma_{max}(A), \sigma_{min}(A)$ denotes the maximum and the minimum singular values of matrix A. We can show that for any time step size $h < \min\{\frac{2\sigma_{min}^G}{\sigma_{max}^H}, 1\}$, the numerical solution $\{\theta_k\}_{k=1}^N$ obtained by applying forward-Euler scheme to (Equation 4.28) *is included in Θ_0. To prove this, we first show $\theta_1 \in \Theta_0$, we consider*

$$\begin{aligned} H(\theta_1) = H(\theta_0 - hG(\theta_0)^{-1}\nabla_\theta H(\theta_0)) =& H(\theta_0) - h\xi^T G(\theta_0)\xi + \frac{h^2}{2}\xi^T\nabla_{\theta\theta}^2 H(\tilde{\theta})\xi \\ \leq& H(\theta_0) - h\sigma_{min}^G|\xi|^2 + \frac{h^2}{2}\sigma_{max}^H|\xi|^2 \leq H(\theta_0) \end{aligned}$$

Here we denote $\xi = G(\theta_0)^{-1}\nabla_\theta H(\theta_0)$. *The second equality is due to* $T_\theta(x) \in C^3(\Theta \times \mathbb{R}^d)$ *and thus* $H(\cdot) \in C^2(\Theta)$. *We notice that* $\tilde{\theta} = \theta_0 + \tau(hG(\theta_0)^{-1}\nabla_\theta H(\theta_0))$ *with* $0 \le \tau \le 1$ *and thus* $\tilde{\theta} \in \Theta_0^l$. *Since* $H(\theta_1) \le H(\theta_0)$, *we know* $\theta_1 \in \Theta_0$. *Applying a similar argument with* θ_0 *being replaced by* θ_1, *we can further prove* $\theta_2 \in \Theta_0$. *By induction, we can prove* $\{\theta_k\}_{k=1}^N \subset \Theta_0$. *Since* $\mathfrak{M}(\theta, s), \mathfrak{N}(\theta, s), \mathfrak{L}(\theta, s)$ *depend continuously on* θ, s, *there supreme values on compact set* $\Theta_0 \times [0, 1]$ *must be finite so we know* $C(\theta_0, h, N)$ *in* (Equation 4.116) *is upper bounded by a constant independent of* h *as well as* N *(recall* $N = \frac{T}{h}$*). Thus the error term* $O(Ch)$ *is of* $O(h)$ *order.*

Similar to the discussion in previous sections, we can naturally extend Theorem 4.6.12 to a *posterior* estimate.

Theorem 4.6.17 (*posterior* error analysis of forward Euler scheme)**.**

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \le (\sqrt{\delta_1}h + Ch^2)\frac{1 - e^{-\lambda t_k}}{1 - e^{-\lambda h}} + e^{-\lambda t_k} W_2(\rho_{\theta_0}, \rho_0) \text{ for any } t_k = kh,\ 0 \le k \le N.$$

The explicit definition of the constant C *is in* (Equation 4.116).

Up to this point, we mainly analyze the error term for the forward Euler scheme. In our numerical implementation, we adopt the scheme (Equation 4.48), which turns out to be a semi-implicit scheme with $O(h^2)$ local error. In the following discussion, we compare the difference between the numerical solutions of our semi-implicit scheme and forward Euler scheme.

Recall that the parametric Fokker-Planck equation (Equation 4.28) is an ODE: $\dot{\theta} = -G(\theta)^{-1}\nabla_\theta H(\theta)$. We consider two numerical schemes:

$$\theta_{n+1} = \theta_n - hG(\theta_n)^{-1}\nabla_\theta H(\theta_n) \quad \theta_0 = \theta,\ 1 \le n \le N \quad \text{forward Euler}, \tag{4.120}$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n - hG(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1}) \quad \hat{\theta}_0 = \theta,\ 1 \le n \le N \quad \text{semi-implicit Euler}. \tag{4.121}$$

We denote $F(\theta') = G(\theta')^{-1}\nabla_\theta F(\theta'')$, and set:

$$L_1 = \max_{1\le n\le N}\left\{\|F(\theta_n) - F(\hat{\theta}_n)\|/\|\theta_n - \hat{\theta}_n\|\right\},$$
$$L_2 = \max_{1\le k\le N}\{\|\nabla_\theta H(\hat{\theta}_n) - \nabla_\theta H(\hat{\theta}_{n+1})\|/\|\hat{\theta}_n - \hat{\theta}_{n+1}\|\},$$
$$M_1 = \max_{1\le n\le N}\{\|G(\hat{\theta}_n)^{-1}\|\}, \quad M_2 = \max_{1\le n\le N}\{\|\nabla_\theta H(\hat{\theta}_n)\|\},$$

where $\|\cdot\|$ is a vector norm (or its corresponding matrix norm).

Theorem 4.6.18 (Relation between forward Euler and proposed semi-implicit schemes)**.** *The numerical solutions θ_n and $\hat{\theta}_n$ of the forward Euler and semi-implicit schemes with time stepsize h and $Nh = T$ satisfy*

$$\|\theta_n - \hat{\theta}_n\| \le ((1 + L_1 h)^n - 1)\frac{M_1^2 M_2 L_2}{L_1}h \quad n = 1, 2, ..., N$$

This result implies that $\|\theta_n - \hat{\theta}_n\|$ can be upper bounded by $(e^{L_1 T} - 1)\frac{M_1^2 M_2 L_2}{L_1}h$. When assuming the upper bounds $L_1, L_2, M_1, M_2 \sim O(1)$ as $h \to 0$ (or equivalently $N \to \infty$), the differences between our proposed semi-implicit scheme and forward Euler scheme can be bounded by $O(h)$. As a consequence, we are able to establish $O(h)$ error bound for our proposed scheme (Equation 4.48).

Proof of Theorem 4.6.18. If we subtract (Equation 4.121) from (Equation 4.120),

$$(\theta_{n+1} - \hat{\theta}_{n+1}) = (\theta_n - \hat{\theta}_n) - h(G(\theta_n)^{-1}\nabla_\theta H(\theta_n) - G(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1}))$$

and denote $e_n = \theta_n - \hat{\theta}_n$ and $F(\theta) = G(\theta_n)^{-1}\nabla_\theta H(\theta)$, we may rewrite this equation as

$$e_{n+1} = e_n - h(F(\theta_n) - F(\hat{\theta}_n) + G(\hat{\theta}_n)^{-1}(\nabla_\theta H(\hat{\theta}_n) - \nabla_\theta H(\hat{\theta}_{n+1}))).$$

Recall the definitions of L_1, L_2, M_1, we have

$$\|e_{n+1}\| \leq \|e_n\| + hL_1\|e_n\| + hM_1L_2\|\hat{\theta}_{n+1} - \hat{\theta}_n\|.$$

By the semi-simplicit scheme, we have

$$\hat{\theta}_{n+1} - \hat{\theta}_n = -hG(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1})$$

Thus $|\hat{\theta}_{n+1} - \hat{\theta}_n\| \leq hM_1M_2$. This gives us a recurrent inequality,

$$\|e_{n+1}\| \leq \|e_n\| + hL_1\|e_n\| + M_1^2M_2L_2h^2,$$

which implies

$$\left(\|e_{n+1}\| + \frac{M_1^2M_2L_2}{L_1}h\right) \leq (1 + hL_1)\left(\|e_n\| + \frac{M_1^2M_2L_2}{L_1}h\right) \quad n = 0, 1, ..., N-1.$$

This leads to:

$$\|e_n\| \leq ((1 + hL_1)^n - 1)\frac{M_1^2M_2L_2}{L_1}h.$$

When we solve the ODE on $[0, T]$ with $h = T/N$, we have $(1 + hL_1)^n \leq (1 + hL_1)^N = \left(1 + \frac{L_1T}{N}\right)^N \leq e^{L_1T}$. This means all terms $\{\|e_n\|\}_{1\leq n\leq N}$ can be upper bounded by $(e^{L_1T} - 1)\frac{M_1^2M_2L_2}{L_1}h$. □

Remark 23. *In order to make our argument clear and concise, we omitted the errors introduced by the approximation of ReLU function ψ_ν. Careful analysis on how well $\nabla\psi_\nu$ can approximate a general gradient field is among our future research directions.*

Remark 24. *The convergence property of the Stochastic Gradient Descent method (mainly Adam method) used in our Algorithm Algorithm 2 is not discussed in details. One can*

check its convergence analysis in the paper [149]. Based on our experiences, for most of the smooth potential functions $V \in \mathcal{V}$ with diffusion coefficient D not too small (i.e. $D > 0.1$), our algorithm shows convergent behavior and produces accurate result when checking against the true solution if it is possible.

4.7 Numerical examples

In this section, we consider solving the Fokker-Planck equation (Equation 4.6) on $\mathbb{R}^d$ with initial condition $\rho_0(x) = \mathcal{N}(0, I_d)$[3] by using Algorithm Algorithm 2. We demonstrate several numerical examples with different potential functions V. In the following experiments, unless specifically stated, we choose the length of normalizing flow T_θ as 60. We set $\psi_\nu : \mathbb{R}^d \to \mathbb{R}$ as ReLU network with number of layers equals 6 and hidden dimension equals 20. We use Adam (Adaptive Moment Estimation) Stochastic Gradient Descent method [149] with default parameters $D_1 = 0.9, D_2 = 0.999; \epsilon = 10^{-8}$. For the parameters of Algorithm Algorithm 2, we choose $\alpha_{\text{out}} = 0.005$, $\alpha_{\text{in}} = 0.0005$. We follow Remark Remark 17 to choose $K_{\text{in}}, K_{\text{out}} = \max\{1000, 300d\}$. Based on our experience, we set $M_{\text{out}} = O(\frac{h}{\alpha_{\text{out}}})$. The suitable value of M_{in} can be chosen after several quick tests to make sure that every inner optimization problem (Equation 4.67) can be solved.

Our Python code is uploaded to Github, which can be downloaded from website https://github.com/LSLSliushu/Parametric-Fokker-Planck-Equation.

4.7.1 Quadratic Potential

Our first set of examples uses quadratic potential V. In this case, we can compute the explicit solution of (Equation 4.6). These examples are used for the verification purpose, because we can check the results with exact solutions.

[3] We can set initial value θ_0 so that $T_{\theta_0} = Id$ and thus $\rho_0 = T_{\theta_0 \sharp} p$ is standard Gaussian distribution.

2D cases

We take $d = 2$, and set $V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$, with $\mu = [3,3]^{\mathrm{T}}$ and $\Sigma = \mathrm{diag}([0.25, 0.25])$. The solution of (Equation 4.6) is:

$$\rho_t = \mathcal{N}(\mu(t), \Sigma(t)) \quad \mu(t) = (1-e^{-4t})\mu,\ \Sigma(t) = \begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-8t} & \\ & \frac{1}{4}+\frac{3}{4}e^{-8t} \end{bmatrix} \quad t \geq 0.$$

We solve the equation in time interval $[0, 0.7]$ with time stepsize 0.01. We set $M_{\text{out}} = 20$ and $M_{\text{in}} = 100$.

To compare against the exact solution, we set $M = 6000$ and sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_M\} \sim T_{\theta_k \sharp} p$ at time t_k and use:

$$\hat{\mu}^k = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{X}_j, \quad \hat{\Sigma}^k = \frac{1}{M-1}\sum_{j=1}^{M}(\boldsymbol{X}_j - \hat{\mu}_k)(\boldsymbol{X}_j - \hat{\mu}_k)^{\mathrm{T}}$$

to compute for its empirical mean and covariance of $\hat{\rho}_k$. We plot the blue curves $\{\hat{\mu}^{(k)}\}$, $\{\hat{\mu}_2^{(k)}\}$, $\{(\hat{\Sigma}_{11}^{(k)}, \hat{\Sigma}_{22}^{(k)})\}$, $\{(\hat{\mu}_1^{(k)}, \hat{\Sigma}_{11}^{(k)})\}$ in Figure 4.13, these plots properly captures the exponential convergence exhibited by the explicit solution (red curves) $\{\mu(t)\}$, $\{\mu_2(t)\}$, $\{(\Sigma_{11}(t), \Sigma_{22}(t))\}$, $\{(\mu_1(t), \Sigma_{11}(t))\}$.

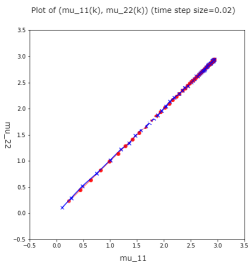


Figure 4.9: $\{\hat{\mu}^{(k)}\}$

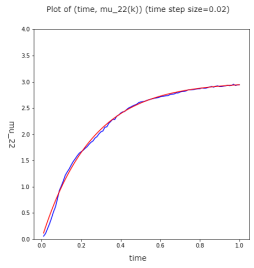


Figure 4.10: $\{\hat{\mu}_1^{(k)}\}$

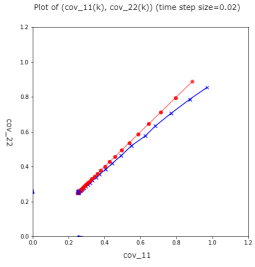


Figure 4.11: $\{(\hat{\Sigma}_{11}^{(k)}, \hat{\Sigma}_{22}^{(k)})\}$

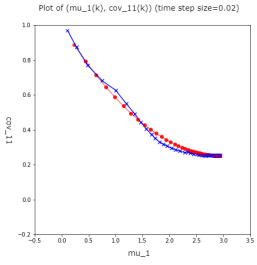


Figure 4.12: $\{(\hat{\mu}_{(k)}, \hat{\Sigma}_{11}^{(k)})\}$

Figure 4.13: Plot of empirical statistics (numerical solution: blue; real solution: red)

We also exam the network $\psi_{\hat{\nu}}$ trained at the end of each outer iteration. Generally

speaking, the gradient field $\nabla\psi_{\hat{\nu}}$ reflects the movements of the particles under the Vlasov-typed dynamic (Equation 4.7) at every time step. Here are the graph of $\psi_{\hat{\nu}}$ at $k = 10, k = 140$ (Figure 4.14, Figure 4.15). As we can see from these graphs, the gradient field is in the same direction, but judging from the variation of two $\psi_{\hat{\nu}}$s, when $k = 10$, $|\nabla\psi_{\hat{\nu}}|$ is much greater than its value at $k = 140$. This is because when $t = 140$, the distribution is already close to the Gibbs distribution, the particles no longer need to move for a long distance to reach their final destination.

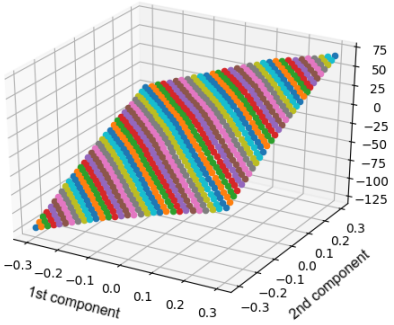


Figure 4.14: Graph of $\psi_{\hat{\nu}}$ after $M_{\text{out}} = 20$ outer iterations at $k = 10$th time step

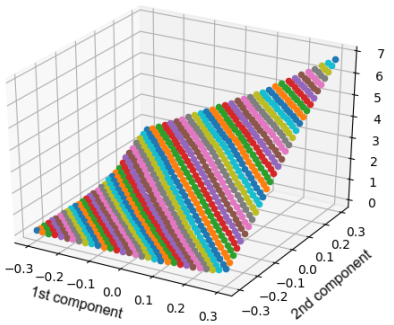


Figure 4.15: Graph of $\psi_{\hat{\nu}}$ after $M_{\text{out}} = 20$ outer iterations at $k = 140$th time step

In the next example, we apply our algorithm to the Fokker-Planck equation with non-isotropic potential

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \quad \mu = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 1 & \\ & \frac{1}{4} \end{bmatrix}.$$

One can verify that the solution to (Equation 4.6) is

$$\rho_t = \mathcal{N}(\mu_t, \Sigma_t) \quad \mu_t = \begin{bmatrix} 3(1-e^{-t}) \\ 3(1-e^{-4t}) \end{bmatrix}, \ \Sigma_t = \begin{bmatrix} 1 & \\ & \frac{1}{4}(1+3e^{-8t}) \end{bmatrix}.$$

We use the same parameters as before. We solve (Equation 4.6) on time interval $[0, 1.4]$ with time step size 0.005.

Similarly, we also plot the empirical mean trajectory, one can compare it with the true solution $\mu(t) = (3(1 - e^{-t}), 3(1 - e^{-4t}))$. Both the curvature and the exponential convergence to μ are captured by our numerical result. To demonstrate the effectiveness of our formulation, we also compare our result with the mean trajectory obtained by computing the flat gradient flow $\dot{\theta} = -\nabla_\theta H(\theta)$, which is plotted in Figure 4.17. It reveals very different behavior of the flat gradient (∇_θ) flow and Wasserstein gradient ($G(\theta)^{-1}\nabla_\theta$) flow. Clearly, our approximation based on Wasserstein gradient flow captures the exact mean function much more accurately. We compare the graph of trained $\psi_{\hat{\nu}}$ at different time steps

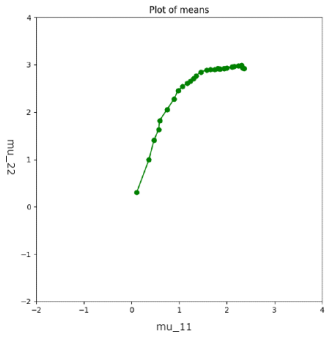


Figure 4.16: mean trajectory of $\{\rho_{\theta_t}\}$ w.r.t. $\dot{\theta} = -G(\theta)^{-1}\nabla_\theta H(\theta)$

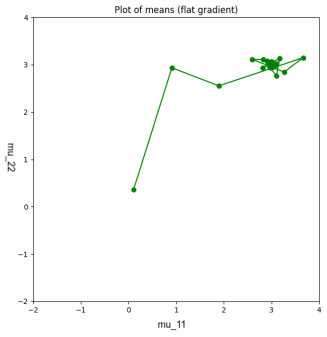


Figure 4.17: mean trajectory of $\{\rho_{\theta_t}\}$ w.r.t. $\dot{\theta} = -\nabla_\theta H(\theta)$

$k = 10, 140$ (Figure 4.18, Figure 4.19). The directions of $\nabla\psi_{\hat{\nu}}$ at $k = 10$ and $k = 140$ is different from the previous example. This is caused by the non-isotropic quadratic (Gaussian) potential V used in this example.

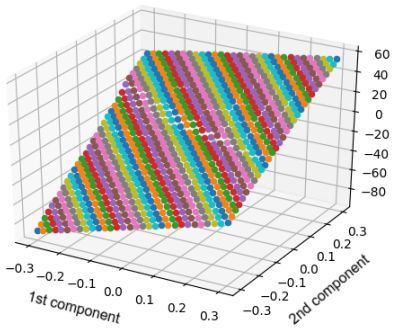


Figure 4.18: Graph of $\psi_{\hat{\nu}}$ after $M_{\text{out}} = 20$ outer iterations at $k = 10$th time step

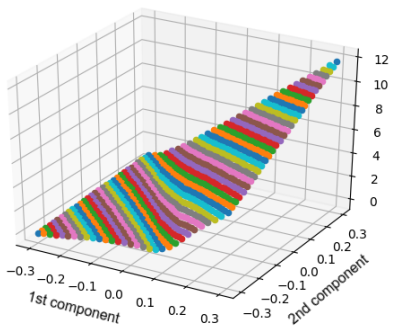


Figure 4.19: Graph of $\psi_{\hat{\nu}}$ after $M_{\text{out}} = 20$ outer iterations at $k = 140$th time step

Verification of the error estimate

We verify the $O(h)$ error estimation discussed in subsubsection 4.6.3 based on numerical experiments with quadratic potentials. We consider $V(x) = |x - \mu|^2$ defined on $\mathbb{R}^2$ with $\mu = (12.0, 12.0)$ and ρ_0 as standard Gaussian on time interval $[0, 1]$. We run our algorithm with several different time step size $h = 0.01, 0.05, 0.08, 0.1, 0.2, 0.3$ and record their corresponding mean trajectory $\{\hat{\mu}^{(k)}\}$ as defined in Section 6.1.1. During this process, we need to adjust our hyperparameters $\alpha_{\text{in}}, \alpha_{\text{out}}, M_{\text{in}}, M_{\text{out}}$ correspondingly in order to guarantee the convergence of Adam method. Denote $\{\mu(t_k)\}$ as the real solution. We compute the average l^2 error of mean values as $\text{AveErr}(h) = \frac{1}{N}\sum_k |\hat{\mu}^{(k)} - \mu(t_k)|$. We pick h in a range larger than 0.01 because when h is smaller, the influence from the approximation error δ_0 of normalizing flow T_θ as well as initial error $W_2(\rho_0, \rho_{\theta_0})$ start to dominate the overall error. Figure 4.20 exhibits the linear relationship between our numerical error $\text{AveErr}(h)$ and time step size h, which confirms our theoretical estimates.

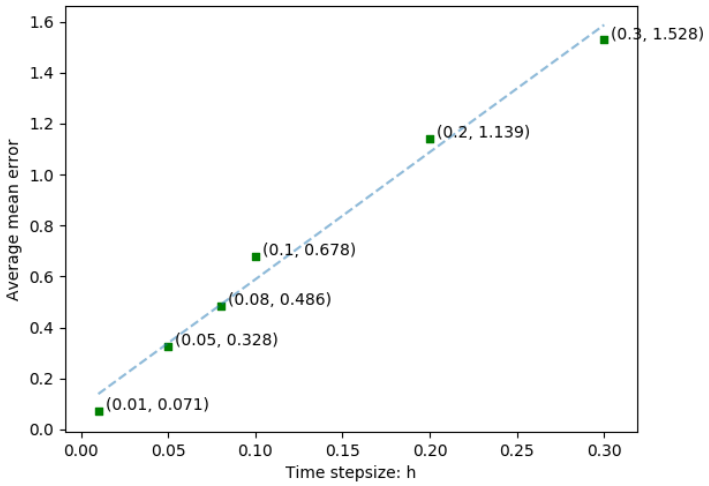


Figure 4.20: Numerical errors versus time stepsize h.

Remark 25. *The reason of choosing quadratic potential is because its corresponding FPE has an explicit solution. The reason that we focus on the average error of mean vectors is mainly due to computational accuracy and convenience: one can approximate the er-*

ror of the mean vector of a distribution by computing the arithmetic average of samples, which is faster and more accurate than computing for the L^2-Wasserstein error among two distributions.

Higher dimension

We implement our algorithm in higher dimensional space. In the next example, we take $d = 10$, and consider the quadratic potential

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \quad \Sigma = \mathrm{diag}(\Sigma_A, I_2, \Sigma_B, I_2, \Sigma_C) \quad \mu = (1,1,0,0,1,2,0,0,2,3)^{\mathrm{T}}.$$

Here we set the diagonal blocks as:

$$\Sigma_A = \begin{bmatrix} \frac{5}{8} & -\frac{3}{8} \\ -\frac{3}{8} & \frac{5}{8} \end{bmatrix} \quad \Sigma_B = \begin{bmatrix} 1 & \\ & \frac{1}{4} \end{bmatrix} \quad \Sigma_C = \begin{bmatrix} \frac{1}{4} & \\ & \frac{1}{4} \end{bmatrix}.$$

We solve the equation in time interval $[0, 0.7]$ with time stepsize 0.005. We set $M_{\text{out}} = 20$ and $M_{\text{in}} = 100$. To demonstrate the results, 6000 samples from the reference distribution p are drawn and pushforwarded by using our computed map T_{θ_k}. We plot a few snapshots of the pushforwarded points (from $t = 0.05$ to $t = 0.70$) in Figure 4.24. One can check that the distribution of our numerical computed samples gradually converges to the Gibbs distribution $\mathcal{N}(\mu, \Sigma)$.

We solve (Equation 4.6) on time interval $[0, 2]$ with time step size $h = 0.005$. We set $K_{\text{in}} = K_{\text{out}} = 3000$ and choose $M_{\text{out}} = 30$, $M_{\text{in}} = 100$. To demonstrate the results, 6000 samples from the reference distribution p are drawn and pushforwarded by using our computed map T_{θ_k}. We exhibit the projection of the samples on $0-1$, $4-5$ and $8-9$ plane in Figure 4.24 at time $t = 2.0$. One can verify that the distribution of our numerical computed samples converges to the Gibbs distribution $\mathcal{N}(\mu, \Sigma)$. The explicit solution to the Fokker-Planck equation is always Gaussian distribution $\mathcal{N}(\mu(t), \Sigma(t))$ with mean $\mu(t)$

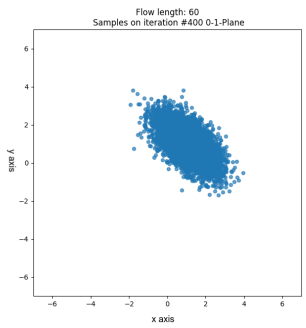


Figure 4.21: projection of samples on 0-1 plane

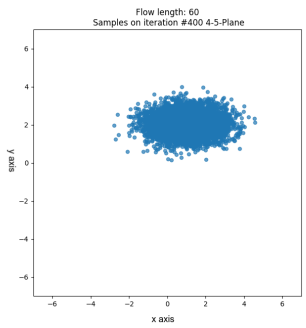


Figure 4.22: projection of samples on 4-5 plane

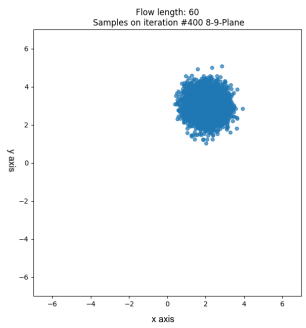


Figure 4.23: projection of samples on 8-9 plane

Figure 4.24: Sample points of computed ρ_{θ_t} projected on different planes at $t = 2.0$

and covariance matrix $\Sigma(t)$:

$$\mu(t) = (1 - e^{-t}, 1 - e^{-t}, 0, 0, 1 - e^{-t}, 2(1 - e^{-4t}), 0, 0, 2(1 - e^{-4t}), 3(1 - e^{-4t}))^{\mathrm{T}},$$

$$\Sigma(t) = \mathrm{diag}(\Sigma_A(t), I, \Sigma_B(t), I, \Sigma_C(t)),$$

$$\text{with } \Sigma_A(t) = \begin{bmatrix} \frac{5}{8} + f(t) & -\frac{3}{8} + f(t) \\ -\frac{3}{8} + f(t) & \frac{5}{8} + f(t) \end{bmatrix}, \ \Sigma_B(t) = \begin{bmatrix} 1 & \\ & \frac{1+3e^{-8t}}{4} \end{bmatrix},$$

$$\Sigma_C(t) = \begin{bmatrix} \frac{1+3e^{-8t}}{4} & \\ & \frac{1+3e^{-8t}}{4} \end{bmatrix},$$

$$\text{here } f(t) = -\frac{2}{7}e^{-t} + \frac{1}{3}e^{-2t} + \frac{55}{168}e^{-8t}.$$

To compare against the exact solution, we set sample size $M = 6000$ and compute the empirical mean $\hat{\mu}^k$ and covariance $\hat{\Sigma}^k$ of our numerical solution $\hat{\rho}_k$ at time t_k. We evaluate the error between $\hat{\mu}^{(k)}$ and $\mu(t_k)$; $\hat{\Sigma}^{(k)}$ and $\Sigma(t_k)$. We plot the error curves of $\|\hat{\mu}^{(k)} - \mu(t_k)\|_2$ (Figure 4.25) and $\|\hat{\Sigma}^{(k)} - \Sigma(t_k)\|_F$ (Figure 4.26). Here $\|\cdot\|_F$ is the matrix Frobenius norm. Figure 4.27 captures the exponential decay of H along its Wasserstein gradient flow, this verifies the entropy dissipation property of the Fokker-Planck equation with convex potential function V.

In this case, we take a closer look at the loss in the inner loops. Figure 4.28 shows the

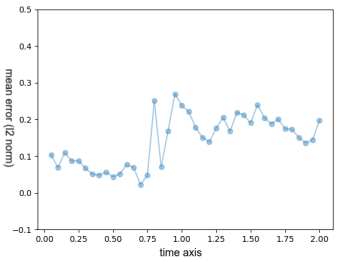


Figure 4.25: mean error (l_2)

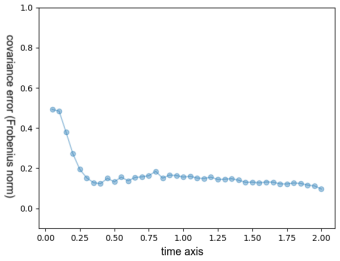


Figure 4.26: covariance error ($\|\cdot\|_F$)

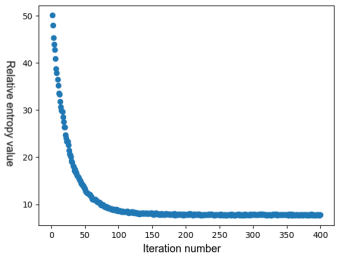
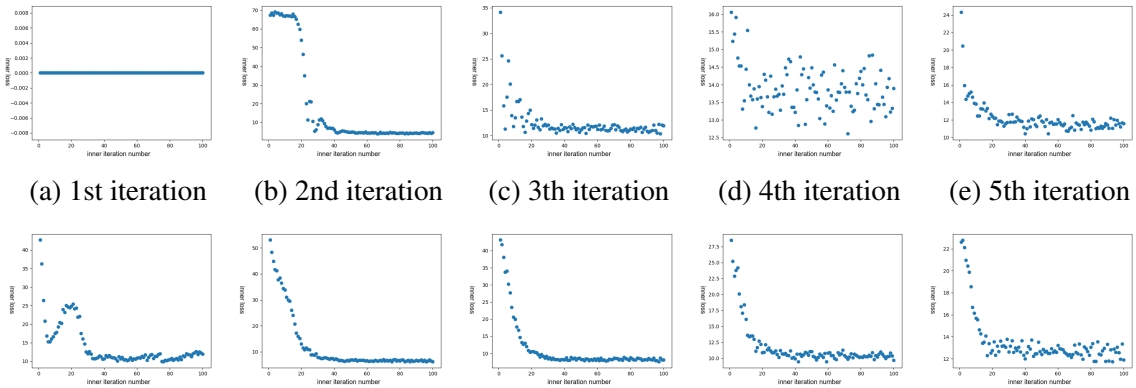


Figure 4.27: Plot of $\{H(\theta)\}$

first 10 (out of 20) loss plots when applying SGD method to solve (Equation 4.70) with $k = 200$ ($t = 200 \cdot h = 1.0$). The remaining loss plots from the 11th outer iteration to 20th iteration are similar to the plots in the second row. The situations are similar for other time step k. We believe that $M_{\text{in}} = 100$ works well in this problem, the SGD method we used can thoroughly solve the variational problem (Equation 4.70) for each outer loop.



(a) 1st iteration (b) 2nd iteration (c) 3th iteration (d) 4th iteration (e) 5th iteration

(f) 6th iteration (g) 7th iteration (h) 8th iteration (i) 9th iteration (j) 10th iteration

Figure 4.28: Plots of inner loop losses

4.7.2 Experiments with more general potentials

In this section, we exhibit two examples with more general potentials in higher dimensional space.

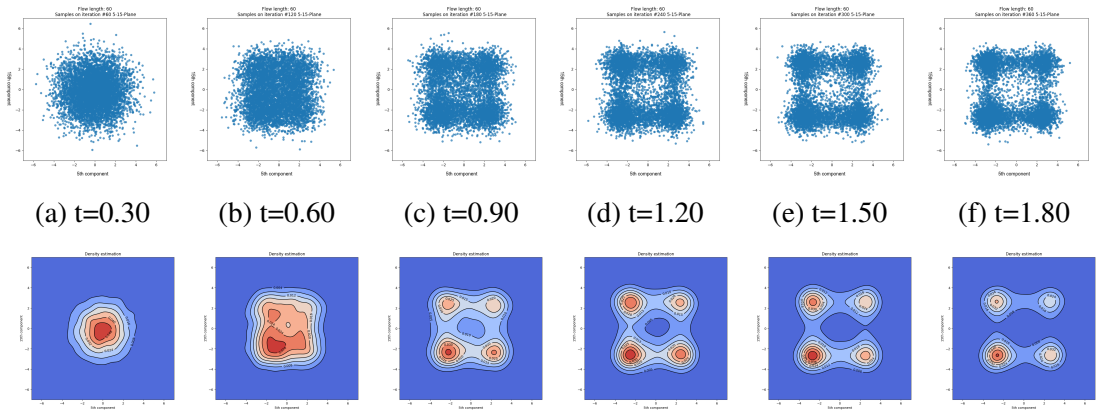
Styblinski-Tang potential

In this example, we set dimension $d = 30$, and consider the Styblinski–Tang function [160]

$$V(x) = \frac{3}{50}\left(\sum_{i=1}^{d} x_i^4 - 16x_i^2 + 5x_i\right).$$

We solve (Equation 4.6) with potential V on time interval $[0, 3]$ with time step size $h = 0.005$. We set $K_{\text{in}} = K_{\text{out}} = 9000$ and $M_{\text{in}} = 100$, $M_{\text{out}} = 30$.

To exhibit sample results, due to the symmetric structure of the potential function, we project the sample points in $\mathbb{R}^{30}$ to some random plane, such as $5-15$ plane in this research. The sample plots and their estimated densities are presented in Figure 4.29.



(a) t=0.30 (b) t=0.60 (c) t=0.90 (d) t=1.20 (e) t=1.50 (f) t=1.80

(g) t=0.30 (h) t=0.60 (i) t=0.90 (j) t=1.20 (k) t=1.50 (l) t=1.80

Figure 4.29: Sample points and estimated densities of ρ_{θ_t} on $5 - 15$ plane at different time nodes

In this special example, the potential function is the direct addition of same functions, we can exploit this property and show that any marginal distribution

$$\varrho_j(x_j, t) = \int \dots \int \rho(x, t)\, dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_d$$

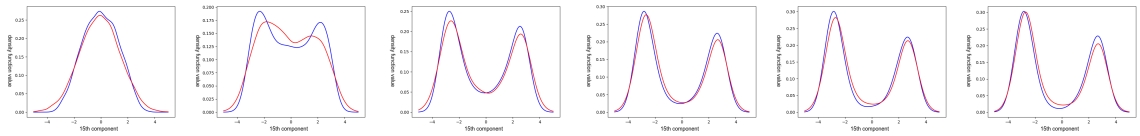
of the solution ρ_t solves the following the 1D Fokker-Planck equation:

$$\frac{\partial \varrho(x,t)}{\partial t} = \frac{\partial}{\partial x}(\varrho(x,t)\, V'(x)) + D\Delta\varrho(x,t) \quad \varrho(\cdot,0) = \mathcal{N}(0,1), \tag{4.122}$$
$$\text{with } V(x) = \frac{3}{50}(x^4 - 16x^2 + 5x).$$

We then solve the SDE associated to (Equation 4.122):

$$dX_t = -V'(X_t)\, dt + \sqrt{2D} dB_t \quad X_0 \sim \mathcal{N}(0,1). \tag{4.123}$$

Since (Equation 4.123) is an SDE in one dimensional space, we can solve it with high accuracy by Euler-Maruyama scheme [161] and use it as a benchmark for our numerical solution. The following Figure 4.30 exhibits both the estimated densities for our numerical solutions (marginal distribution on the 15th component) and the solution of (Equation 4.123) given by Euler-Maruyama scheme with step size 0.005. The sample sizes for both solutions equal to 6000.



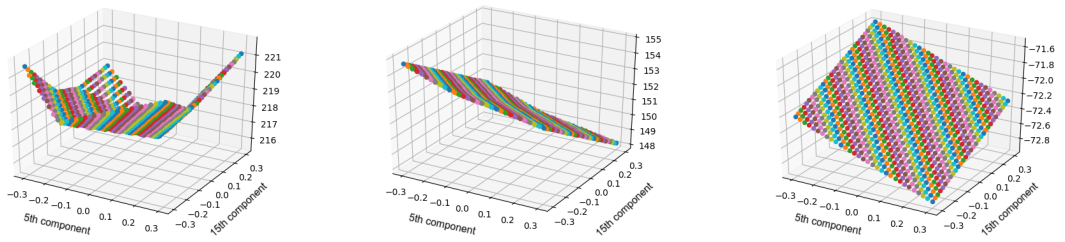
t=0.30 t=0.60 t=0.90 t=1.20 t=1.50 t=1.80

Figure 4.30: Estimated densities of our numerical solution(red) (projected onto the 15th component) and the solution given by Euler Maruyama scheme(blue)

We also illustrate the graphs of $\psi_{\hat{\nu}}$ on $5-15$ plane trained at different time steps in Figure 4.31.

Affects of different initial distributions

Different initial conditions ρ_0 affect the behavior of solutions of neural parametric FPE differently, especially on the convergence speed to the Gibbs distribution. Here is an example. We consider V as Styblinski-Tang potential in $\mathbb{R}^2$. We compute the solutions with three



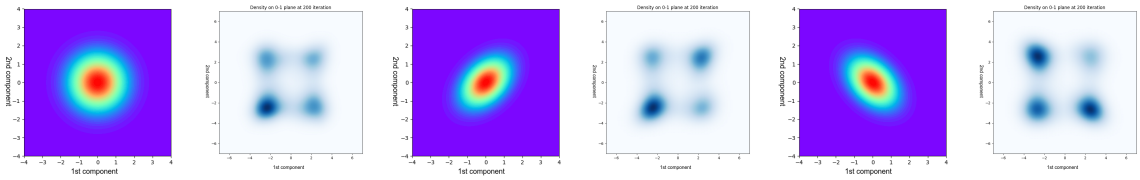
$\psi_{\hat{\nu}}$ at $k = 30$ $\psi_{\hat{\nu}}$ at $k = 150$ $\psi_{\hat{\nu}}$ at $k = 360$

Figure 4.31: Graph of $\psi_{\hat{\nu}}$ on $5 - 15$ plane trained at different time steps

different initial distributions given as Gaussian distributions with covariances

$$\Sigma_1 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \frac{13}{8} & \frac{5}{8} \\ \frac{5}{8} & \frac{13}{8} \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \frac{13}{8} & -\frac{5}{8} \\ -\frac{5}{8} & \frac{13}{8} \end{bmatrix},$$

respectively. Although the solutions converge to the Gibbs distribution, as expected from the theory, regardless of the initial density. Their convergence speed may be different. Figure 4.32 shows the initial distributions and the corresponding densities (which are the estimations of the samples obtained from our algorithm) at $t = 1.0$. As we can observe, the numerical result produced by $\rho_0 = \mathcal{N}(0, \Sigma_1)$ is already close to Gibbs distribution at $t = 1.0$, while numerical results associated to Σ_2, Σ_3 still have noticeable differences from Gibbs. They seems to be trapped in intermediate metastable statuses that are clearly influenced by the orientations in initial distributions.



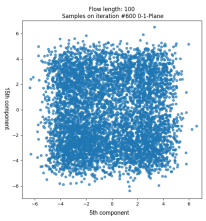
$\rho_0 = \mathcal{N}(0, \Sigma_1)$ $t = 1.0$ $\rho_0 = \mathcal{N}(0, \Sigma_2)$ $t = 1.0$ $\rho_0 = \mathcal{N}(0, \Sigma_3)$ $t = 1.0$

Figure 4.32: Different behaviors of numerical solution with different ρ_0s

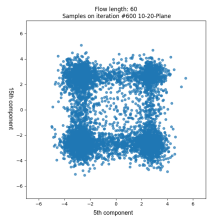
In general, we believe that the choice of ρ_0 affects the behavior of numerical solution. Choosing a suitable ρ_0 may shorten the computing time in the training process.

Solving the equation with different diffusion coefficients

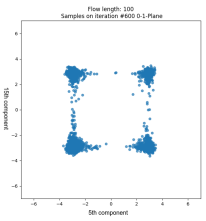
The different behaviors of the Fokker-Planck equation caused by different diffusion coefficients D can be captured by our algorithm. As the following figure shows, we apply our method to solve Fokker-Planck equation with Styblinski-Tang potential function with $D = 0.1, 1.0, 10.0$ and exhibit samples points and estimated density surfaces at the time $t = 3.0$.



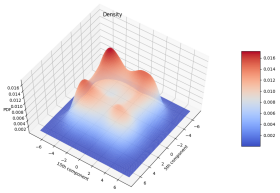
(a) Samples $D = 10$



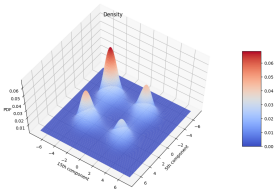
(b) Samples $D = 1$



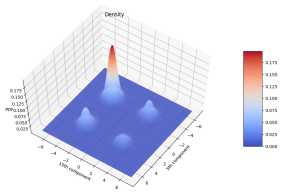
(c) Samples $D = 0.1$



(d) Density $D = 10$



(e) Density $D = 1$



(f) Density $D = 0.1$

Figure 4.33: Samples and estimated densities at $t = 3.0$, from left to right: $D = 10$, $D = 1.0$, $D = 0.1$

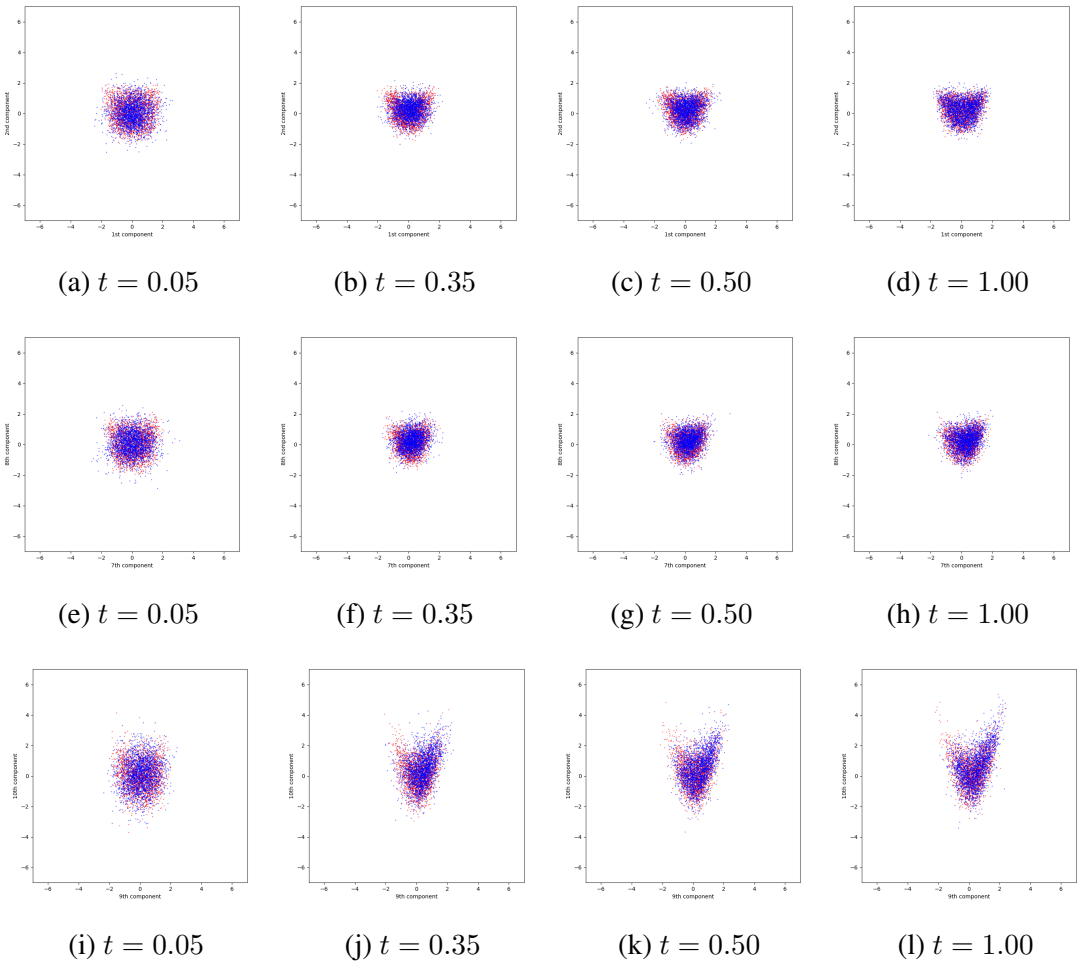
Rosenbrock potential

In this example, we set dimension $d = 10$. We consider the Rosenbrock typed function [162]:

$$V(x) = \frac{3}{50}\left(\sum_{i=1}^{d-1} 10(x_{k+1} - x_k^2)^2 + (x_k - 1)^2\right),$$

which involve interactions among its coordinates. We solve the corresponding (Equation 4.6) on time interval $[0, 1]$ with step size $h = 0.005$. We set the length of normalizing flow T_θ as 100. We set $K_{\text{in}} = K_{\text{out}} = 3000$ and $M_{\text{in}} = 100$, $M_{\text{out}} = 60$.

Here are the sample results, we exhibit the projection of sample points on the $1-2$, $7-8$ and $9-10$ plane in Figure 4.34. Blue samples are obtained from our numerical solution while red samples are obtained by applying Euler-Maruyama scheme with the same step size.



(a) $t = 0.05$ (b) $t = 0.35$ (c) $t = 0.50$ (d) $t = 1.00$

(e) $t = 0.05$ (f) $t = 0.35$ (g) $t = 0.50$ (h) $t = 1.00$

(i) $t = 0.05$ (j) $t = 0.35$ (k) $t = 0.50$ (l) $t = 1.00$

Figure 4.34: Samples of our numerical solution (blue) and Euler-Maruyama (red) on different planes at different time nodes

4.7.3 Discussion on time consumption

we should point out that the running time of our algorithm depends on the following three aspects:

(i) Dimension d of the problem; potential function V;

(ii) The size of normalizing flow T_θ and fully connected neural network ψ_ν;

(iii) Number of time steps N; outer iterations M_{out}; inner iterations M_{in}; sample size K_{out} and K_{in}.

Among them, the networks in (ii) are selected according to (i). The hyper-parameters $M_{\text{out}}, M_{\text{out}}, K_{\text{out}}, K_{\text{in}}$ in (iii) are chosen based on our trial and error as well as Remark 17 stated earlier in this paper.

All numerical examples reported in this paper are computed on a Laptop with Intel Core™ i5-8250U CPU @ 1.60GHz × 8 processor. For most of the high dimensional examples ($d \geq 10$), we choose the length of T_θ between 60 and 100; for the ReLU network ψ_ν, we set its number of layers equal to 6 with hidden dimension 20. We set $M_{\text{out}} \sim 50, M_{\text{in}} \sim 100$ and choose sample sizes $K_{\text{out}}, K_{\text{in}}$ according to Remark 17. The total running time is ranged in $20 - 40$ hours.

We observe that the running time of our algorithm is dominated by the inner loop of Algorithm 2, i.e. the part for optimizing over ψ_ν. The cost associated with this part can be estimated as $O(N \cdot M_{\text{out}} \cdot M_{\text{in}} \cdot (K_{\text{in}} t_a + t_b))$, where t_a denotes the time cost of using backpropagation to evaluate the gradient w.r.t. ν of each $|\nabla\psi_\nu(T_{\theta_0}(\mathbf{X}_k)) - (T_\theta(\mathbf{X}_k) - T_{\theta_0}(\mathbf{Y}_k))|^2$ in every inner loop of Algorithm 2, and t_b denotes the time for updating ν by Adam method. Here t_a, t_b both depend on d, V and the sizes of networks T_θ, ψ_ν. According to our experiences, for most of the cases, t_a is of the order of magnitude around $10^{-5}s$ and t_b is around $10^{-2}s$.

Although the cost for our current implementation of the train process is still high, we want to remind that there is a distinct advantage in the sampling application, namely that the

network training just needs to be done once. The trained network can be reused to generate samples, regardless the sample size, from distribution ρ_t by pushing forward samples from the reference distribution p with negligible additional cost. This is in sharp contrast to the classical MCMC sampling techniques, which requires to solve the SDE associated with FPE by numerical methods, such as Euler-Maruyama scheme, for every sample.

4.8 Discussion

In this paper, we design and analyze an algorithm for computing the high dimensional Fokker-Planck equations. Our approach is based on transport information geometry with probability formulations arisen in deep learning generative models. We first introduce the parametric Fokker-Planck equations, a set of ODE, to approximate the original Fokker-Planck equation. The ODE can be viewed as the "spatial" discretization of the PDE using neural networks. We propose a variational version of the semi-implicit Euler scheme and design a discrete time updating algorithm to compute the solution of the parametric Fokker-Planck equations. Our method is a sampling based approach that is capable to handle high dimensional cases. It can also be viewed as an alternative of the JKO scheme used in conjunction with neural networks. More importantly, we prove the asymptotic convergence and error estimates, both under the Wasserstein metric, for our proposed scheme.

We hope that our study may shed light on principally designing deep neural networks and other machine learning approaches to compute solutions of high dimensional PDEs, and systematically analyzing their error bounds for understandable and trustworthy computations. Our parametric Fokker-Planck equations are derived by approximating the density function in free energy using neural networks, and then following the rules in calculus of variation to get its Euler-Langrange equation. The energy law and principles in variational framework build a solid foundation for our "spatial" discretization that is able to inherit many desirable physical properties shared by the PDEs, such as relative entropy dissipation in a neural network setting. Our numerical scheme provides a systemic mechanism to de-

sign sampling efficient algorithms, which are critical for high dimensional problems. One distinction of our method is that, contrary to the data dependent machine learning studies in the literature, our approach does not require any knowledge of the ”data” from the PDEs. In fact, we generate the “data” to compute the numerical solutions, just like the traditional numerical schemes do for PDEs. More importantly, we carried out the numerical analysis, using tools such as KL divergence and Wasserstein metric from the transport information geometry, to study the the asymptotic convergence and error estimates in probability space. We emphasize that the Wasserstein metric provides a suitable geometric structure to analyze the convergence behavior in generative models, which are widely used in machine learning field. For this reason, we believe that our investigations can be adopted to understand many machine learning algorithms, and to design efficient sampling strategies based on pushforward maps that can generate flows of samples in generative models.

We also believe that the approaches in algorithm design and error analysis developed in this study can be extended to other types of equations. On one hand, our method is ready to be applied to equations such as porous media equation and aggregate equation, which possess gradient flow structures; On the other hand, we are also working on applying similar technique to high dimensional Hamiltonian flows instead of gradient flows. This could be more challenging from both computational and theoretical aspects since Hamiltonian is a second order differential system. Typically, we are interested in applying our computational tool to deal with Schrödinger equation as well as Schrödinger bridge systems in high dimensional space, and many more. Those topics are worth to be further investigated in the future.

CHAPTER 5
HAMILTONIAN PROCESS ON FINITE GRAPHS VIA WASSERSTEIN HAMILTONIAN THEORY

5.1 Introduction

Hamiltonian systems, including both ordinary or partial differential equations (ODEs or PDEs respectively), are ubiquitous in applications. Their mathematical studies have a long and rich history (see e.g., [163, 164, 165]). Traditionally, the ambient space on which to define a Hamiltonian system is continuous, such as Euclidean space $\mathbb{R}^n$ or smooth manifolds like torus $\mathbb{T}^2$. What is a Hamiltonian process if the underlying space becomes discrete, such as a finite graph? This is the question that we would like to explore within the framework of optimal transport (OT) in this study.

Our motivation to consider this question is 3-fold. Curiosity is at the first place. Secondly, the notion of gradient flow on graph has been investigated extensively using OT theory (see e.g. [166, 27] and references therein). For example, an irreducible and reversible continuous time Markov chain on graph can be viewed as the gradient flow of entropy with respect to the discrete Wasserstein metric [166]. Naturally, we are inspired to ask whether the concept of Hamiltonian process on graph exists or not. To the best of our knowledge, the Hamiltonian mechanics on graph has not been explored yet. Finally and most importantly, recent developments in several practical problems, which can be defined in both continuous and discrete spaces, demonstrate Hamiltonian principles on graph. They are

(i) the OT problem (see e.g. [167]),

$$W_2^2(\rho_0,\rho_1)=\inf_v\{\int_0^1 \mathbb{E}[|\dot{X}_t|^2]dt \ : \ \dot{X}_t=v(t,X_t), X_0\sim\rho^0, X_1\sim\rho^1\}, \tag{5.1}$$

(ii) the SBP (see e.g. [168]),

$$\inf_{v} \Big\{ \int_0^1 \frac{1}{2}\mathbb{E}[|v(t,X_t)|^2]dt : \dot{X}_t = v(t,X_t) + \sqrt{\hbar}\dot{B}_t,\ X_0 \sim \rho^0,\ X_1 \sim \rho^1 \Big\} \tag{5.2}$$

and (iii) the Schrödinger equation (see e.g. [169, 170, 171]),

$$\inf_{v} \Big\{ \int_0^T \frac{1}{2}\mathbb{E}[|\dot{X}_t|^2]dt : \dot{X}_t = v(t,X_t) + \sqrt{\hbar}\dot{B}_t,\ X_0 \sim \rho^0,\ X_1 \sim \rho^1 \Big\}. \tag{5.3}$$

The above formulations are presented in Euclidean space where $v \in \mathbb{R}^d$ can be any smooth vector field, X_t is a stochastic process with prescribed probability densities ρ^0 and ρ^1 at time 0 and 1 respectively, B_t is the standard Brownian motion and $\hbar > 0$ is a constant.

A common property shared by these problems is that their critical points obey the Hamiltonian principle. For instance, the minimizer of OT problem (Equation 5.1) satisfies a Hamiltonian PDE with the Hamiltonian $H(x,v,t) = \frac{1}{2}|v|^2$ (see e.g. [172]). The minimizer of SBP (Equation 5.2) is the solution of a Hamiltonian PDE with $H(x,v) = \frac{1}{2}|v|^2 - \frac{1}{8}\hbar\frac{\delta}{\delta\rho}\mathcal{I}(\rho)(t,x)$ where the Fisher information $\mathcal{I}(\rho) = \int_{\mathbb{R}^d} |\nabla \log \rho(x)|^2 \rho(x)dx$ (see e.g. [173, 174]). Needless to say, the critical point of (Equation 5.3) satisfies the the Schrödinger equation, which is a well-known Hamiltonian system. The problems stated in (Equation 5.1), (Equation 5.2) and (Equation 5.3) can be posed, with nominal changes, on a graph, and the density functions of their critical points have been studied on the Wasserstein manifold (see [175], [176, 177], [171]) showing that they satisfy Hamiltonian ODEs. Based on those results, we investigate the properties of stochastic process $X(t)$ and provide a definition of Hamiltonian process on graph within the optimal transport framework.

Defining Hamiltonian process on graph must face several intrinsic difficulties. The most obvious one is that $X(t)$ is a stochastic process jumping from node to node on the graph, while its continuous space counterpart trajectory is a spatial-temporal continuous function. Another challenge is about characteristic line. In fact, it is not clear how to define characteristic on graph. Furthermore, there is no reported result about examining whether

a stochastic process, such as discrete OT and SBP, can preserve Hamiltonian along its trajectory, just like a classical Hamiltonian system does in continuous space.

To fill the gaps on finite graph, our idea is lifting the process on graph into a motion on its density manifold. To be more precise, we define the Hamiltonian process by a random process whose density and generators of instantaneous transition rate matrix form a Wasserstein Hamiltonian flow on the cotangent bundle of density manifold. Meanwhile, we show that such defined Hamiltonian processes exist in numerous practical problems, such as the discrete OT problem and SBP. Two important classes of Hamiltonian processes, namely the stationary Hamiltonian process and the periodic Hamiltonian process, are also discussed via the framework of SBP. They correspond to the invariant measure and the periodic solution of the Hamiltonian flow on the density space. We would like to mention that the Wasserstein Hamiltonian flow is firstly studied by Nelson's mechanics (see e.g. [169, 178]). It is also pointed out that the Hamiltonian flows in density space are probability transition equations of classical Hamiltonian ODEs (see [167, 179] and references therein).

There are several works with titles related to Hamiltonian systems on graphs, like the port-Hamiltonian system on graphs (see e.g. [180, 181] and the references therein). Our current work is different from them. The port-Hamiltonian systems are the generalization of classical Hamiltonian system which describes the dynamics in interaction with control units, energy dissipating or energy storing units. The graph structure is used to characterize the interaction of the systems with ports, and their underlying phase variables are still in continuous spaces, like $\mathbb{R}^d$ or smooth manifold.

This chapter is organized as follows. In section 2, we use the discrete optimal transport problem as the motivation of studying the Hamiltonian process on finite graph. In section 3, we present the definition and several properties of the Hamiltonian process on graph. In section 4, we study several different Hamiltonian dynamics derived from the discrete SBP from two different perspectives. We also discuss the existence of stationary and periodic

Hamiltonian processes of the discrete SBP. We provide more examples of Hamiltonian process on graph in section 5.

5.2 Preliminary knowledge

In this section, we first briefly recall the relationship between the continuous OT problem and Hamiltonian systems. Then we introduce our motivation example on a graph and review some notations for inhomogeneous Markov process, which is used in our definition for Hamiltonian process.

It is known that in a continuous OT problem (Equation 5.1) with given marginal distributions ρ^0 and ρ^1, the optimal transfer $\{X_t\}_{t\in[0,1]}$ induces a trajectory concentrating on the geodesic path whose position and momentum obey the Hamiltonian principle (see e.g. [167]). More precisely, recalling that $H(x,v)=\frac{1}{2}|v|^2$, the critical point of the OT problem (Equation 5.1) in density manifold satisfies the Wasserstein–Hamiltonian flow,

$$\begin{aligned} \partial_t\rho+\nabla\cdot(\frac{\partial H}{\partial v}(x,\nabla S)\rho)=0,\\ \partial_t S+H(x,\nabla S)=C(t), \end{aligned} \tag{5.4}$$

where $C(t)$ is a function depending only on t and $v=\nabla S$ with $|\nabla S|^2=\nabla S\cdot\nabla S$. Being a Hamiltonian system on its own, (Equation 5.4) can also be connected to the following classic Hamiltonian system closely (see e.g. [171]):

$$\begin{aligned} d_t v=-\frac{\partial H}{\partial x}(X,v),\\ d_t X=\frac{\partial H}{\partial v}(X,v), \end{aligned} \tag{5.5}$$

where $X\in\mathbb{R}^d$, the conjugate momenta $v\in\mathbb{R}^d, d\in\mathbb{N}^+$, and the Hamiltonian H is smooth. If the initial position $X(0)$ is random following a distribution with density ρ^0, the trajectory X_t is random too. Its density function ρ, defined by the pushforward operator induced by the X_t, $\rho_t=X_t^\sharp\rho^0$, satisfies the Wasserstein-Hamiltonian flow (Equation 5.4).

However, directly mimicking the relationship between (Equation 5.4) and (Equation 5.5) is impossible if the underlying space become a graph. In the next subsection, we illustrate the challenges in detail by an example on graph.

5.2.1 A motivation example

Consider a graph $G = (V, E, \mathbf{W})$ with a node set $V = \{a_i\}_{i=1}^N$, an edge set E, and $w_{jl} \in \mathbf{W}$ are the weights of the edges: $w_{jl} = w_{lj} > 0$, if there is an edge between a_j and a_l, and 0 otherwise. Below, we write $(i, j) \in E$ to denote the edge in E between the vertices a_i and a_j. We assume that G is an undirected and connected graph with no self loops or multiple edges for simplicity. Let us denote the set of discrete probabilities on the graph by:

$$\mathcal{P}(G) = \{(\rho)_{j=1}^N \ : \ \sum_j \rho_j = 1, \rho_j \geq 0, \ \text{for } j \in V\},$$

and let $\mathcal{P}_o(G)$ be its interior (i.e., all $\rho_j > 0$, for $a_j \in V$). Inspired by [171, 182, 176], we consider the following discrete OT problem whose minimizer is the so-called geodesic random walk.

Example 5.2.1. *OT on G (geodesic random walk).*

The OT problem on a finite graph is related to the Wasserstein distance on $\mathcal{P}(G)$, which can be defined by the discrete Benamou–Brenier formula:

$$W(\rho^0, \rho^1) := \inf_{v,\rho} \Big\{ \sqrt{\int_0^1 \langle v, v \rangle_{\theta(\rho)} dt} \ : \ \frac{d\rho}{dt} + div_G^\theta(\rho v) = 0, \ \rho(0) = \rho^0, \ \rho(1) = \rho^1 \Big\}.$$

where $\rho^0, \rho^1 \in \mathcal{P}(G)$, $\rho \in H^1([0,1], \mathbb{R}^N)$ and v is a skew matrix valued function. The inner product of two vector fields u, v is defined by

$$\langle u, v \rangle_{\theta(\rho)} := \frac{1}{2} \sum_{(j,l) \in E} u_{jl} v_{jl} \theta_{jl} w_{jl}$$

with the weight θ_{ij} depending on ρ_i and ρ_j. The divergence of the flux function ρv is defined

as

$$(div_G^{\theta}(\rho v))_j := -(\sum_{l \in N(j)} w_{jl} v_{jl} \theta_{jl}), \tag{5.6}$$

where $N(i) = \{a_j \in V : (i,j) \in E\}$ *is the adjacency set of node* a_i*. Then its critical point* (ρ, v)*, with*

$$v = \nabla_G S := (S_j - S_l)_{(j,l) \in E} \tag{5.7}$$

for some function S *on* V*, satisfies the following discrete Wasserstein-Hamiltonian flow on the graph* G*:*

$$\begin{aligned} &\frac{d\rho_i}{dt} + \sum_{j \in N(i)} w_{ij}(S_j - S_i)\theta_{ij}(\rho) = 0, \\ &\frac{dS_i}{dt} + \frac{1}{2}\sum_{j \in N(i)} w_{ij}(S_i - S_j)^2 \frac{\partial \theta_{ij}(\rho)}{\partial \rho_i} = 0. \end{aligned} \tag{5.8}$$

We may view this equation as a discrete analog of (Equation 5.4)*. Consequently, its Hamiltonian only consists of the kinetic energy*

$$\mathscr{H}(\rho, S) = \frac{1}{4}\sum_{i,j}(S_i - S_j)^2 \theta_{ij}(\rho) w_{ij}.$$

As discussed in [176], the goal of the discrete OT problem is to find an optimal transport of the informal minimization problem

$$\inf_Q \Big\{ \int_0^T \frac{1}{2}\sum_{i,j}(v_{ij})^2 \theta_{ij} w_{ij} dt : d\rho_t = \rho_t Q_t dt,\ X(0) \sim \rho_0,\ X(T) \sim \rho_T \Big\}, \tag{5.9}$$

where the transition rate matrix Q_t *may be written as*

$$\begin{aligned} (Q_t)_{ii} &= \frac{1}{2}\sum_{j \in N(i)} w_{ij} \frac{\theta_{ij}(\rho)}{\rho_i} v_{ij}, \\ (Q_t)_{ji} &= -\frac{1}{2} w_{ji} \frac{\theta_{ji}(\rho)}{\rho_j} v_{ji}, \end{aligned}$$

if $\theta_{ij} = \theta_{ji}$. In [176], the minimizer of the above discrete OT problem is called the geodesic random walk which is defined as a random walk whose marginal probability is supported on the set of geodesic paths on $\mathcal{P}(G), i.e, X_t$ is determined by the marginal distribution and the instantaneous transition rate matrix Q_t. However, examining the transition rate matrix, we can find that the geodesic random walk X_t may not be well-defined, because there may not exist such a stochastic process due to possible negative probability and transition probability (See Remark 27 for more details).

This example illustrates that when compared to the continuous case, where the Hamiltonian system (Equation 5.5) on the phase space corresponds to the Hamiltonian PDEs (Equation 5.4) on Wasserstein manifold, such a correspondence in discrete space can't be easily established, because the counterpart of (Equation 5.5) requires more careful treatments.

5.2.2 Inhomogeneous Markov process

In order to define a stochastic process which plays the role of the Hamiltonian mechanics (Equation 5.5) on a finite graph, we recall the definition of the **inhomogeneous Markov process** in [183]. The linear master equation

$$\frac{d\rho}{dt} = \rho Q$$

determines a linear Markov process. When $Q = Q(t)$, it corresponds to a time inhomogeneous Markov process. Here $Q(t)$ is a family of infinitesimal generators of the stochastic matrix or Kolmogorov matrix, namely, a square matrix which has non-positive (resp. non-negative) elements on the main diagonal (resp. off the main diagonal), and the sum of each row is zero. Among different types of inhomogeneous Markov process, the **nonlinear Markov processes** [183] whose transition rate matrix Q may depend not only on the current state x of the process but also on the current distribution ρ of the process is of

particular interest to us.

Given an initial distribution ρ_0, a time inhomogeneous Markov process $\{X_t\}_{t\geq 0}$ can be defined as a process which has ρ_0 as the distribution of X_0 and $(s,t) \to P_{s,t}$ as its transition mechanism in the sense that

$$\mathbb{P}(X_0 = a_i) = \rho_i,\ \mathbb{P}(X_t = a_j | X_\sigma, \sigma \in [0,s]) = (P_{s,t})_{X(s)a_j},$$

where $(P_{s,t})_{a_i a_j} = \mathbb{P}(X_t = a_j | X_s = a_i)$. The corresponding forward Kolmogorov equation can be rewritten as

$$d_t P_{s,t} = P_{s,t} Q_t.$$

If $t \in [s,\infty) \mapsto \rho_t$ is continuously differentiable, then $\dot{\rho}_t = \rho_t Q_t$ is equivalent to $\rho_t = \rho_s P_{s,t}$ for $t \geq s$. Given $\{Q_t\}_{t\geq 0}$, ρ_0, there exists an inhomogeneous Markov process X_t related to the transition rate matrix Q_t and the marginal distribution ρ_t. On the other hand, given an inhomogeneous Markov process with transition matrices $P_{s,t}$, it will induce the equation of ρ with Q_t (see e.g. [183]).

5.3 Hamiltonian process on a finite graph

As shown in Example Example 5.2.1, although it may not be possible to find a stochastic process for every discrete optimal transport problem, it reveals two key features that the density of such a stochastic process, if exists, satisfies the generalized master equation and that its Q_t-matrix is determined by a potential S_t, where S_t satisfies a discrete Hamiltonian Jacobi equation. Inspired by these properties, we introduce the definition of stochastic Hamiltonian process.

Definition 5.3.1. *A stochastic process $\{X_t\}_{t\geq 0}$ is called a Hamiltonian process on the graph if*

1. *The density ρ_t of X_t satisfies the following generalized Master equation,*

$$d_t\rho_t = \rho_t Q_t(v, \rho_t),$$

with

$$(Q_t(v,\rho))_{ij} = w_{ji} f_{ji}(v_{ji}, \rho, t), (Q_t(v))_{ii} = -\sum_{j\in N(i)} w_{ij} f_{ji}(v_{ji}, \rho, t),$$

where the skew-matrix v is induced by a potential function S, i.e., $v = \nabla_G S + u$, with $\rho_t Q_t(\nabla_G S, \rho_t) = \rho_t Q_t(v, \rho_t)$. And $f_{ji} : \mathbb{R} \times \mathbb{R}^+ \cup \{0\} \times \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, is a real-valued measurable function which is piece-wise continuous in the first component $x \in \mathbb{R}$.

2. *The density ρ and the potential S form a Hamiltonian system on the cotangent bundle of the density space.*

The following theorem gives the structure of the Hamiltonian on the density manifold of the Hamiltonian process.

Theorem 5.3.1. *Suppose that the stochastic process $\{X_t\}_{t\geq 0}$ with density $\{\rho_t\}_{t\geq 0}$ and potential $\{S_t\}_{t\geq 0}$ defined in Definition 5.3.1 forms a Hamiltonian process on the graph G. In addition assume that the antiderivative F_{ij} of f_{ij} exists for $ij \in E$. Then the Hamiltonian always have the form*

$$\mathscr{H}(\rho, S) = \sum_{i\in V} \sum_{j\in N(i)} \rho_i F_{ji}(S_j - S_i, \rho, t) w_{ji} + \mathcal{V}(\rho, t) \tag{5.10}$$

where $\mathcal{V}$ is a function depending ρ and t. Moreover, the Hamiltonian system on the cotan-

gent bundle of $\mathcal{P}(G)$ can be written as:

$$\frac{\partial}{\partial t}\rho_i(t) = \sum_{j\in N(i)} w_{ij}f_{ij}(S_i - S_j, \rho, t)\rho_j - w_{ji}f_{ji}(S_j - S_i, \rho, t)\rho_i \tag{5.11}$$
$$\frac{\partial}{\partial t}S_i(t) = -\sum_{j\in N(i)} \left(w_{ji}F_{ji}(S_j - S_i, \rho, t) + \rho_j \frac{\partial}{\partial \rho_i} F_{ji}(S_j - S_i, \rho, t) w_{ji} \right) - \frac{\partial}{\partial \rho_i}\mathcal{V}(\rho, t).$$

Proof. According to Definition 5.3.1, we have $\frac{\partial}{\partial t}\rho_i(t) = \sum_{j\in N(i)} w_{ij}f_{ij}(S_i - S_j, \rho, t)\rho_j - w_{ji}f_{ji}(S_j - S_i, \rho, t)\rho_i$. Since $\{\rho_t, S_t\}$ forms a Hamiltonian system, we are able to state

$$\frac{\partial}{\partial S_i}\mathscr{H}(\rho, S, t) = \sum_{j\in N(i)} w_{ij}f_{ij}(S_i - S_j, \rho, t)\rho_j - w_{ji}f_{ji}(S_j - S_i, \rho, t)\rho_i, \quad i \in V.$$

Considering the following quantity,

$$\mathscr{H}_0(\rho, S, t) = \sum_{i\in V}\sum_{j\in N(i)} \rho_i F_{ji}(S_j - S_i, \rho, t)w_{ji},$$

we can directly verify that $\frac{\partial}{\partial S}(\mathscr{H}(\rho, S, t) - \mathscr{H}_0(\rho, S, t)) = 0$. This suggests that there exists some function $\mathcal{V}$ depending on ρ and t such that $\mathscr{H}(\rho, S, t) - \mathscr{H}_0(\rho, S, t) = \mathcal{V}(\rho, t)$. This directly leads to form of Hamiltonian $\mathscr{H}(\rho, S, t) = \sum_{i\in V}\sum_{j\in N(i)} \rho_i F_{ji}(S_j - S_i, \rho, t)w_{ji} + \mathcal{V}(\rho, t)$. Furthermore, the discrete Hamiltonian Jacobi equation is derived as

$$\frac{\partial}{\partial t}S_t = -\frac{\partial}{\partial \rho}\mathscr{H}(\rho, S, t).$$

□

As a direct consequence, we have the following properties of Hamiltonian process.

Proposition 5.3.1 (Properties of Hamiltonian process)**.** *Assume that a stochastic process X_t on a finite graph is a Hamiltonian process. Then it holds that*

1. *there exists a Hamiltonian $\mathscr{H}$ on the density space such that its marginal distribution $\rho_t = X_t^\sharp \rho_0$ and the generator S_t of the transition rate matrix Q_t forms a Hamiltonian*

system;

2. *the symplectic structure on the density space is preserved, i.e.,*

$$\omega_{g(\rho,S)}(g'(\rho,S)\xi, g'(\rho,S)\eta) = \omega_{(\rho,S)}(\xi,\eta),$$

where ω denotes the symplectic form on $\mathcal{T}^*\mathcal{P}(G)$, $\xi,\eta \in \mathcal{T}_{(\rho,S)}(\mathcal{T}^*P(G))$ and $g'(\rho,S)$ is the Jacobi matrix of the Hamiltonian flow on the density space;

3. *$\mathscr{H}(t) = \mathscr{H}(0)$, if the Hamiltonian $\mathscr{H}$ is independent of t;*

4. *and X_t is mass-preserving, i.e., $\sum_{i=1}^N \rho_i(t) = \sum_{i=1}^N \rho_i(0)$.*

Remark 26 (Particle-level properties of Hamiltonian process)**.** *Consider the Hamiltonian with specific form*

$$\mathscr{H}(\rho,S) = \sum_{i\in V}\sum_{j\in N(i)} \rho_j F_{ji}(S_j - S_i)w_{ji} + \sum_{i\in V}\rho_i V_i.$$

Suppose that $\{X(t)\}$ is a Hamiltonian process on G associated to the Hamiltonian $\mathscr{H}$. Then one can verify $\mathbb{E}[H(X(t),S(t))]$ with

$$H(X(t),S(t)) = \sum_{j\in N(X(t))} F_{jX(t)}(S_j(t) - S_{X(t)}(t))w_{jX(t)} + V_{X(t)}.$$

remains constant as time t evolves.

Based on the definition of Hamiltonian process, we are able to construct the discrete optimal transport problem which retains the property that the minimizer is a stochastic process on the graph for Example 5.2.1.

Proposition 5.3.2. *There always exists a density dependent weight θ such that the geodesic random walk in Example 5.2.1 is a Hamiltonian process.*

Proof. Define $\theta_{ij}^U = \theta_S^U(\rho_i, \rho_j)$, where $\theta_S^U(\rho_i, \rho_j) = \rho_i$ if $S_j > S_i$. Denote $(x)^+ = \max(0, x), (x)^- = \min(0, x)$. Using the notations in Example Example 5.2.1, the geodesic random walk on G with the probability weight $\theta = \theta^U$ satisfies

$$d\rho_i = \sum_{j \in N(i)} w_{ij}(v_{ij})^+ \rho_j + \sum_{j \in N(i)} w_{ij}(v_{ij})^- \rho_i. \tag{5.12}$$

From the discrete Hodge decomposition on the graph [171], for any skew matrix v and probability density $\rho \in \mathcal{P}_o(G)$, there exists a decomposition $v = \nabla_G S + u$ with $\text{div}_G^\theta(\rho u) = 0$. Here $(\nabla_G S)_{ij} := (S_i - S_j)$ and $\text{div}_G^\theta(\rho u) := -(\sum_{j \in N(i)} u_{ij} \theta_{ij}^U(\rho))$. To see this fact, it suffices to prove that there exists a unique solution of S such that $\text{div}_G^\theta(\rho \nabla_G S) = \text{div}_G^\theta(\rho v)$. The connectivity of the graph and the fact that $\rho \in \mathcal{P}_o(G)$ implies that if

$$\langle \text{div}_G^\theta(\rho \nabla_G S), S \rangle = \frac{1}{2} \sum_{(i,j) \in E} ((S_i - S_j)^-)^2 \theta_{ij}(\rho) = 0,$$

then 0 must be a simple eigenvalue of $\text{div}_G^\theta(\rho \nabla_G)$ with eigenvector $(1, \cdots, 1)$. Thus S is unique up to a constant shift and the skew matrix $v_t = \nabla_G S_t + u$ satisfies

$$d(S_t)_i = -\frac{1}{2} \sum_{j \in N(i)} w_{ij}((S_i - S_j)^-)^2 + C(t), \ \text{div}_G^\theta(\rho u) = 0,$$

where $C(t)$ is independent of nodes. Meanwhile, f_{ij} can be selected to achieve $f_{ij}(S_i - S_j) = (S_i - S_j)^+$ and thus

$$\begin{aligned}
&(Q_t)_{ii} = \sum_{j \in N(i)} w_{ij}(S_i - S_j)^- = \sum_{j \in N(i)} w_{ij} f_{ji}(S_j - S_i), \\
&(Q_t)_{ji} = w_{ji}(S_j - S_i)^+ = w_{ji} f_{ji}(S_j - S_i), \ ij \in E, \ \text{otherwise } Q_{ji} = 0.
\end{aligned}$$

We can define a time inhomogenous Markov process as follows by the transition matrix

$\mathbb{P}(X_t = v_j | X_\tau, \tau \in [0,s]) = (P_{s,t})_{X(s)v_j}$. Given the past $\sigma(\{X_\tau : \tau \in [0,t]\})$ of X up to time $t \geq 0$, the probability of its having moved away from X_t at the time $t+h$ with h small enough can be approximated by $1 - (Q_t)_{X_t X_t} h$, i.e.,

$$\left| \mathbb{P}(X(t+h) = X_t | X_\tau, \tau \leq t) - 1 - (Q_t)_{X_t X_t} h \right| = o(h).$$

Here $\{-(Q_t)_{ii}\}_i$ is often called as the transition rate of X_t. Given the history that the jump appeared $\sigma(\{X_\tau : \tau \in [0,t]\} \cup \{X_{t+h} \neq X_t\})$, the probability that $X_{t+h} = a_j$ is approximately $(P_{t,t+h})_{X_t a_j}$, which implies that

$$\left| \mathbb{P}(X(t+h) = a_j | X_\tau, \tau \leq t) - h(Q_t)_{X_t a_j} \right| = o(h).$$

□

Remark 27. *It is worth mentioning that the Hamiltonian system on $\mathcal{P}(G)$ does not necessarily induce a stochastic process on G. This can also be illustrated by using the optimal transport problem introduced in Example Example 5.2.1. Let us take $w_{ij} = 1$ if $ij \in E$ for simplicity. In order to define a Hamiltonian process on G, the probability weight θ can not be chosen arbitrarily here. For example, if we take the probability weight $\theta_{ij} = \theta^A(\rho_i, \rho_j) = \frac{1}{2}(\rho_i + \rho_j)$ in [171], the density equation can be rewritten as*

$$d_t \rho_t = \rho_t Q_t,$$

where

$$(Q_t)_{ii} = \frac{1}{2} \sum_{j \in N(i)} (S_i - S_j),$$
$$(Q_t)_{ij} = \frac{1}{2}(S_j - S_i),\ ij \in E,\ \textit{otherwise } Q_{ij} = 0.$$

The function $f_{ij}(x) = \frac{1}{2}x$.

When $\theta_{ij} = \theta^L(\rho_i, \rho_j) = \frac{\rho_i - \rho_j}{\log(\rho_i) - \log(\rho_j)}$ in [27], the density equation can be rewritten as

$$d_t \rho_t = \rho_t Q_t,$$

where

$$(Q_t)_{ii} = \sum_{j \in N(i)} \frac{(S_i - S_j)}{\log(\rho_i) - \log(\rho_j)},$$
$$(Q_t)_{ij} = -\frac{(S_i - S_j)}{\log(\rho_i) - \log(\rho_j)},\ ij \in E,\ \textit{otherwise } Q_{ij} = 0.$$

The function $f_{ij}(x) = \frac{x}{\log(\rho_i) - \log(\rho_j)}$.

In both cases, there is no guarantee that the off-diagonal of Q_t is non-positive. Hence, Q_t is unable to admit a stochastic process X_t which is time inhomogeneous Markov due to the appearance of negative transition probabilities. For valid choices of θ that may admit stochastic processes, we refer to [27], [166] and references therein.

Remark 28. *If $\theta_{ij} > 0$ for all $ij \in E$, then the Hodge decomposition yield a unique potential S up to a constant which induces v. If there exists $ij \in E$ such that $\theta_{ij} = 0$, then the generator S may be not unique. Meanwhile, the Hamiltonian Jacobi equation may become one-side inequality,*

$$v_{ij} = S_i - S_j,\ \partial_t S_i + \frac{\partial}{\partial \rho_i} \mathscr{H}(\rho, S) \leq 0.$$

Remark 29. *The initial value problem of the Hamiltonian system of ρ, S may develop singularity at a finite time $T > 0$, i.e, either $\lim_{t \to T} S_i(t) = \infty$ or $\lim_{t \to T} \rho_i \leq 0$.*

We would like to emphasize that a Hamiltonian process is not Markov in general. The sufficient and necessary conditions when a Hamiltonian process gives a Markov process are presented as follows.

Theorem 5.3.2. *Given a Hamiltonian process $\{X_t\}_{t\geq 0}$ on the graph with a Hamiltonian $\mathscr{H}(\rho,S)=\sum_{i=1}^{N}\sum_{j\in N(i)}F_{ij}(\rho,S)w_{ij}\rho_i$. If X_t is a Markov process, then (ρ,S) in Definition* (Definition 5.3.1) *satisfies the following system,*

$$\begin{aligned}
&\frac{\partial^2 F_{ij}}{\partial S_i\partial\rho_i}\Big(\sum_{l\in N(i)}\frac{\partial F_{il}}{\partial S_i}\rho_i w_{il}+\sum_{l\in N(i)}\frac{\partial F_{li}}{\partial S_i}\rho_l w_{li}\Big) \\
&+\frac{\partial^2 F_{ij}}{\partial S_i\partial\rho_j}\Big(\sum_{k\in N(j)}\frac{\partial F_{jk}}{\partial S_j}\rho_j w_{jk}+\sum_{k\in N(j)}\frac{\partial F_{kj}}{\partial S_j}\rho_l w_{kj}\Big) \\
&-\frac{\partial^2 F_{ij}}{\partial S_i\partial S_i}\Big(\sum_{l\in N(i)}\frac{\partial F_{il}}{\partial\rho_i}\rho_i w_{il}+\sum_{l\in N(i)}\frac{\partial F_{li}}{\partial\rho_i}\rho_l w_{li}+\sum_{l\in N(i)}(F_{il}w_{il}+F_{li}w_{li})\Big) \\
&-\frac{\partial^2 F_{ij}}{\partial S_i\partial S_j}\Big(\sum_{k\in N(j)}\frac{\partial F_{jk}}{\partial\rho_j}\rho_j w_{jk}+\sum_{k\in N(j)}\frac{\partial F_{kj}}{\partial\rho_j}\rho_k w_{kj}+\sum_{k\in N(j)}(F_{jk}w_{jk}+F_{ki}w_{ki})\Big)=0
\end{aligned} \tag{5.13}$$

for $i,j\in V$. Conversely, if (ρ,S) satisfies (Equation 5.13)*, then there exists a Markov process which is Hamiltonian.*

Proof. Since X_t is a Hamiltonian process, the transition matrix is determined by $\rho_t Q_t=\frac{\partial H}{\partial S}=d_t\rho_t$. This implies that

$$(\rho_t Q_t)_i=\sum_{j\in N(i)}\frac{\partial F_{ij}(\rho,S)}{\partial S_i}w_{ij}\rho_i+\sum_{j\in N(i)}\frac{\partial F_{ji}(\rho,S)}{\partial S_i}w_{ji}\rho_j.$$

Therefore, $(Q_t)_{ii}=\sum_{j\in N(i)}\frac{\partial F_{ij}(\rho,S)}{\partial S_i}w_{ij}$, $(Q_t)_{ij}=\frac{\partial F_{ij}(\rho,S)}{\partial S_j}w_{ij}$. Since X_t preserves the mass, it holds that $\sum_{j\in N(i)}(\frac{\partial F_{ij}(\rho,S)}{\partial S_i}+\frac{\partial F_{ij}(\rho,S)}{\partial S_j})w_{ij}=0$ for every $i\leq N$.

Notice that X_t is Markov implies that $d_tQ_{ij}=0$, for $i,j\leq N$, that is

$$d_t\frac{\partial F_{ij}}{\partial S_i}=0, d_t\frac{\partial F_{ji}}{\partial S_j}=0.$$

Direct calculation leads to

$$
\begin{aligned}
d_t\frac{\partial F_{ij}}{\partial S_i} &= \frac{\partial^2 F_{ij}}{\partial S_i \partial \rho_i} d_t\rho_i + \frac{\partial^2 F_{ij}}{\partial S_i \partial \rho_j} d_t\rho_j \\
&+ \frac{\partial^2 F_{ij}}{\partial S_i \partial S_i} d_t S_i + \frac{\partial^2 F_{ij}}{\partial S_i \partial S_j} d_t S_j \\
&= \frac{\partial^2 F_{ij}}{\partial S_i \partial \rho_i}\Big(\sum_{l\in N(i)} \frac{\partial F_{il}}{\partial S_i}\rho_i w_{il} + \sum_{l\in N(i)} \frac{\partial F_{li}}{\partial S_i}\rho_l w_{li}\Big) \\
&+ \frac{\partial^2 F_{ij}}{\partial S_i \partial \rho_j}\Big(\sum_{k\in N(j)} \frac{\partial F_{jk}}{\partial S_j}\rho_j w_{jk} + \sum_{k\in N(j)} \frac{\partial F_{kj}}{\partial S_j}\rho_l w_{kj}\Big) \\
&- \frac{\partial^2 F_{ij}}{\partial S_i \partial S_i}\Big(\sum_{l\in N(i)} \frac{\partial F_{il}}{\partial \rho_i}\rho_i w_{il} + \sum_{l\in N(i)} \frac{\partial F_{li}}{\partial \rho_i}\rho_l w_{li} + \sum_{l\in N(i)} (F_{il}w_{il} + F_{li}w_{li})\Big) \\
&- \frac{\partial^2 F_{ij}}{\partial S_i \partial S_j}\Big(\sum_{k\in N(j)} \frac{\partial F_{jk}}{\partial \rho_j}\rho_j w_{jk} + \sum_{k\in N(j)} \frac{\partial F_{kj}}{\partial \rho_j}\rho_k w_{kj} + \sum_{k\in N(j)} (F_{jk}w_{jk} + F_{ki}w_{ki})\Big),
\end{aligned}
$$

which yields the desired result. Conversely, if (ρ, S) satisfies (Equation 5.13), the previous arguments leads to the equation of ρ becomes a linear Master equation. Then there always exists a Markov process which is a stochastic representation of linear Master equation. Meanwhile, it can be verified that this Markov process satisfies all the conditions in Definition 5.3.1 and is Hamiltonian. □

Corollary 5.3.2.1. *Given a Hamiltonian* $\mathscr{H}(\rho, S) = \sum_{i=1}^{N}\sum_{j\in N(i)} F_{ij}(\rho, S)w_{ij}\rho_i$. *Assume that there exists* $(\rho^*, S^*(t))$ *satisfies the following conditions,*

1. $\sum_{j\in N(i)} \frac{\partial F_{ij}(\rho,S)}{\partial S_i} + \frac{\partial F_{ij}(\rho,S)}{\partial S_j} = 0,$

2. ρ^* *is independent of* t *and* $(\rho^*, S^*(t))$ *solves*

$$
\begin{aligned}
&\sum_{l\in N(i)} \frac{\partial F_{il}}{\partial S_i}\rho_i w_{il} + \sum_{l\in N(i)} \frac{\partial F_{li}}{\partial S_i}\rho_l w_{li} = 0, \\
&\frac{\partial^2 F_{ij}}{\partial S_i \partial S_i}\Big(\sum_{l\in N(i)} \frac{\partial F_{il}}{\partial \rho_i}\rho_i w_{il} + \sum_{l\in N(i)} \frac{\partial F_{li}}{\partial \rho_i}\rho_l w_{li} + \sum_{l\in N(i)} (F_{il}w_{il} + F_{li}w_{li})\Big) \\
&+ \frac{\partial^2 F_{ij}}{\partial S_i \partial S_j}\Big(\sum_{k\in N(j)} \frac{\partial F_{jk}}{\partial \rho_j}\rho_j w_{jk} + \sum_{k\in N(j)} \frac{\partial F_{kj}}{\partial \rho_j}\rho_k w_{kj} + \sum_{k\in N(j)} (F_{jk}w_{jk} + F_{ki}w_{ki})\Big) = 0
\end{aligned}
$$

Then there exists a Hamiltonian process which is Markov and preserves the mass. Furthermore, the Hamiltonian process is invariant with respect to ρ^*.

5.4 Hamiltonian process via discrete SBP on graphs

Although the SBP [168] has a history close to 100 years, it has received revived attention from control theory and machine learning communities recently, see [174, 173]. For convenience, the background of continuous SBP is presented in the appendix.

For the discrete counterpart of SBP on graph, there are two different treatments reported in the literature.

1. One is to consider a reference path measure R (induced by a reversible random walk) on the graph and then study the optimization problem involving the relative entropy between the reference measure R and the path measure P with given initial and terminal distributions [174, 176].) In this framework, the reference random walk is often related to a discrete version of (Equation D.5) (For example, the linear discretization of the Laplacian gives the time homogenous Markov chain as the reference in [184]).

2. Another way is proposed by the discrete version of (Equation D.2) or (Equation D.4) directly [177].

We shall show that different treatments create differences on the structure and formulation of equations, in particular the discrete Laplacian operator. Each of these formulations can determine its corresponding Hamiltonian process on graph.

5.4.1 Discrete SBP based on relative entropy and reference Markov measure

In the following discussion, we always assume that $w_{ij} = 1$ if $ij \in E$ for conciseness of formulations. By using the discrete Girsanov theorem on graph, the discrete SBP in the

form of relative entropy (A) becomes the following control problem

$$\min_{\widehat{m}^t \geq 0} \left\{ \int_0^1 \sum_{i \in V} \rho(i,t) \sum_{j \in N(i)} \left(\frac{\widehat{m}_{ij}^t}{m_{ij}^t} \log \left(\frac{\widehat{m}_{ij}^t}{m_{ij}^t} \right) - \frac{\widehat{m}_{ij}^t}{m_{ij}^t} + 1 \right) m_{ij}^t \, dt \right\} \tag{5.14}$$
$$\text{subject to:} \quad \frac{d}{dt}\rho(i,t) = \sum_{j \in N(i)} \widehat{m}_{ji}^t \rho_j - \widehat{m}_{ij}^t \rho_i \quad \rho(\cdot,0) = \rho^0, \ \rho(\cdot,1) = \rho^1.$$

where the reference measure R is determined by the master equation $d_t \widetilde{\rho}_i = \sum_{j \in N(i)} m_{ji}^t \widetilde{\rho}_j - m_{ij}^t \widetilde{\rho}_i$.

Remark 30. *The formula for relative entropy between path measure P and reference path measure R is formulated as*

$$\mathcal{H}(P|R) = \int_0^1 \sum_{i \in V} \rho(i,t) \sum_{j \in N(i)} \left(\frac{\widehat{m}_{ij}^t}{m_{ij}^t} \log \left(\frac{\widehat{m}_{ij}^t}{m_{ij}^t} \right) - \frac{\widehat{m}_{ij}^t}{m_{ij}^t} + 1 \right) m_{ij}^t \, dt.$$

This result is provided in [174],[176]. A rigorous proof for this formula originates from Theorem 2.9 of [185].

Let us denote $u(x) = x \log x - x + 1$. By introducing Lagrange multiplier ψ, we obtain the following Lagrangian functional

$$\begin{aligned}
\mathscr{L}(\rho, \widehat{m}, \psi) &= \int_0^1 \sum_{i \in V} \rho(i,t) \sum_{j \in N(i)} u \left(\frac{\widehat{m}_{ij}^t}{m_{ij}^t} \right) m_{ij}^t \, dt \\
&+ \int_0^1 \sum_{i \in V} -\rho(i,t) \frac{\partial}{\partial t} \psi(i,t) - \psi(i,t) \left(\sum_{j \in N(i)} \widehat{m}_{ji}^t \rho_j - \widehat{m}_{ij}^t \rho_i \right) \, dt \\
= \int_0^1 &- \sum_{i \in V} \rho(i,t) \frac{\partial}{\partial t} \psi(i,t) - \frac{1}{2} \sum_{(i,j) \in E} \left[\frac{\widehat{m}_{ji}}{m_{ji}} (\psi(i,t) - \psi(j,t)) - u(\frac{\widehat{m}_{ji}}{m_{ji}}) \right] m_{ji} \rho(j,t) \\
&+ \left[\frac{\widehat{m}_{ij}}{m_{ij}} (\psi(j,t) - \psi(i,t)) - u(\frac{\widehat{m}_{ij}}{m_{ij}}) \right] m_{ij} \rho(i,t) \, dt.
\end{aligned}$$

When solving the above saddle point problem, we minimize over $\widehat{m}$ and get

$$\int_0^1 -\sum_{i\in V}\rho(i,t)\frac{\partial}{\partial t}\psi(i,t) - \frac{1}{2}\sum_{(i,j)\in E}[u^*(\psi(i,t)-\psi(j,t))m_{ji}\rho(j,t) + u^*(\psi(j,t)-\psi(i,t))m_{ij}\rho(i,t)]\,dt.$$

Here u^* is the Legendre dual of u: $u^*(x) = \sup_y\{x\cdot y - u(y)\}$, leading to $u^*(x) = e^x - 1$. By formulating the Lagrangian, we can identify the Hamiltonian of this control problem, which can be written as:

$$\mathscr{H}(\rho,\psi) = \sum_{i\in V}\sum_{j\in N(i)}(\exp(\psi(j,t)-\psi(i,t))-1)m_{ij}\rho(i,t). \tag{5.15}$$

Then the above control problem implies the following Hamiltonian system

$$\partial_t\rho = \frac{\partial\mathscr{H}(\rho,\psi)}{\partial\psi}, \quad \partial_t\psi = -\frac{\partial\mathscr{H}(\rho,\psi)}{\partial\rho},$$

that is,

$$\begin{aligned}\frac{\partial}{\partial t}\rho(i,t) &= \sum_{j\in N(i)} -e^{\psi(j,t)-\psi(i,t)}m_{ij}\rho(i,t) + e^{\psi(i,t)-\psi(j,t)}m_{ji}\rho(j,t), \\ \frac{\partial}{\partial t}\psi(i,t) &= -\sum_{j\in N(i)}(e^{\psi(j,t)-\psi(i,t)}-1)m_{ij}.\end{aligned} \tag{5.16}$$

By using the Hopf-Cole transform, we can further verify that the discrete SBP problem determines a Hamiltonian process on the graph. Let us consider the following transform $\tau : \mathcal{T}^*\mathcal{P}(G) \to \mathcal{T}^*\mathcal{P}(G)$ as:

$$\tau[(\rho,\psi)] = (\rho, \psi - \frac{1}{2}\ln\rho) \tag{5.17}$$

Let us denote $g'(\rho,\psi) = D\tau(\rho,\psi)$. Then the symplectic form ω is unchanged in the sense

that

$$\omega_{g(\rho,\psi)}(g'(\rho,\psi)\xi, g'(\rho,\psi)\eta) = \omega_{(\rho,\psi)}(\xi,\eta),$$

where $(\xi,\eta) \in T_{(\rho,\psi)}\mathcal{T}^*\mathcal{P}(G)$. By using the symplectic submersion from $\mathcal{P}(G)$ to $\mathbb{R}^N$, the symplectic form can be represented by $(g'(\rho,\psi)\xi)^T J g'(\rho,\psi)\eta = \xi^T J \eta$, where J is the standard symplectic matrix. Since $d_t\tau(\rho,\psi)^T = \tau' d_t(\rho,\psi)^T$ and that $(\tau')^T J \tau' = J$, we conclude that the Hopf–Cole transformation (Equation 5.17) is a symplectic transformation on the cotangent bundle of the density manifold. Denote (ρ, S) as the new coordinate. Then $\{\rho_t, S_t\}$ satisfies the following Hamiltonian system:

$$\begin{aligned}\frac{\partial \rho(i,t)}{\partial t} &= \frac{\partial \tilde{\mathscr{H}}(\rho,S)}{\partial S}\\ \frac{\partial S(i,t)}{\partial t} &= -\frac{\partial \tilde{\mathscr{H}}(\rho,S)}{\partial \rho}\end{aligned}$$

with

$$\tilde{\mathscr{H}}(\rho,S) = \mathscr{H}(\tau^{-1}(\rho,S)) = \sum_{i\in V}\sum_{j\in N(i)} e^{(S_j - S_i)} m_{ij}\sqrt{\rho_i\rho_j}, \tag{5.18}$$

that is

$$\begin{aligned}\frac{\partial S(i,t)}{\partial t} &= -m_{ii} - \frac{1}{2}\sum_{j\in N(i)} e^{S_j - S_i} m_{ij}\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}} - \frac{1}{2}\sum_{j\in N(i)} e^{S_i - S_j} m_{ji}\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}},\\ \frac{\partial \rho(i,t)}{\partial t} &= \sum_{j\in N(i)} e^{S_i - S_j} m_{ji}\sqrt{\rho_j}\sqrt{\rho_i} - \sum_{j\in N(i)} e^{S_j - S_i} m_{ij}\sqrt{\rho_i}\sqrt{\rho_j}.\end{aligned} \tag{5.19}$$

As a consequence, we verify that, as reported in [174], the discrete SBP corresponds to a Hamiltonian process with the transition rate matrix Q ($Q_{ij} = \widehat{m}_{ij}$) defined by $Q_{ii} = -\sum_{j\in N(i)} e^{S_j - S_i}\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}} m_{ij}$, $Q_{ij} = e^{S_i - S_j}\frac{\sqrt{\rho_i}}{\sqrt{\rho_j}} m_{ji}$ if $ij \in E$.

Using the above procedures, we can naturally extend the original SBP problem to the following generalized control problem

$$\min_{\hat{m}^t \geq 0} \left\{ \int_0^1 \sum_{i \in V} \rho(i,t) \sum_{j \in N(i)} u\left(\frac{\hat{m}_{ij}^t}{m_{ij}^t}\right) m_{ij}^t \, dt \right\} \tag{5.20}$$
$$\text{subject to:} \quad \frac{d}{dt}\rho(i,t) = \sum_{j \in N(i)} \hat{m}_{ji}^t \rho_j - \hat{m}_{ij}^t \rho_i \quad \rho(\cdot,0) = \rho^0, \ \rho(\cdot,1) = \rho^1.$$

Here u is an arbitrary convex function. Then the Hamiltonian associated with this general control problem is

$$\mathscr{H}(\rho, \psi) = \sum_{i \in V} \sum_{j \in N(i)} u^*(\psi(j,t) - \psi(i,t)) m_{ij} \rho(i,t), \tag{5.21}$$

where $\lambda_{ij} = (u')^{-1}(\psi_j - \psi_i)$.

For the sake of completeness of our discussion, we also reveal the relations among the so-called Schrödinger system [179, 186, 187] and our derived systems (Equation 5.16) and (Equation 5.19). All three PDE systems are derived from the SBP. We introduce the Madelung Transform $\phi : \mathcal{T}^*\mathcal{P}(G) \to \mathcal{T}^*\mathcal{P}(G)$

$$(f, g) = \phi(\rho, S) = (\sqrt{\rho} e^{-S}, \sqrt{\rho} e^{S}), \tag{5.22}$$

or equivalently,

$$(f, g) = \tilde{\phi}(\rho, \psi) = (\rho e^{-\psi}, e^{\psi}). \tag{5.23}$$

Combining (Equation 5.22) with (Equation 5.19), or combining (Equation 5.23) with (Equation 5.16) yields the Schrödinger system:

$$\begin{aligned} \frac{\partial}{\partial t} f(i,t) &= \sum_{j \in N(i)} (f(j,t) - f(i,t)) m_{ij}^t, \\ \frac{\partial}{\partial t} g(i,t) &= -\sum_{j \in N(i)} (g(j,t) - g(i,t)) m_{ij}^t. \end{aligned} \tag{5.24}$$

Similar to our previous analysis, we can verify that both transforms ϕ and $\tilde{\phi}$ preserves the symplectic form. And we know that (Equation 5.24) is a Hamiltonian system and its corresponding Hamiltonian is

$$\widehat{\mathscr{H}}(f,g)=\sum_{i\in V}\sum_{j\in N(i)}f_ig_jm_{ij}^t.$$

By applying Theorem 5.3.2, we obtain the following result about the conditions under which the Hamiltonian process in SBP enjoys the stationary measure and Markov property.

Proposition 5.4.1. *Assume that the reference process is mass-preserving, i.e.,* $\sum_i\widetilde{\rho}(i,t)=\sum_i\widetilde{\rho}^0(i)$*, and possesses a stationary measure* ρ^**. Then there exists a stationary point* (ρ^*,S^*) *of the Hamiltonian system* (Equation 5.19) *on the density manifold.*

Proof. Take $\frac{\partial\tilde{H}}{\partial S}=0$ and $\frac{\partial\tilde{H}}{\partial\rho}=0$ such that (ρ,S) is independent of time. The equation of ρ leads to

$$\sum_{j\in N(i)}e^{S_i-S_j}m_{ji}\sqrt{\rho_j}=\sum_{j\in N(i)}e^{S_j-S_i}m_{ij}\sqrt{\rho_j}.$$

Due to $m_{ii}=-\sum_{j\in N(i)}m_{ij}$, the equation of S becomes

$$\frac{1}{2}\sum_{j\in N(i)}(e^{S_i-S_j}m_{ji}+e^{S_j-S_i}m_{ij})\sqrt{\rho_j}=\sum_{j\in N(i)}m_{ij}\sqrt{\rho_i}.$$

Applying the above relationships, we obtain that

$$\sum_{j\in N(i)}e^{S_i-S_j}m_{ji}\sqrt{\rho_j}=\sum_{j\in N(i)}m_{ij}\sqrt{\rho_i}.$$

This immediately implies that

$$\sum_{j\in N(i)}e^{-S_j}m_{ji}\sqrt{\rho_j}=\sum_{j\in N(i)}e^{-S_i}m_{ij}\sqrt{\rho_i}.$$

Now by taking $e^{S_j^*}\sqrt{\rho_j^*} = e^{S_i^*}\sqrt{\rho_i^*}$ for all $ij \in E$, the first equation is reduced to

$$\sum_{j\in N(i)} m_{ji}\rho_j^* = \sum_{j\in N(i)} m_{ij}\rho_i^*.$$

This leads to

$$\sum_{j\in N(i)} m_{ji}\rho_j^* + m_{ii}\rho_i^* = 0,$$

which is the sufficient and necessary condition that the reference process admits the stationary measure ρ^*. From the above arguments, there always exists a stationary point (ρ^*, S^*) which refers to the reference process itself and $\rho^0 = \rho^1 = \rho^*$ in the SBP. □

In the following, we show that if the solution process of the SBP is Markov, then its density function ρ must be invariant with respect to time.

Corollary 5.4.0.1. *Assume there exists a Markov process solving the SBP and that the reference process is mass-preserving, then for all $ij \in E$, $c_{ij} = \frac{e^{S_i}\sqrt{\rho_i}}{e^{S_j}\sqrt{\rho_j}}$ is the solution of*

$$-\sum_{k\in N(i)} c_{ki}m_{ik} + \sum_{l\in N(j)} c_{lj}m_{jl} - m_{jj} + m_{ii} = 0. \tag{5.25}$$

Moreover, ρ is the invariant measure of the solution process in SBP.

Proof. Since the solution process of the SBP is time homogenous Markov, we can verify that $\frac{e^{S_i}\sqrt{\rho_i}}{e^{S_j}\sqrt{\rho_j}} = c_{ij} > 0$ is independent of time and that

$$d_t\rho = \rho Q,$$

where $Q_{ii} = -\sum_{j\in N(i)} c_{ji}m_{ij}$, $Q_{ij} = c_{ji}m_{ij}$. Let $e^{\psi_i} = e^{S_i}\sqrt{\rho_i}$. Then it holds that

$$\frac{\partial}{\partial t}\rho(i,t) = \sum_{j\in N(i)} -e^{\psi(j,t)-\psi(i,t)}m_{ij}\rho(i,t) + e^{\psi(i,t)-\psi(j,t)}m_{ji}\rho(j,t)$$
$$\frac{\partial}{\partial t}\psi(i,t) = -\sum_{j\in N(i)} (e^{\psi(j,t)-\psi(i,t)} - 1)m_{ij}.$$

As a consequence, for $ij \in E$,

$$\begin{aligned} d_t c_{ij} &= d_t[(e^{\psi_i-\psi_j})] \\ &= c_{ij}(-\sum_{l\in N(i)} e^{\psi_l-\psi_i}m_{il} + \sum_{k\in N(j)} e^{\psi_k-\psi_j}m_{jk}) + c_{ij}(-m_{jj} + m_{ii}) \\ &= c_{ij}(-\sum_{l\in N(i)} c_{li}m_{il} + \sum_{k\in N(j)} c_{kj}m_{jk} - m_{jj} + m_{ii}) = 0. \end{aligned} \tag{5.26}$$

Since $c_{ij} > 0$ for $ij \in E$, we obtain (Equation 5.25). Next we show that the density function ρ is invariant with respect to time.

Notice that $e^{S_i-S_j} = \frac{\sqrt{\rho_j}}{\sqrt{\rho_i}}c_{ij}$ leads to

$$d(S_i - S_j) = \frac{1}{2}\frac{d\rho_j}{\rho_j} - \frac{1}{2}\frac{d\rho_i}{\rho_i} + d\ln(c_{ij}) = \frac{1}{2}\frac{d\rho_j}{\rho_j} - \frac{1}{2}\frac{d\rho_i}{\rho_i}.$$

This implies that

$$\begin{aligned} &- m_{ii} - \frac{1}{2}\sum_{k\in N(i)} e^{S_k-S_i}m_{ik}\frac{\sqrt{\rho_k}}{\sqrt{\rho_i}} - \frac{1}{2}\sum_{k\in N(i)} e^{S_i-S_k}m_{ki}\frac{\sqrt{\rho_k}}{\sqrt{\rho_i}} \\ &+ m_{jj} + \frac{1}{2}\sum_{l\in N(j)} e^{S_l-S_j}m_{jl}\frac{\sqrt{\rho_l}}{\sqrt{\rho_j}} + \frac{1}{2}\sum_{l\in N(j)} e^{S_j-S_l}m_{lj}\frac{\sqrt{\rho_l}}{\sqrt{\rho_j}} \\ &= -\frac{1}{2\rho_i}(\sum_{k\in N(i)} e^{S_i-S_k}m_{ki}\sqrt{\rho_i\rho_k} - \sum_{k\in N(i)} e^{S_k-S_i}m_{ik}\sqrt{\rho_i\rho_k}) \\ &+ \frac{1}{2\rho_j}(\sum_{l\in N(j)} e^{S_j-S_l}m_{lj}\sqrt{\rho_j\rho_l} - \sum_{l\in N(j)} e^{S_l-S_j}m_{jl}\sqrt{\rho_j\rho_l}), \end{aligned}$$

which is equivalent to (Equation 5.25). Using (Equation 5.26), it yields that

$$\sum_{k\in N(i)} c_{ik}m_{ki}\rho(k,t) - c_{ki}m_{ik}\rho(i,t) = \frac{1}{\rho_j}\Big(\sum_{l\in N(j)} c_{jl}m_{lj}\rho(l,t) - c_{lj}m_{jl}\rho(j,t)\Big),$$

that is,

$$d_t\rho_i = d_t \ln(\rho_j).$$

Similarly, we have $d_t\rho_j = d_t \ln(\rho_i)$, which implies that

$$d_t\rho_i = \rho_i\rho_j d_t\rho_i.$$

Now we claim that ρ must be invariant with respect to t. Indeed, if there exists $ij \in E$ such that $d\rho_i \neq 0$, then we have that $\rho_i\rho_j = 1$. However, this contradicts the mass conservation $\sum_{i=1}^N \rho_i = 1$. It follows that $d_t\rho_i = 0$, and therefore ρ should be invariant with respect to time. We conclude that ρ must be the invariant measure of the solution process in the SBP.

□

5.4.2 Discrete SBP based on minimum action with Fisher information

Another way (B) to describe the discrete SBP (see e.g. [177]) lies on the discretization of the variational problem (Equation D.4). Consider the following control problem by directly discretizing the Fisher information $\mathcal{I}(\rho)$ in (Equation D.4):

$$J_1 = \min_{\rho,v}\Big\{\int_0^1 (\frac{1}{2}\langle v,v\rangle_{\theta(\rho)} + \frac{1}{8}\mathcal{I}(\rho))dt + \frac{1}{8}\sum_i(\rho^1(i)\log(\rho^1(i)) - \rho^0\log(\rho^0(i)))\Big\}, \tag{5.27}$$

where $\rho_i \in H^1((0,1))$, $v_{ij} \in L^2((0,1);\theta_{ij}(\rho))$ and

$d_t\rho_t = \rho_t Q_t = -\text{div}_G^\theta(\rho_t v_t)$

with $\rho^0, \rho^1 \in \mathcal{P}_o(G)$. In this case, we look for a stochastic process which obeys the above master equation and minimize the action with the Fisher information

$$\mathcal{I}(\rho) := \frac{1}{2}\sum_{ij\in E}(\log(\rho_i) - \log(\rho_j))^2\widetilde{\theta}_{ij}(\rho).$$

Here $\widetilde{\theta}$ is another density dependent weight which may be different from earlier defined θ on the graph G.

By using Lagrangian multiplier method, the critical point of the discrete variational approach should satisfies

$$\begin{aligned} &v_{ij}(t) = (S_i(t) - S_j(t)),\\ &d_t\rho_i - \sum_{j\in N(i)}(S_i - S_j)\theta_{ij}(\rho) = 0,\\ &d_tS_i + \frac{1}{2}\sum_{j\in N(i)}(S_i - S_j)^2\frac{\partial\theta_{ij}}{\partial\rho_i} = \frac{1}{8}\frac{\partial}{\partial\rho_i}\mathcal{I}(\rho). \end{aligned} \tag{5.28}$$

It forms a Hamiltonian system on the density space with the Hamiltonian $\frac{1}{4}\sum_{i,j}(S_i - S_j)^2\theta_{ij}(\rho) - \frac{1}{8}\mathcal{I}(\rho)$. In other words, the critical point gives a Hamiltonian process on the graph.

We can also reformulate the above system (Equation 5.28) in the form of Schrödinger system (Equation D.5). By taking differential on $f = \sqrt{\rho}e^S$ and $g = \sqrt{\rho}e^{-S}$, we get

$$\begin{aligned} d_tf &= e^{(\frac{1}{2}\log(\rho)+S)}(\frac{1}{2}\frac{d_t\rho}{\rho} + d_tS)\\ &= e^{(\frac{1}{2}\log(\rho)+S)}\Big(\frac{1}{2}\frac{\sum_{j\in N(i)}w_{ij}(S_i - S_j)\theta_{ij}(\rho)}{\rho} - \frac{1}{2}\sum_{j\in N(i)}w_{ij}(S_i - S_j)^2\frac{\partial\theta_{ij}}{\partial\rho_i} + \frac{1}{8}\frac{\partial}{\partial\rho_i}\mathcal{I}(\rho)\Big),\\ d_tg &= e^{(\frac{1}{2}\log(\rho)-S)}(\frac{1}{2}\frac{d_t\rho}{\rho} - d_tS)\\ &= e^{(\frac{1}{2}\log(\rho)-S)}\Big(\frac{1}{2}\frac{\sum_{j\in N(i)}w_{ij}(S_i - S_j)\theta_{ij}(\rho)}{\rho} + \frac{1}{2}\sum_{j\in N(i)}w_{ij}(S_i - S_j)^2\frac{\partial\theta_{ij}}{\partial\rho_i} - \frac{1}{8}\frac{\partial}{\partial\rho_i}\mathcal{I}(\rho)\Big). \end{aligned}$$

Rewriting the above systems into compact form leads to

$$\begin{aligned} d_t f &= -\frac{1}{2}\Delta_G f, \\ d_t g &= \frac{1}{2}\Delta_G g, \end{aligned} \tag{5.29}$$

where Δ_G is the nonlinear discretization of the Laplacian operator,

$$\begin{aligned} &(\Delta_G f)_j = \\ &- f_j \Bigg(\frac{1}{f_j g_j} \sum_{l\in N(j)} \Big(\widetilde{w}_{jl}(\log(f_j/g_j) - \log(f_l/g_l))\widetilde{\theta}_{ij}(fg) + w_{jl}(\log(f_j g_j) - \log(f_l g_l))\theta_{ij}(fg) \Big) \\ &\quad + \sum_{l\in N(j)} \Big(\widetilde{w}_{jl} |\log(f_j/g_j) - \log(f_l/g_l)|^2 \frac{\partial \widetilde{\theta}_{ij}(fg)}{\partial f_j g_j} + w_{jl} |\log(f_j g_j) - \log(f_l g_l)|^2 \frac{\partial \theta_{ij}(fg)}{\partial f_j g_j} \Big) \Bigg). \end{aligned}$$

Remark 31. *In approach (A), the Hamiltonian systems (*(Equation 5.16)*,* (Equation 5.19) *and* (Equation 5.24)*) are corresponding to the control problem* (Equation 5.14)*, which is derived from discretizing the relative entropy $\mathcal{H}(P|R)$ in* (Equation D.1)*; In approach (B), the Hamiltonian systems (*(Equation 5.28) *and* (Equation 5.29)*) are corresponding to the control problem* (Equation 5.27)*, which is derived via discretizing the Fisher information $\mathcal{I}(\rho)$ in* (Equation D.2)*. It worth mentioning that under continuous cases,* (Equation D.1) *and* (Equation D.2) *are equivalent under the transform* (Equation D.3) *and their corresponding Hamiltonian systems are also equivalent. However, this is not true for discrete cases. Discretizing the SBP at different stages leads to different Hamiltonian systems.*

Remark 32 (Nonlinear Markov process as reference process in approach (B))**.** *Let us recall that in continuous space $\mathbb{R}^d$, f, g solve the Schrödinger system*

$$\frac{\partial}{\partial t} f_t = \mathcal{L}_t f_t, \quad \frac{\partial}{\partial t} g_t = -\mathcal{L}_t g_t. \quad \textit{with } f_0,\ g_1 \textit{ are given},$$

with $\mathcal{L}_t$ corresponds to the generator of the reference process R (c.f. Equation (32) of [174]).

By comparing the systems (Equation 5.24) *and* (Equation 5.29) *related to f, g, it is*

observed that $\mathcal{L}_t$ in approach (A) can be viewed as a linear approximation of Laplacian operator, which is associated to the Markov reference process R with transition rate matrix $\{m_{ij}^t\}$; On the other hand, $\mathcal{L}_t = \Delta_G$ in approach (B) is a nonlinear approximation of Laplacian operator. We can thus interpret Δ_G as a nonlinear generator depending on both the state and the distribution. According to the definition of nonlinear Markov process mentioned in subsection 5.2.2, we can associate approach (B) with a nonlinear Markov reference process R generated by Δ_G even though such reference process is not needed in the original control formulation (Equation 5.27).

We end this subsection by presenting the table comparing the two SBPs (Equation 5.19) and (Equation 5.28) discussed in our thesis.

Table 5.1: Comparing two SBPs on graph

	Entropy-minimization SBP	Action-minimization SBP
Origin	Derived from (Equation 5.14)	Derived from (Equation 5.27)
Hamiltonian system	$\frac{d}{dt}\rho_t = \rho_t Q(S_t, t)$ $\frac{d}{dt}S_i = -\sum_{j\in N(i)}(e^{S_j - S_i} - 1)m_{ij}^t$	$\frac{d}{dt}\rho_t = \rho_t Q(S_t)$ $\frac{dS_i}{dt} + \frac{1}{2}\sum_{j\in N(i)}((S_j - S_i)^+)^2 = \frac{1}{8}\frac{\partial}{\partial \rho_i}\mathcal{I}_G(\rho)$
$\mathscr{H}$	$\sum_{i\in V}\sum_{j\in N(i)}(\exp(S_j - S_i) - 1)m_{ij}^t\rho_i$	$\frac{1}{2}\sum_{i\in V}\sum_{j\in N(i)}\rho_i((S_j - S_i)^+)^2 - \frac{1}{8}\mathcal{I}_G(\rho)$
$Q_{ji}, j \neq i$	$e^{S_i - S_j}m_{ji}^t \geq 0$, Hamiltonian process exists	$(S_i - S_j)^+ \geq 0$, Hamiltonian process exists
Reference R	stochastic process induced by **linear generator** $Q = \{m_{ij}^t\}$	stochastic process induced by **nonlinear generator** related to the Fisher Information $\mathcal{I}_G(\rho)$

5.4.3 Periodic marginal distribution of Hamiltonian process in SBP

The periodic solution, as one classical topic of Hamiltonian systems, has been studied for many decades (see e.g. [188, 163, 165]). For our considered Hamiltonian process, the periodicity of the solution appears in the density evolution. Below, we present several examples of periodic reference process, and prove that if the periodic Hamiltonian process exists, it coincides with the reference process in SBP.

By using the Floquet theorem in [189], the fundamental matrix $X(t)$ satisfies $X(t + T) = X(t)\exp(LT)$, where $\exp(LT)$ is a non-singular constant matrix. The Floquet expo-

nents of $d_t\rho = \rho Q_t$ are the eigenvalues $\mu_i, i \leq k \leq N$ of the matrix L. If there exists some i such that $\exp(\mu_i T) = 1$ or -1, then there exists periodic density function with period T or $2T$. As a consequence, we obtain the following results.

Lemma 5.4.1. *Assume that $\{Q_t\}_{t\geq 0}$ is transition rate matrix and Q_t is T-periodic. If there exists a Floquet exponent $\mu = \frac{k\pi \mathbf{i}}{T}, k \in \mathbb{Z}$, then $d_t\rho = \rho Q_t$ has a periodic density.*

Example 5.4.1. *Consider a 2-nodes graph G. Given a reference measure which possesses the marginal distribution as follows,*

$$d_t\rho_1 = \rho_1 m_{11} + \rho_2 m_{21},$$

$$d_t\rho_2 = \rho_1 m_{12} + \rho_2 m_{22},$$

where $m_{21} = -m_{11}$, $m_{22} = -m_{12}$, $m_{11} = -\frac{\frac{1}{2} - \frac{1}{4}\cos(t) + \frac{1}{8}\sin(t) - \frac{1}{16}\sin(t)\cos(t)}{(\frac{1}{2} + \frac{1}{4}\cos(t))^2}$ and $m_{22} = -\frac{1}{\frac{1}{2} + \frac{1}{4}\cos(t)}$.

There exists a nontrivial periodic solution $\rho_1(t) = \frac{1}{2} + \frac{1}{4}\cos(t), \rho_2(t) = 1 - \rho_1(t)$. And the periods of ρ_1 and ρ_2 are both $T = 2\pi$. Therefore, there exists a time inhomogenous Markov process X_t with periodic marginal distribution ρ_t on G with the transition rate matrix $Q_t = (m_{ij})_{i,j\leq 2}$.

We can also show the existence of time inhomogeneous Markov process with periodic marginal distribution on any fully-connected graph.

Proposition 5.4.2. *Suppose G is a fully connected graph, and $\{\rho_t\}$ is a periodic density trajectory (with period T) in $\mathcal{P}_o(G)$, then we can always find a transition rate matrix $Q(t)$ such that ρ_t is the solution to the master equation $\dot{\rho}_t = \rho_t Q(t)$.*

Proof. Assume G contains n vertices. Let us assume the non-diagonal entries of $Q(t)$ to be $\{m_{ij}\}$, we rearrange these entries to form a $n(n-1)$ dimensional vector as:

$$m = (m_{12}, ..., m_{1n}, m_{21}, m_{23}, ..., m_{2n}, ..., m_{n1}, ..., m_{n\,n-1})^T.$$

Plugging m into the Master's equation, we derive the linear equation for m:

$$P(t)\, m = \begin{pmatrix} \dot{\rho_1} & \dot{\rho_2} & \dots & \dot{\rho_n} \end{pmatrix}^T. \tag{5.30}$$

Where P is an $n \times n(n-1)$ matrix defined as

$$P(t) = \left(\begin{array}{c|c|c|c} P_1(t) & P_2(t) & \dots & P_n(t) \end{array} \right).$$

$$P_m(t) = \begin{pmatrix} \rho_m(t) I_m & 0_{m\times(n-m-1)} \\ -\rho_m(t)\boldsymbol{e}_m^T & -\rho_m(t)\boldsymbol{e}_{n-m-1}^T \\ 0_{(n-m-1)\times m} & \rho_m(t) I_{n-m-1} \end{pmatrix}_{n\times(n-1)} \qquad \text{for } 1 \le m \le n$$

Here we denote $\boldsymbol{e}_m^T = \underbrace{(1, ..., 1)}_{m \text{ 1s}}$. We can verify that

$$m^0 = \left(\frac{1}{(n-1)\rho_1}\boldsymbol{e}_{n-1}^T, \frac{1}{(n-1)\rho_2}\boldsymbol{e}_{n-1}^T, ..., \frac{1}{(n-1)\rho_n}\boldsymbol{e}_{n-1}^T \right)$$

belongs to the kernel of $P(t)$, and that $P(t)$ is a full rank matrix. There must exist a solution m^* to (Equation 5.30), where its entries are expressions of $\{\rho_i, \dot{\rho_i}\}_{i\in V}$. In other words, we can directly give such a solution. To be more specific, let's consider the transport process on the loop from vertex 1 to 2, 2 to 3,... n-1 to n and n to 1. This corresponds to setting m_{ij} to 0 except $m_{12}, m_{23}, ..., m_{n-1\ n}$, and m_{n1}. Now (Equation 5.30) becomes:

$$\begin{pmatrix} -\rho_1 & & & & \rho_n \\ & -\rho_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & -\rho_n \end{pmatrix} \begin{pmatrix} m_{12} \\ m_{23} \\ \vdots \\ m_{n-1\ n} \\ m_{n1} \end{pmatrix} = \begin{pmatrix} \dot{\rho_1} \\ \dot{\rho_2} \\ \vdots \\ \dot{\rho}_{n-1} \\ \dot{\rho_n} \end{pmatrix}$$

Therefore the solution is $(-\frac{\dot{\rho_1}-\dot{\rho_n}}{\rho_1}, -\frac{\dot{\rho_2}}{\rho_2}, \cdots, -\frac{\dot{\rho}_{n-1}}{\rho_{n-1}}, -\frac{\dot{\rho_n}}{\rho_n})^T$.

Then we can directly take $m(t) = Km^0(t) + m^*(t)$, since $\{\rho_t\}$ is in the interior of $\mathcal{P}(G)$, we can always find a large enough $K > 0$ that guarantees the entries of $m(t)$ to be always non negative. And $m(t)$ forms the transition rate matrix $Q(t)$ whose master equation admits the periodic solution $\{\rho_t\}$. □

Example 5.4.2. *Consider the periodic marginal distribution ρ_t:*

$$\rho_t = (\frac{\cos t}{2\sqrt{6}} + \frac{\sin t}{6\sqrt{2}} + \frac{1}{3}, -\frac{\cos t}{2\sqrt{6}} + \frac{\sin t}{6\sqrt{2}} + \frac{1}{3}, -\frac{\sin t}{3\sqrt{2}} + \frac{1}{3}).$$

which is a circle centered at $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ with radius $\frac{1}{2\sqrt{3}}$ on $\mathcal{P}(G)$. Following the idea of Proposition Proposition 5.4.2, one may take

$$\begin{aligned}
m_{11}(t) &= -\frac{6\sqrt{2} + \sqrt{3}\sin t - 3\cos t}{\sqrt{3}\cos t + 4\sin t + 2\sqrt{2}},\ m_{12} = -m_{11},\ m_{13} = 0,\\
m_{22}(t) &= -\frac{24 - 4\sqrt{2}\cos t}{-\sqrt{6}\cos t + \sqrt{2}\sin t + 4},\ m_{21} = -\frac{1}{2}m_{22},\ m_{23} = -\frac{1}{2}m_{22},\\
m_{33}(t) &= -\frac{3\sqrt{2}}{\sqrt{2} - \sin t}\ m_{13} = 0, m_{23} = -m_{33}.
\end{aligned}$$

such that $d_t = \rho_t Q_t$ with $Q_t = (m_{ij})_{i,j\leq 3}$.

Next we aim to use general SBP (Equation 5.20) to produce a Hamiltonian process with periodic marginal distribution on G. In particular, when the convex function $u = x\log(x) - x - 1$, by using the Nelson's transformation $\psi_i = \sqrt{\rho_i}e^{S_i}$, the Hamiltonian system can be also rewritten as

$$\begin{aligned}
dS_i &= -m_{ii} - \frac{1}{2}\sum_{j\in N(i)} e^{S_j - S_i} m_{ij}(t)\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}} - \frac{1}{2}\sum_{j\in N(i)} e^{S_i - S_j} m_{ji}(t)\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}},\\
d\rho_i &= \sum_{j\in N(i)} e^{S_i - S_j} m_{ji}(t)\sqrt{\rho_j}\sqrt{\rho_i} - \sum_{j\in N(i)} e^{S_j - S_i} m_{ij}(t)\sqrt{\rho_i}\sqrt{\rho_j}.
\end{aligned}$$

with the Hamiltonian $\tilde{\mathscr{H}}(\rho, S, t) = \sum_{i\in V}\sum_{j\in N(i)} e^{(S_j - S_i)} m_{ij}(t)\sqrt{\rho_i\rho_j}$. Taking ψ as a time-independent potential and choosing ρ^0, ρ^1 as the initial and terminal distribution from

the periodic solution, then the distribution of the solution process is exactly same as the reference process. Thus it induces a Hamiltonian system which is periodic in time. Therefore there exists SBP with the given ρ^0, ρ^1 such that the solution process is Hamiltonian and its marginal distribution is periodic in time.

In the following, we assume that the Legendre transformation u^* of u in (Equation 5.20) is continuous differentiable and satisfies

$$
\begin{aligned}
&u^*(x) \geq 0, \text{ if } x \leq 0, \ u^*(x) \leq 0, \text{ if } x \geq 0, \\
&\frac{\partial u^*}{\partial x}(0) = 1, \ \lim_{x \to -\infty} \left| \frac{\partial u^*}{\partial x}(x) \right| < \infty, \ \lim_{x \to +\infty} \frac{\partial u^*}{\partial x}(x) = +\infty.
\end{aligned}
$$

Now we are able to give the characterization of the periodic Hamiltonian process on finite graph via general SBP.

Theorem 5.4.2. *Assume that the reference process is periodic with the marginal distribution and its period $T > 0$. There always exists ρ^0, ρ^1 such that the critical point of the general SBP problem* (Equation 5.20) *is a Hamiltonian process and its marginal distribution is periodic in time.*

Proof. Notice that the critical point of SBP satisfies

$$
\begin{aligned}
&\frac{\partial}{\partial t}\rho(i,t) = \sum_{j \in N(i)} -\frac{\partial u^*}{\partial x}(\psi_j - \psi_i) m_{ij} \rho(i,t) + \frac{\partial u^*}{\partial x}(\psi_i - \psi_j) m_{ji} \rho(j,t), \\
&\frac{\partial}{\partial t}\psi(i,t) = -\sum_{j \in N(i)} u^*(\psi_j - \psi_i) m_{ij},
\end{aligned}
$$

where $\rho(0) = \rho^0, \rho(1) = \rho^1$. Choosing ρ^0, ρ^1 as two different distribution at different time of the reference process, and taking $\psi_i = \psi_j$, we get

$$
\begin{aligned}
&\frac{\partial}{\partial t}\rho(i,t) = \sum_{j \in N(i)} -m_{ij}\rho(i,t) + m_{ji}\rho(j,t), \\
&\frac{\partial}{\partial t}\psi(i,t) = 0.
\end{aligned}
$$

This implies that the critical point forms a Hamiltonian system with Hamiltonian defined as $H(\rho, \psi, t) = \sum_{i,j} m_{ij}(t)\rho_i$. Due to the fact that the marginal distribution of reference process is periodic in time, the critical point is exactly equal to the reference process and its marginal distribution is periodic. □

One may wonder whether there exists certain Hamiltonian process whose marginal distribution is periodic but is not the reference process. We first use a 2-nodes graph example to point out it is not possible to get such Hamiltonian by using SBP when $u(x) = x\log(x) - x - 1$. Even worse, we show that for general finite graph, the periodic Hamiltonian process exists if and only if it equals to a reference process in general SBP.

Example 5.4.3. *Given G consisting of 2 nodes. Assume the reference process with transition rate matrix m is periodic with period $T > 0$ and $\{t \in [0, T] | m_{ij}(t) = 0, ij \in E\}$ has Lebesgue measure zero. Notice that ρ, S of the Hamiltonian process $X(t)$ satisfies*

$$\begin{aligned}
&\frac{\partial}{\partial t}\rho(1,t) = -e^{\psi(2,t)-\psi(1,t)}m_{12}\rho(1,t) + e^{\psi(1,t)-\psi(2,t)}m_{21}\rho(2,t),\\
&\frac{\partial}{\partial t}(\psi(1,t)-\psi(2,t)) = -(e^{\psi(2,t)-\psi(1,t)}-1)m_{12} + (e^{\psi(1,t)-\psi(2,t)}-1)m_{21}.
\end{aligned}$$

Since m_{12}, $m_{21} \geq 0$, then $\psi(1)-\psi(2)$ equals to constant if and only if $\psi(1) = \psi(2)$. Meanwhile, if $\psi_1-\psi_2 > 0$, then $\psi_1-\psi_2$ is increasing to $+\infty$, and $\psi_1-\psi_2$ is decreasing to $-\infty$ if $\psi_1 < \psi_2$. Then we claim that ρ_1 is not periodic in time. If we assume that ρ_1 is periodic with period T_1, then it holds true $\int_{kT_1}^{(k+1)T_1} -e^{\psi(2,t)-\psi(1,t)}m_{12}\rho(1,t)+e^{\psi(1,t)-\psi(2,t)}m_{21}\rho(2,t)dt = 0$. Without losing generality, let us assume that $\psi_1 - \psi_2 > 0$. It is not hard to see that $e^{\psi(1,t)-\psi(2,t)}$ is increasing to $+\infty$ and $e^{\psi(1,t)-\psi(2,t)}$ is decreasing to 0 as $t \to \infty$. The boundedness of $\rho(1,t), \rho(2,t)$ yield that there exists large enough k such that

$$\int_{kT_1}^{(k+1)T_1} -e^{\psi(2,t)-\psi(1,t)}m_{12}\rho(1,t) + e^{\psi(1,t)-\psi(2,t)}m_{21}\rho(2,t)dt > 0,$$

which leads to a contradiction. Therefore, $\rho(t)$ is periodic in time if and only if $\psi_1 = \psi_2$.

This implies that $X(t)$ is exactly the reference process.

Theorem 5.4.3. *Assume the reference process with transition rate matrix m is periodic with period $T > 0$ and $\{t \in [0,T] | m_{ij}(t) = 0, ij \in E\}$ has Lebesgue measure zero. Then the Hamiltonian process which has periodic density distribution in general SBP problem* (Equation 5.20) *is equal to the reference process which has the periodic density distribution.*

Proof. Assume that there is a maximum $\psi_{i^*} \geq \psi_i$, $i \neq i^*$ and $\psi_{i^*} > \psi_{i_{min}}$. Then according to the evolution of ψ,

$$\frac{\partial}{\partial t}\psi(i,t) = -\sum_{j \in N(i)} u^*(\psi(j,t) - \psi(i,t))m_{ij},$$

then the maximum principle holds, i.e., $\psi_{i^*}(t) \geq \psi_i(t) \geq \psi_{i_{min}}(t)$. Notice that

$$\frac{d}{dt}\rho(i,t) = \sum_{j \in N(i)} -\frac{\partial u^*}{\partial x}(\psi_j - \psi_i)m_{ij}\rho(i,t) + \frac{\partial u^*}{\partial x}(\psi_i - \psi_j)m_{ji}\rho(j,t).$$

The periodicity of ρ_i implies that there exists $T_1 > 0$ for any $k \in \mathbb{N}^+$ such that

$$\int_{kT_1}^{(k+1)T_1} \sum_{j \in N(i)} -\frac{\partial u^*}{\partial x}(\psi_j - \psi_i)m_{ij}\rho(i,t) + \frac{\partial u^*}{\partial x}(\psi_i - \psi_j)m_{ji}\rho(j,t)dt = 0.$$

Due to the maximum principle, if there exists one node l with a local maximum of ψ_l connected with another node k with a local minimum of ψ_k, it will lead to $\psi_l - \psi_k \to +\infty$ as $t \to \infty$. This contradicts with the periodicity of ρ_k and ρ_l. If any node l with a local maximum of ψ_l is not connected with another node k with a local minimum of ψ_k, we pick a road $l, j_1, \cdots, j_w, k$ which connects l and k. Notice that $\psi_l \to +\infty$, $\psi_k \to -\infty$, $\psi_{j_m} \in (\psi_k, \psi_l), m \leq w$. Then there must exists j_m such that m is the smallest number which satisfies $\psi_k - \psi_{j_m} \to -\infty$. Now consider the periodicity of ρ_{j_m}. There exists k'

large enough such that

$$\int_{k'T_1}^{(k'+1)T_1} \sum_{j\in N(j_m)} -\frac{\partial u^*}{\partial x}(\psi_j - \psi_{j_m}) m_{j_m j}\rho(j_m,t) + \frac{\partial u^*}{\partial x}(\psi_{j_m} - \psi_j) m_{jj_m}\rho(j,t)dt > 0.$$

This leads to a contradiction, we complete the proof.

5.5 More examples and future work

In this section, we conclude this chapter by presenting a few more examples of Hamiltonian processes on graph and more questions to be considered in the future.

Example 5.5.1. *(Euler-Lagrangian equations [175]) Assume that the Lagrangian in density mainifold is given by* $\mathscr{L}(\rho_t, \dot{\rho}_t) = \frac{1}{2} g_W(\dot{\rho}_t, \dot{\rho}_t) - \mathcal{F}(\rho_t)$. *Here* $g_W(\sigma_1, \sigma_2) := -\sigma_1(\Delta_\rho)^+\sigma_2$ *where* $\sigma_k \in \mathcal{T}_\rho\mathcal{P}_o(G), k = 1, 2$ *and* $(\Delta_\rho)^+$ *is the pseudo inverse of the weight graph Laplacian matrix* $\Delta_\rho(\cdot) := div_G^\theta(\rho\nabla_G(\cdot))$. *Then the critical point of*

$$\inf_{\rho_t} \int_0^T \mathscr{L}(\rho_t, \partial_t\rho_t)dt$$

with given ρ_0 *and* ρ_T *satisfies the Euler-Lagrangian equation*

$$\partial_t \frac{\delta}{\delta \partial_t\rho_t}\mathscr{L}(\rho_t, \partial_t\rho_t) = \frac{\delta}{\delta\rho_t}\mathscr{L}(\rho_t, \partial_t\rho_t) + C(t).$$

By introducing the Legendre transform $S_t = (-\Delta_{\rho_t})^+\partial_t\rho_t$, *it can be rewritten as a Hamiltonian system. That is*

$$\begin{aligned}
&\partial_t\rho_t + div_G^\theta(\rho\nabla_G S) = 0,\\
&\partial_t S_t + \frac{1}{4}\sum_{j\in N(i)} (S_i - S_j)^2(\partial_{\rho_i}\theta(\rho_i,\rho_j) + \partial_{\rho_i}\theta(\rho_j,\rho_i)) + \frac{\delta}{\delta\rho_t}\mathcal{F}(\rho_t) = C(t),
\end{aligned}$$

with the Hamiltonian $\mathscr{H}(\rho, S) = \frac{1}{4}\sum_{ij}(S_i - S_j)^2\theta_{ij}w_{ij} + \mathcal{F}(\rho_t)$. *Therefore, if the transition rate matrix in generalized master equation is well-defined, the Euler-Lagrangian equation*

in density space determines a Hamiltonian process on G.

Example 5.5.2. *(Madelung system [179]) The energy is given by*

$$\mathscr{H}(\rho, S) = \frac{1}{4}\sum_{ij\in E}(S_i - S_j)^2\theta_{ij}w_{ij} + \mathcal{F}(\rho_t) + \beta\mathcal{I}(\rho_t), \beta > 0.$$

Here $\mathcal{F}(\rho) = \sum_i \rho_i\mathbb{V}_i + \sum_{i,j}\rho_i\rho_j\mathbb{W}_{ij}$, and $\mathcal{I}(\rho) = \frac{1}{2}\sum_{ij\in E}(\log(\rho_i) - \log(\rho_j))^2\widetilde{\theta}_{ij}$. Here $\widetilde{\theta}_{ij}$ is another density dependent weight on the graph that can be the same or different from θ_{ij}. The Madelung system is

$$\begin{aligned}
&\partial_t\rho_t + div_G^\theta(\rho\nabla_G S) = 0,\\
&\partial_t S_t + \frac{1}{4}\sum_{j\in N(i)}(S_i - S_j)^2(\partial_{\rho_i}\theta(\rho_i,\rho_j) + \partial_{\rho_i}\theta(\rho_j,\rho_i)) + \frac{\delta}{\delta\rho_t}\mathcal{F}(\rho_t) + \beta\frac{\delta}{\delta\rho_t}\mathcal{I}(\rho_t) = C(t).
\end{aligned}$$

When taking $\theta = \theta^U$, the Madelung system in density space determines a Hamiltonian process on G. This system has a close relationship with the discrete Schrödinger equation [171].

Example 5.5.3. *(p-Wasserstein distance) The L^p Wasserstein distance, $p \in (1,\infty)$, is related to the following minimization problem,*

$$W_p^p(\rho^0,\rho^1) = \inf_v\{\int_0^1\sum_{i=1}^N\sum_{j\in N(i)}\frac{1}{2}\theta_{ij}(\rho)v_{ij}^p dt : \partial_t\rho + div_G^\theta(\rho v) = 0, \rho(0) = \rho^0, \rho(1) = \rho^1\}.$$

We refer to [190] for a continuous version of p-Wasserstein distance. Its critical point is related to the Hamiltonian system in density space

$$\begin{aligned}
&\partial_t\rho_t + div_G^\theta(\rho_t|\nabla_G S|^{q-2}\nabla_G S) = 0,\\
&\partial_t(S_i) + \frac{1}{2q}\sum_{j\in N(i)}|(\nabla_G S)_{ij}|^q(\partial_1\theta_{ij} + \partial_2\theta_{ji}) = 0,
\end{aligned}$$

with the Hamiltonian

$$\mathscr{H}(\rho, S) = \frac{1}{2q} \sum_{i,j} |\nabla_G S|^q \theta_{ij}, \frac{1}{q} + \frac{1}{p} = 1, p \in (1, \infty).$$

When the equation of ρ is determined by a transition rate matrix, this leads to a Hamiltonian process.

To end the discussion, we want to mention two problems that are worth to be studied further.

- As shown in [179], the classical Hamiltonian ODEs induce the Wasserstein–Hamiltonian flows on the density manifold. There are many special properties for Hamiltonian system in continuous space, such as conservation of energy, preservation of the volume etc. The particle-level counterpart on graph is the Hamiltonian process introduced in Definition 5.3.1. In addition to the conservation property discussed in Remark 26, are there other quantities or structures being preserved by the Hamiltonian process on the graph G?

- As discussed in [191], stochastic differential equations can be well approximated by continuous time random walk on the lattices. Then it is natural to ask whether the proposed Hamiltonian process on a lattice can be used to approximate a Hamiltonian system in $\mathbb{R}^d$. If so, how well is the approximation?

Appendices

Table 2: Notations frequently used in this thesis

Indices	Meaning
$T_\sharp \mu$	pushforward of measure μ by the map T, c.f. Equation 2.1
$\mathscr{L}^d$	Lebesgue measure on $\mathbb{R}^d$
$\mu \ll \nu$	Measure μ is absolute continuous to ν
$\text{Law}(\boldsymbol{X})$	Probability distribution of the random variable $\boldsymbol{X}$
$\frac{d\mu}{d\nu}$	Radon-Nikodym derivative between measure μ, ν
$W_2(\cdot, \cdot)$	2-Wasserstein distance, c.f. Example 2.1.1
$\mathcal{P}_2$	Wasserstein manifold, c.f. Equation 2.49
$\mathcal{T}\mathcal{P}_2$	Tangent bundle of Wasserstein manifold
$\mathcal{T}_\rho\mathcal{P}_2$	Tangent space of Wasserstein manifold at ρ
$\mathcal{T}^*\mathcal{P}_2$	Cotangent bundle of Wasserstein manifold
g^W	Wasserstein metric, c.f. Equation 2.58
$o(h^\alpha)$	Higher order term of h^α ($\alpha > 0$)
$O(h^\alpha)$	Same order term of h^α ($\alpha > 0$)
grad_W	Wasserstein gradient, c.f. Equation 2.61
$\mathcal{H}(\rho)$	Relative entropy of ρ
$\mathcal{D}_{\text{KL}}(\cdot\|\cdot)$	Kullbeck-Leibler divergence
$L(x, v)$	Lagrandian defined on particle space
$\mathscr{L}(\rho, \partial_t\rho)$	Lagrangian defined on $\mathcal{T}\mathcal{P}_2$
$H(x, p)$	Hamiltonian defined on particle space
$\mathscr{H}(\rho, S)$	Hamiltonian defined on $\mathcal{T}^*\mathcal{P}_2$
$\succ, \succeq$	For square matrix A, B, $A \succ B$ indicates $A - B$ is positive definite, $A \succeq B$ indicates $A - B$ is semi-positive definite
I_n	$n \times n$ identity matrix
O_n	$n \times n$ zero matrix
π_1, π_2	$\pi_1 : \mathbb{R}^d \times \mathbb{R}^d :\to \mathbb{R}^d, (x, y) \mapsto x$ is the projection onto the first coordinate component; similarly π_2 is the projection onto the second component

Table 3: Notations frequently used in this thesis (continued)

Indices	Meaning
Θ	Parameter space of pushforward map T_θ
θ	Parameter of pushforward map T_θ
p	Reference probability distribution used in Chapter 4
ρ_θ	Pushforwarded distribution of p by T_θ, i.e., $\rho_\theta = T_{\theta\sharp}p$
$H(\theta)$	Relative entropy of ρ_θ
$\mathcal{I}(\mu\|\nu)$	Fisher information of distribution μ with reference measure ν
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$G(V, E)$	Finite graph with vertices set V, and edge set E
∇_G	Gradient on graph, c.f. Equation 5.7
div_G^θ	Divergence operator on graph with weight function θ, c.f. Equation 5.6
$N(i)$	Set of neighbouring vertices of vertex i of graph G
Q_t	Transition rate matrix of Markov process

APPENDIX A

APPENDIX FOR PART 2

Definition A.0.1 (Superdifferentiablity)**.** *For function $f : \mathbb{R}^n \to \mathbb{R}$, we say f is superdifferentiable at x, if there exists $p \in \mathbb{R}^n$, such that*

$$f(z) \geq f(x) + \langle p, z - x \rangle + o(|z - x|).$$

Definition A.0.2 (Locally Lipschitz)**.** *Let $U \subset \mathbb{R}^n$ be open and let $f : \mathbb{R}^n \to \mathbb{R}$ be given. Then*

(1) f is Lipschitz if there exists $L < \infty$ such that

$$\forall x, z \in \mathbb{R}^n, \quad |f(z) - f(x)| \leq L|x - z|.$$

(2) f is said to be locally Lipschitz if for any $x_0 \in \mathbb{R}^n$, there is a neighbourhood O of x_0 in which f is Lipschitz.

Definition A.0.3 (Distance function)**.** *For a set X, a distance function $d : X \times X \to \mathbb{R}$ is a function satisfies*

- $d(x, y) = 0$ *if and only if* $x = y$*;*
- $d(x, y) = d(y, x)$ *for any* $x, y \in X$*;*
- $d(x, y) + d(y, z) \geq d(x, z)$.

APPENDIX B

APPENDIX FOR PART 3

B.1 Scalable computation of Monge maps with general costs

B.1.1 Relation between our method and generative adversarial networks

It is worth pointing out that our scheme and Wasserstein generative adversarial networks (WGAN) [8] are similar in the sense that they are both doing minimization over the generator/map and maximization over the discriminator/dual potential. However, there are two main distinctions between them. Such differences are not reflected from the superficial aspects such as the choice of reference distributions ρ_a, but come from the fundamental logic hidden behind the algorithms.

- We want to first emphasize that the mechanisms of two algorithms are different: Typical Wasserstein GANs (WGAN) are usually formulated as

$$\min_{G} \underbrace{\max_{\|D\|_{\text{Lip}\leq 1}} \int D(y)\rho_b(y)dy - \int D(G(x))\rho_a(x)dx}_{1-\text{Wasserstein distance } W_1(G_\sharp\rho_a,\rho_b)} \tag{B.1}$$

  and ours reads

$$\underbrace{\max_{f}\min_{T} \int f(y)\rho_b(y)dy - \int f(T(x))\rho_a(x)dx + \int c(X,T(x))\rho_a(x)dx}_{\text{general Wasserstein distance } C(\rho_a,\rho_b)} \tag{B.2}$$

  The inner maximization of (Equation B.1) computes W_1 distance via Kantorovich duality and the outer loop minimize the W_1 gap between desired ρ_b and $G_\sharp\rho_a$; However, the logic behind our scheme (Equation B.2) is different: the inner optimization computes for the $c-$transform of f, i.e. $f^{c,-}(x) = \sup_\xi(f(\xi) - c(x,\xi))$; And

the outer maximization computes for the Kantorovich dual problem $C(\rho_a, \rho_b) = \sup_f \{\int f(y)\rho_b(y)dy - \int f^{c,-}(x)\rho_a(x)dx\}$.

Even under W_1 circumstance, one can verify the intrinsic difference between two proposed methods: when setting the cost $c(x, y) = \|x - y\|$, and $\rho_a = G\sharp\rho_a$ in (Equation B.2), the entire "max-min" optimization of (Equation B.2) (underbraced part) is equivalent to the inner maximization problem of (Equation B.1) (underbraced part), but not for the entire saddle point scheme.

It is also important to note that WGAN aims to minimize the distance between generated distribution and the target distribution and the ideal value for (Equation B.1) is 0. On the other hand, one of our goal is to estimate the optimal transport distance between the initial distribution ρ_a and the target distribution ρ_b. Thus the ideal value for (Equation B.2) should be $C(\rho_a, \rho_b)$, which is not 0 in most of the cases.

- We then argue about the optimality of the computed map G and T: In (Equation B.1), one is trying to obtain a map G by minimizing $W_1(\rho_b, G_\sharp\rho_a)$ w.r.t. G, and hopefully, $G_\sharp\rho_a$ can approximate ρ_b well. However, there isn't any restriction exerted on G, thus one can not expect the computed G to be the optimal transport map between ρ_a and ρ_b; On the other hand, in (Equation B.2), we not only compute T such that $T_\sharp\rho_a$ approximates ρ_b , but also compute for the optimal T that minimizes the transport cost $\mathbb{E}_{\rho_a}[c(X, T(X))]$. In (Equation B.2), the computation of T is naturally incorporated in the max-min scheme and there exists theoretical result (recall (Theorem 3.2.1)) that guarantees T to be the optimal transport map.

In summary, even though the formulation of both algorithms are similar, the designing logic (minimizing distance vs computing distance itself) and the purposes (computing arbitary pushforward map vs computing the optimal map) of the two methods are distinct. Thus the theoretical and empirical study of GANs cannot be trivially translated to proposed method. In addition to the above discussions, we should also refer the readers to [71], in which a

comparison between a similar saddle point method and the regularized GANs are made in section 6.2 and summarized in Table 1.

B.1.2 Proof of Theorem 3.1.2

We consider the Monge problem from $\mathbb{R}^d$ to $\mathbb{R}^d$ with the cost function $c \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$ satisfying the conditions mentioned in Theorem 3.2.1. Recall that we assume c satisfies:

$$\partial_{xy}c(x,y), \text{ as an } n \times n \text{ matrix, is invertible and self-adjoint.} \quad \text{(Equation 3.5)}$$

$$\partial_{yy}c(x,y) \text{ is independent of } x; \quad \text{(Equation 3.6)}$$

We further denote

$$\sigma(x,y) = \sigma_{\min}(\partial_{xy}c(x,y)) \quad \text{(B.3)}$$

as the minimum singular value of matrix $\partial_{xy}c(x,y)$, since the matrix is invertible, $\sigma(x,y) > 0$ for any $x, y \in \mathbb{R}^n$.

Theorem 3.1.2 (Posterior Error Analysis via Duality Gaps)**.** *Assume $f \in C^2(\mathbb{R}^d)$ is a c-concave function and assume that there exists $\varphi \in C^2(\mathbb{R}^d)$ such that $f(y) = \inf_x\{\varphi(x) + c(x,y)\}$. Suppose $\varphi(x) + c(x,y)$ has a unique minimizer $\hat{x}_y$ for arbitrary $y \in \mathbb{R}^d$. We further assume there exists function $\lambda(\cdot) > 0$ such that the Hessian of $\varphi(\cdot) + c(\cdot, y)$ at minimizer $\hat{x}_y$ is positive definite and bounded from above:*

$$\lambda(y)I_n \succeq \nabla^2_{xx}(\varphi(x) + c(x,y))|_{x=\hat{x}_y} \succ O_n, \quad \text{(Equation 3.7)}$$

where I_n, O_n denotes $n \times n$ identity matrix and zero matrix.

We denote the duality gaps

$$\mathcal{E}_1(T,f) = \mathcal{L}(T,f) - \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},f), \quad \mathcal{E}_2(f) = \sup_{\widetilde{f}} \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},\widetilde{f}) - \inf_{\widetilde{T}} \mathcal{L}(\widetilde{T},f)$$

Denote T_* as the Monge map of (Equation 2.2)*. Then there exists a strict positive weight function $\beta(\cdot) > \min_y\{\frac{\sigma(x,y)}{2\lambda(y)}\}$ (β depends on c, T_*, f and φ), such that the weighted L^2 error between computed map T and optimal map T_* is upper bounded by*

$$\|T - T_*\|_{L^2(\beta\mu)} \leq \sqrt{2(\mathcal{E}_1(T, f) + \mathcal{E}_2(f))}.$$

Lemma B.1.1. *Suppose $n \times n$ matrix A is self-adjoint, i.e. $A = A^T$, with minimum singular value $\sigma_{\min}(A) > 0$. Also assume $n \times n$ matrix H is self-adjoint and satisfies $\lambda I_n \succeq H \succ O_n$. Then $AH^{-1}A \succeq \frac{\sigma_{\min}(A)^2}{\lambda} I_n$.*

Proof of Lemma B.1.1 . Firstly, one can verify that $H^{-1} \succeq \frac{1}{\lambda} I_n$ by digonalizing H^{-1}. To prove this lemma, we only need to verify that for arbitrary $v \in \mathbb{R}^n$,

$$v^{\mathrm{T}} AH^{-1}Av = (Av)^{\mathrm{T}} H^{-1} Av \geq \frac{|Av|^2}{\lambda} \geq \frac{\sigma_{\min}(A)^2}{\lambda} |v|^2$$

Thus $AH^{-1}A - \frac{\sigma_{\min}(A)^2}{\lambda} I_n$ is non-negative definite. □

The following lemma is crucial for proving our results, it analyzes the concavity of the target function $f(\cdot) - c(\cdot, y)$ with f to be c-concave.

Lemma B.1.2 (Concavity of $f(\cdot) - c(x, \cdot)$ when f is c-concave)**.** *Suppose the cost function $c(\cdot, \cdot)$ and f satisfy the conditions mentioned in Theorem 3.2.2. Denote the function $\Psi_x(y) = f(y) - c(x, y)$, keep all notations defined before, we have*

$$\nabla^2 \Psi_x(y) \preceq -\frac{\sigma(x, y)^2}{\lambda(y)} I_n.$$

Proof of Lemma B.1.2. First, we notice that f is c-convex, thus, there exists φ such that $f(y) = \inf_x\{\varphi(x) + c(x, y)\}$. Let us also denote $\Phi(x, y) = \varphi(x) + c(x, y)$.

Now for a fixed $y \in \mathbb{R}^n$, We pick one

$$\hat{x}_y \in \operatorname{argmin}_x \{\varphi(x) + c(x, y)\}$$

Since we assumed that $\varphi \in C^2(\mathbb{R}^n)$ and $c \in C^2(\mathbb{R}^n \times \mathbb{R}^n)$, we have

$$\partial_x \Phi(\hat{x}_y, y) = \nabla \varphi(\hat{x}_y) + \partial_x c(\hat{x}_y, y) = 0 \tag{B.4}$$

At the same time, since $\hat{x}_y$ is the minimum point of the C^2 function $\Phi(\cdot, y)$, then the Hessian of $\Phi(\cdot, y)$ at $\hat{x}_y$ is positive definite,

$$\partial^2_{xx} \Phi(\hat{x}_y, y) = \nabla^2_{xx}(\varphi(x) + c(x, y))\Big|_{x=\hat{x}_y} = \nabla^2 \varphi(\hat{x}_y) + \partial^2_{xx} c(\hat{x}_y, y) \succ 0.$$

Since $\partial^2_{xx} \Phi(\hat{x}_y, y)$ is positive definite, it is also invertible. We can now apply the implicit function theorem to show that the equation $\partial_x \Phi(x, y) = 0$ determines an implicit function $\hat{x}(\cdot)$, which satisfies $\hat{x}(y) = \hat{x}_y$ in a small neighbourhood $U \subset \mathbb{R}^n$ containing y. Furthermore, one can show that $\hat{x}(\cdot)$ is continuously differentiable at y. We will denote $\hat{x}_y$ as $\hat{x}(y)$ in our following discussion.

Now differentiating (Equation B.4) with respect to y yields

$$\partial^2_{xx} \Phi(\hat{x}(y), y) \nabla \hat{x}(y) + \partial^2_{xy} c(\hat{x}(y), y) = 0 \tag{B.5}$$

On one hand, (Equation B.5) tells us

$$\nabla \hat{x}(y) = -\partial_{xx} \Phi(\hat{x}(y), y)^{-1} \partial_{xy} c(\hat{x}(y), y). \tag{B.6}$$

On the other hand, notice that $c \in C^2(\mathbb{R}^n \times \mathbb{R}^n)$, thus $\partial_{xy} c = \partial_{yx} c$. By (Equation B.5), we have

$$\begin{aligned} \partial^2_{yx} c(\hat{x}(y), y) \nabla \hat{x}(y) =& - \partial^2_{xx} \Phi(\hat{x}(y), y) \nabla \hat{x}(y) \nabla \hat{x}(y) \\ =& - (\nabla^2 \varphi(\hat{x}(y)) + \partial^2_{xx} c(\hat{x}(y), y)) \nabla \hat{x}(y) \nabla \hat{x}(y). \end{aligned} \tag{B.7}$$

Now we are able to prove our theorem, we directly compute

$$\nabla^2 \Psi_x(y) = \nabla^2 f(y) - \partial^2_{yy} c(x, y). \tag{B.8}$$

in order to compute $\nabla^2 f(y)$, we first compute $\nabla f(y)$

$$\nabla f(y) = \nabla(\varphi(\hat{x}(y)) + c(\hat{x}(y), y)) = \partial_y c(\hat{x}(y), y). \tag{B.9}$$

the second equality is due to the envelope theorem [143]. Then $\nabla^2 f(y)$ can be computed as

$$\nabla^2 f(y) = \partial_{yx} c(\hat{x}(y), y) \nabla \hat{x}(y) + \partial_{yy} c(\hat{x}(y), y). \tag{B.10}$$

Plugging (Equation B.7) into (Equation B.10), recall (Equation B.8), this yields

$$\nabla^2 \Psi_x(y) = -(\nabla^2 \varphi(\hat{x}(y)) + \partial^2_{xx} c(\hat{x}(y), y)) \nabla \hat{x}(y) \nabla \hat{x}(y) + \partial^2_{yy} c(\hat{x}(y), y) - \partial^2_{yy} c(x, y)$$

Now by (Equation 3.6), $\partial_{yy} c(x, y)$ is independent of x, thus $\partial^2_{yy} c(\hat{x}(y), y) - \partial^2_{yy} c(x, y) = 0$. As a result we obtain

$$\begin{aligned} \nabla^2 \Psi_x(y) &= -(\nabla^2 \varphi(\hat{x}(y)) + \partial^2_{xx} c(\hat{x}(y), y)) \nabla \hat{x}(y) \nabla \hat{x}(y) \\ &= -\partial^2_{xx} \Phi(\hat{x}(y), y) \nabla \hat{x}(y) \nabla \hat{x}(y). \end{aligned} \tag{B.11}$$

To further simplify (Equation B.11), recall (Equation B.6), we have

$$\nabla^2 \Psi_x(y) = -\partial_{xy} c(\hat{x}(y), y) \partial_{xx} \Phi(\hat{x}(y), y)^{-1} \partial_{xy} c(\hat{x}(y), y).$$

By (Equation 3.7), $\lambda(y) I_n \succeq \partial_{xx} \Phi(\hat{x}(y), y) \succ O_n$. Recall condition (Equation 3.5), $\partial_{xy} c$ is self-adjoint, and (Equation B.3) leads to $\sigma_{\min}(\partial_{xy} c(x, y)) = \sigma(x, y)$. Now applying

Lemma B.1.1 yields

$$\nabla^2 \Psi_x(y) \preceq -\frac{\sigma(x,y)^2}{\lambda(y)} I_n.$$

□

Now we can prove main result in Theorem 3.2.2:

Proof of Theorem 3.2.2. In this proof, we denote $\int$ as $\int_{\mathbb{R}^d}$ for simplicity.

We first recall

$$\mathcal{L}(T,f) = \int f(y) d\nu(y) - \int (f(T(x)) - c(x,T(x))) d\mu(x),$$

also recall definition (Equation 3.3), $f^{c,-}(x) = \sup_y \{f(y) - c(x,y)\}$, we can write

$$\begin{aligned}
\mathcal{E}_1(T,f) &= -\int [f(T(x)) - c(x,T(x))] d\mu(x) + \inf_{\widetilde{T}} \left\{ \int [f(\widetilde{T}(x)) - c(x,\widetilde{T}(x))] d\mu(x) \right\} \\
&= \int [f^{c,-}(x) - (f(T(x)) - c(x,T(x)))] d\mu(x)
\end{aligned}$$

We denote

$$T_f(x) = \text{argmax}_y \{f(y) - c(x,y)\} = \text{argmax}_y \{\Psi_x(y)\},$$

then we have

$$\nabla \Psi_x(T_f(x)) = 0. \tag{B.12}$$

On the other hand, one can write:

$$\begin{aligned}
\mathcal{E}_1(T,f) &= \int [(f(T_f(x)) - c(x,T_f(x))) - (f(T(x)) - c(x,T(x)))] d\mu(x) \\
&= \int [\Psi_x(T_f(x)) - \Psi_x(T(x))] d\mu(x)
\end{aligned}$$

For a fixed x, since $\Psi_x(\cdot) \in C^2(\mathbb{R}^n)$, then

$$\begin{aligned}\Psi_x(T(x)) - \Psi_x(T_f(x)) =& \nabla\Psi_x(T_f(x))(T(x) - T_f(x)) \\ &+ \frac{1}{2}(T(x) - T_f(x))^{\mathrm{T}}\nabla^2\Psi_x(\eta(x))(T(x) - T_f(x))\end{aligned}$$

with $\eta(x) = (1 - \theta_x)T(x) + \theta_x T_f(x)$ for certain $\theta_x \in (0, 1)$. By (Equation B.12) and Lemma B.1.2, we have

$$\Psi_x(T(x)) - \Psi_x(T_f(x)) \leq -\frac{\sigma(x, \eta(x))^2}{2\lambda(\eta(x))}|T(x) - T_f(x)|^2.$$

Thus we have:

$$\mathcal{E}_1(T, f) = \int [\Psi_x(T_f(x)) - \Psi_x(T(x))]\mu(x)(x)\, dx \geq \int \frac{\sigma(x, \eta(x))^2}{2\lambda(\eta(x))}|T(x) - T_f(x)|^2 d\mu(x) \tag{B.13}$$

On the other hand, let us denote the optimal Monge map from $\mu(x)$ to ν as T_*, by Kontorovich duality, we have

$$\sup_f \inf_T \mathcal{L}(T, f) = \inf_{T, T_\sharp\mu(x)=\nu} \int c(x, T(x))d\mu(x) = \int c(x, T_*(x))d\mu(x)$$

Thus we have

$$\begin{aligned}\mathcal{E}_2(f) &= \int c(x, T_*(x))d\mu(x) - \left(\int f(y)d\nu(y) - \int f^{c,-}(x)d\mu(x)\right) \\ &= \int c(x, T_*(x))d\mu(x) - \left(\int f(T_*(x))d\mu(x) - \int f^{c,-}(x)d\mu(x)\right) \\ &= \int [f^{c,-}(x) - (f(T_*(x)) - c(x, T_*(x)))]d\mu(x)\end{aligned}$$

Similar to the previous treatment, we have

$$\mathcal{E}_2(f) = \int [\Psi_x(T_f(x)) - \Psi_x(T_*(x))]d\mu(x)$$

Apply similar analysis as before, we will also have

$$\mathcal{E}_2(f) \geq \int \frac{\sigma(x, \xi(x))^2}{2\lambda(\xi(x))} |T_*(x) - T_f(x)|^2 d\mu(x) \tag{B.14}$$

with $\xi(x) = (1 - \tau_x)T_*(x) + \tau_x T_f(x)$ for certain $\tau_x \in (0, 1)$.

Now we set

$$\beta(x) = \min \left\{ \frac{\sigma(x, \eta(x))}{2\lambda(\eta(x))}, \frac{\sigma(x, \xi(x))}{2\lambda(\xi(x))} \right\}, \tag{B.15}$$

combining (Equation B.13) and (Equation B.14), we obtain

$$\begin{aligned} \mathcal{E}_1(T, f) + \mathcal{E}_2(f) &\geq \int \beta(x)(|T(x) - T_f(x)|^2 + |T_*(x) - T_f(x)|^2) d\mu(x) \\ &\geq \int \frac{\beta(x)}{2} |T(x) - T_*(x)|^2 d\mu(x) \end{aligned}$$

This leads to $\|T - T_*\|_{L^2(\beta\mu(x))} \leq \sqrt{2(\mathcal{E}_1(T, f) + \mathcal{E}_2(f))}$.

□

B.1.3 Experiment details

For general settings, for all experiments we use the Adam optimizer [149] and vanilla feed-forward networks unless specified. The activation functions are all PReLu unless specified.

Unequal dimensions

We ran all experiments in this part on NVIDIA RTX 2080 GPU.

For the incomplete ellipse example, the networks T_θ and f_η each has 5 layers with 10 hidden neurons. The batch size $N = 100$. $K_1 = 6, K_2 = 1$. The learning rate is 10^{-3}. The number of iterations $K = 12000$.

For the ball example, the networks T_θ has 12 layers and f_η has 5 layers. Both of them have 32 hidden neurons. The batch size $N = 100$. $K_1 = 4, K_2 = 1$. The learning rate is 10^{-3}. The number of iterations $K = 15000$.

Effects of different cost

We ran all experiments in this part on CPU.

Decreasing cost function In this example, we set $T_\theta(x) = x + F_\theta(x)$ and optimize over θ. For either $\frac{1}{|x-y|^2}$ or $|x-y|^2$ case we set both F_θ and the Lagrange multiplier f_η as six layers fully connected neural networks, with PReLU and Tanh activation functions respectively, each layer has 36 nodes. The training batch size $N = 2000$. We set $K = 2000$, $K_1 = 8, K_2 = 6$.

On sphere In this example, we set $T_\theta(x) = x + F_\theta(x)$ and optimize over θ. We set both F_θ and the Lagrange multiplier f_η as six layers fully connected neural networks, with PReLU activation functions, each layer has 8 nodes The training batch size $N = 200$. We set $K = 4000$, $K_1 = 8, K_2 = 4$. We choose rather small learning rate in this example to avoid gradient blow up, we set 0.5×10^{-5} as the learning rate for θ and 10^{-5} as the learning rate for η.

Examples in 256D space

We ran all experiments in this part on NVIDIA RTX 2080 GPU. For both L^2 and KL divergence experiments, the network T_θ has 4 layers and f_η has 5 layers. Both of them have 512 hidden neurons. The batch size $N = 100$. $K_1 = 10, K_2 = 1$. The learning rate is 10^{-4}. The number of iterations $K = 10000$. The running time is about 30 minutes.

B.2 Approximating the Optimal Transport Plan via Particle-Evolving Method

B.2.1 Existence and uniqueness of optimal solution to ET problem

In this part, we present the following theorem that guarantees the existence and uniqueness of the solution to the Entropy Transport problem (Equation 3.16):

Theorem B.2.1 (Existence and uniqueness of optimal solution to ET problem)**.** *We consider the general ET problem* (Equation 3.16)*. Suppose the divergence function F satisfies* (Equation 3.11)*,* (Equation 3.12)*, and the cost c satisfies* (Equation 3.13)*. We further assume that there exists at least one $\gamma \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}^d)$ such that $\mathcal{E}(\gamma|\mu,\nu) < +\infty$, and the marginal distribution μ is absolute continuous with respect to the Lebesgue measure $\mathscr{L}^d$ on $\mathbb{R}^d$. Then there exists a unique optimal solution to* (Equation 3.16)*.*

This theorem is a direct result of Theorem 3.3, Corollary 3.6, and Example 3.7 of [75].

B.2.2 Γ-convergence results

Despite the discussion for a fixed Λ, we also establish asymptotic results for (Equation 3.15) as $\Lambda \to +\infty$. We consider $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ equipped with the topology of weak convergence. Before we work on the proofs of Theorem 3.3.2 Theorem 3.3.3, we should briefly introduce the definition of Γ convergence (c.f. Definition 1.5 of [192]):

Definition B.2.1 (Definition of Γ convergence)**.** *Suppose X be a metric space equipped with the distance d. Denote $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Then we say that a sequence $f_n : X \to \bar{\mathbb{R}}$ Γ-converges in X to $f : X \to \bar{\mathbb{R}}$ if for all $x \in X$ we have*

1. *(lim inf inequality) for every sequence $\{x_n\}$ converging to x*

$$f(x) \leq \liminf_n f_n(x_n);$$

2. (lim sup inequality) there exists a sequence $\{x_n\}$ converging to x such that

$$f(x) \geq \limsup_n f_n(x_n).$$

The function f is called the $\Gamma-$limit of $\{f_n\}$, and we write $f = \Gamma - \lim_n f_n$.

We are able to establish the following Γ-convergence results for the functional $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot|\mu,\nu)$ defined on $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$:

Theorem 3.2.2 (Γ-convergence)**.** *Suppose $c(x,y) = |x-y|^2$. Assume that we are given $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, $\mu \ll \mathscr{L}^d, \nu \ll \mathscr{L}^d$, and at least one of μ and ν satisfies the Logarithmic Sobolev inequality with constant $K > 0$. Let $\{\Lambda_n\}$ be a positive increasing sequence, satisfying $\lim_{n\to\infty} \Lambda_n = +\infty$. We consider the sequence of functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$. Recall the functional $\mathcal{E}_\iota(\cdot|\mu,\nu)$ defined in* (Equation 3.10)*. Then $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ Γ- converges to $\mathcal{E}_\iota(\cdot|\mu,\nu)$ on $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$.*

Before we present the proof, we introduce the Logarithmic Sobolev inequality [7]:

Definition B.2.2 (Log-Sobolev)**.** *We say a probability distribution μ satisfying the Logarithmic Sobolev inequality with constant $K > 0$, if for any probability measure $\tilde{\mu} \ll \mu$, we have*

$$\mathcal{D}_{\mathrm{KL}}(\tilde{\mu}\|\mu) \leq \frac{1}{2K}\mathcal{I}(\tilde{\mu}|\mu).$$

Here $\mathcal{I}(\bar{\mu}|\mu)$ is the Fisher information defined as

$$\mathcal{I}(\tilde{\mu}|\mu) = \int \left|\nabla \log\left(\frac{d\tilde{\mu}}{d\mu}\right)\right|^2 d\tilde{\mu}.$$

We also need the following Talagrand inequality [7]:

Theorem B.2.2 (Talagrand)**.** *Suppose $\mu \in \mathscr{P}_2(\mathbb{R}^m)$ satisfies the Logarithmic Sobolev inequality with constant $K > 0$. Then μ also satisfies the following Talagrand inequality:*

for any $\tilde{\mu} \in \mathscr{P}_2(\mathbb{R}^m)$,

$$W_2(\tilde{\mu}, \mu) \leq \sqrt{\frac{2\mathcal{D}_{\mathrm{KL}}(\tilde{\mu}\|\mu)}{K}}. \tag{B.16}$$

Now we can prove Theorem 3.2.2.

Proof of Theorem 3.2.2. First, we notice that $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ equipped with the topology of weak convergence is metrizable by the 2-Wasserstein distance [7]. Thus $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ is metric space and is first countable. For first countable space, we only need to verify the upper bound inequality and the lower bound inequality in order to prove Γ-convergence.

1) Upper bound inequality: For every $\gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$, there is a sequence $\{\gamma_n\}$ converging to γ such that

$$\limsup_{n\to\infty} \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\mu,\nu) \leq \mathcal{E}_\iota(\gamma|\mu,\nu). \tag{B.17}$$

We set $\gamma_n = \gamma$ for all $n \geq 1$, now there are two cases:

(a) If γ doesn't satisfy at least one of the marginal constraints, i.e. $\pi_{1\sharp}\gamma$[1] $\neq \mu$ or $\pi_{2\sharp}\gamma \neq \nu$, then $\mathcal{E}_\iota(\gamma|\mu,\nu) = +\infty$ and the inequality (Equation B.17) definitely holds;

(b) If γ satisfies the marginal constraints, $\pi_{1\sharp}\gamma = \mu$, $\pi_{2\sharp}\gamma = \nu$, then $\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma|\mu,\nu) = \mathcal{E}_\iota(\gamma|\mu,\nu)$, (Equation B.17) also holds.

2) Lower bound inequality: For every sequence $\{\gamma_n\}$ converging to γ,

$$\liminf_{n\to\infty} \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\mu,\nu) \geq \mathcal{E}_\iota(\gamma|\mu,\nu). \tag{B.18}$$

We still separate our discussion into two cases:

[1]Here we define $\pi_1 : \mathbb{R}^d \times \mathbb{R}^d :\to \mathbb{R}^d, (x,y) \mapsto x$ as the projection onto the first coordinate component; similarly π_2 is the projection onto the second component.

(a) If γ satisfies the marginal constraints, we have:

$$\begin{aligned}
&\liminf_{n\to\infty} \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\ \mu,\nu)\\
=&\liminf_{n\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\gamma_n(x,y) + \Lambda_n\mathcal{D}_{\mathrm{KL}}(\pi_{1\sharp}\gamma_n\|\mu) + \Lambda_n\mathcal{D}_{\mathrm{KL}}(\pi_{2\sharp}\gamma_n\|\nu)\\
\geq&\liminf_{n\to\infty} \int_{\mathcal{M}\times\mathcal{M}} c(x,y)d\gamma_n(x,y)\\
=&\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\gamma(x,y)\\
=&\mathcal{E}_\iota(\gamma|\ \mu,\nu).
\end{aligned}$$

Here we use the fact that $\mathcal{D}_{\mathrm{KL}}(\mu_1\|\mu_2) \geq 0$ for any $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R}^d)$ with densities.

(b) If γ doesn't satisfy at least one of the marginal constraints, without loss of generality, assume that $W_2(\pi_{1\sharp}\gamma, \mu) = \delta > 0$. We have:

$$W_2(\pi_{1\sharp}\gamma,\mu) \leq W_2(\pi_{1\sharp}\gamma,\pi_{1\sharp}\gamma_n) + W_2(\pi_{1\sharp}\gamma_n,\mu) \leq W_2(\gamma,\gamma_n) + W_2(\pi_{1\sharp}\gamma_n,\mu).$$

We can choose large enough N such that when $n > N$, $W_2(\gamma,\gamma_n) \leq \delta/2$, then we have $W_2(\pi_{1\sharp}\gamma_n,\mu) \geq \delta/2$.

According to Talagrand inequality (Equation B.16), we have:

$$\sqrt{\frac{2\mathcal{D}_{\mathrm{KL}}(\pi_{1\sharp}\gamma_n\|\mu)}{K}} \geq W_2(\pi_{1\sharp}\gamma_n,\mu) \geq \frac{\delta}{2},$$

i.e., when $n > N$, $\mathcal{D}_{\mathrm{KL}}(\pi_{1\sharp}\gamma_n\|\mu) \geq K\frac{\delta^2}{8}$. This implies:

$$\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\mu,\nu) \geq \Lambda_n K\frac{\delta^2}{8}.$$

Therefore we show that:

$$\liminf_{n\to\infty} \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\mu,\nu) = +\infty = \mathcal{E}_\iota(\gamma|\mu,\nu).$$

Thus, combining (a) and (b), we have proved (Equation B.18). Combining (Equation B.17), (Equation B.18), we have shown that $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ Γ-converges to $\mathcal{E}_\iota(\cdot|\mu,\nu)$. $\square$

We then establish the equi-coercive property for $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}_n$. We can apply the Fundamental Theorem of Γ-convergence [81] [82] to establish the following asymptotic result:

Theorem 3.2.3 (Property of Γ-convergence)**.** *Suppose $c(x,y) = |x-y|^2$. Assuming $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$, $\mu \ll \mathscr{L}^d, \nu \ll \mathscr{L}^d$, and both μ,ν satisfy the Logarithmic Sobolev inequality with constants $K_\mu, K_\nu > 0$. According to Corollary 3.3.1.1, problem* (Equation 3.15) *with functional $\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)$ admits a unique optimal solution, let us denote it as γ_n. According to Theorem 2.1.2, the Kantorovich problem* (Equation 2.6) *also admits a unique solution, we denote it as γ_{OT}. Then $\lim_{n\to\infty}\gamma_n = \gamma_{OT}$ in $\mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$.*

Before we prove this theorem, we introduce the definition of equi-coerciveness:

Definition B.2.3. *A family of functions $\{F_n\}$ on X is said to be equi- coercive, if for every $\alpha\in\mathbb{R}$, there is a compact set C_α such that the sublevel sets $\{F_n \le \alpha\} \subset C_\alpha$ for all n.*

To prove Theorem 3.2.3, we first establish the following two lemmas:

Lemma B.2.3. *Suppose $d_0 > 0$. Denote*

$$C = \{\gamma \in \mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d)\ |W_2(\pi_{1\sharp}\gamma,\mu) \le d_0,\ W_2(\pi_{2\sharp}\gamma,\nu) \le d_0\}.$$

Then C is compact set of $\mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$. Recall that $\mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$ is equipped with the topology of weak convergence.

Proof of the Lemma B.2.3. According to Prokhorov's Theorem [193], we only need to show that C is tight. That is: for any $\epsilon > 0$, we can find a compact set $E_\epsilon \subset \mathbb{R}^d\times\mathbb{R}^d$, such that

$$\gamma(E_\epsilon) \ge 1-\epsilon \quad \forall\,\gamma\in C.$$

Let us denote $B_R^d \subset \mathbb{R}^d$ as the ball centered at origin with radius R in $\mathbb{R}^d$. Since μ, ν are probability measures, for arbitrary $\epsilon > 0$, we can pick $R(\mu, \epsilon), R(\nu, \epsilon) > 0$ such that

$$\mu(B^d_{R(\mu,\epsilon)}) \geq 1 - \epsilon, \quad \nu(B^d_{R(\nu,\epsilon)}) \geq 1 - \epsilon.$$

Now for any chosen $\epsilon > 0$, we choose

$$R = \sqrt{\frac{4d_0^2}{\epsilon}} \quad \text{and} \quad \tilde{R} = \sqrt{(R(\mu, \frac{\epsilon}{4}) + R)^2 + (R(\nu, \frac{\epsilon}{4}) + R)^2}.$$

Now we prove $\gamma(B^{2d}_{\tilde{R}}) \geq 1 - \epsilon$ for any $\gamma \in C$:

Denote $\gamma_1 = \pi_{1\sharp}\gamma$, let γ_{OT} be the optimal coupling of γ_1 and μ, i.e.

$$\gamma_{OT} = \underset{\pi \in \Pi(\gamma_1, \mu)}{\operatorname{argmin}} \left\{ \iint c(x, y)\, d\pi(x, y) \right\}.$$

Then (here, we denote $R_\mu = R(\mu, \frac{\epsilon}{4})$ for short hand):

$$\begin{aligned} d_0^2 \geq W_2^2(\gamma_1, \mu) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x - y|^2 d\gamma_{OT}(x, y) &\geq \int_{\overline{B^d_{R_\mu + R}}} \int_{B^d_{R_\mu}} |x - y|^2 \, d\gamma_{OT}(x, y) \\ &\geq R^2 \int_{\overline{B^d_{R_\mu + R}}} \int_{B^d_{R_\mu}} d\gamma_{OT}(x, y). \end{aligned}$$

This gives:

$$\int_{\overline{B^d_{R_\mu + R}}} \int_{B^d_{R_\mu}} d\gamma_{OT}(x, y) \leq \frac{d_0^2}{R^2} = \frac{\epsilon}{4}. \tag{B.19}$$

On the other hand, one have:

$$\int_{\overline{B^d_{R_\mu + R}}} \int_{\overline{B^d_{R_\mu}}} d\gamma_{OT}(x, y) \leq \int_{\overline{B^d_{R_\mu}}} d\mu(y) = 1 - \mu(B^d_{R_\mu}) \leq \frac{\epsilon}{4}. \tag{B.20}$$

Now sum (Equation B.19) and (Equation B.20) together, we have:

$$\gamma_1 \left(\overline{B^d_{R_\mu + R}} \right) = \int_{\overline{B^d_{R_\mu + R}}} \int_{\mathbb{R}^d} d\gamma_{OT} = \iint_{\overline{B^d_{R_\mu + R}} \times B^d_{R_\mu}} d\gamma_{OT} + \iint_{\overline{B^d_{R_\mu + R}} \times \overline{B^d_{R_\mu}}} d\gamma_{OT} \leq \frac{\epsilon}{2}.$$

Similarly, denote $\gamma_2 = \pi_{2\sharp}\gamma$, we have:

$$\gamma_2\left(\overline{B^d_{R_\nu+R}}\right) \leq \frac{\epsilon}{2}.$$

As a result, for any $\epsilon > 0$, we can pick the compact ball $B^{2d}_{\tilde{R}} \subset \mathbb{R}^d \times \mathbb{R}^d$, so that for any $\gamma \in C$,

$$\begin{aligned}\gamma(B^{2d}_{\tilde{R}}) = 1 - \gamma\left(\overline{B^{2d}_{\tilde{R}}}\right) \geq & 1 - \gamma\left(\left(\overline{B^d_{R_\mu+R}} \times \mathbb{R}^d\right) \bigcup \left(\mathbb{R}^d \times \overline{B^d_{R_\nu+R}}\right)\right) \\ \geq & 1 - \gamma_1\left(\overline{B^d_{R_\mu+R}}\right) - \gamma_2\left(\overline{B^d_{R_\nu+R}}\right) \geq 1 - \epsilon, \end{aligned} \tag{B.21}$$

here we are using the fact:

$$\overline{B^{2d}_{\tilde{R}}} \subset \left(\overline{B^d_{R_\mu+R}} \times \mathbb{R}^d\right) \bigcup \left(\mathbb{R}^d \times \overline{B^d_{R_\nu+R}}\right).$$

The inequality (Equation B.21) proves the tightness of set C and thus C is compact set in $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$. □

Lemma B.2.4. *Assuming $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and both μ, ν satisfies the Logarithmic Sobolev inequality with constants $K_\mu, K_\nu > 0$. The sequence of functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ defined on $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ with positive increasing sequence $\{\Lambda_n\}$ is equi-coercive.*

proof of (Lemma B.2.4). By Talagrand inequality (Equation B.16) involving μ, ν:

$$\mathcal{D}_{\mathrm{KL}}(\rho\|\mu) \geq \frac{K_\mu}{2} W_2^2(\rho,\mu) \quad \mathcal{D}_{\mathrm{KL}}(\rho\|\nu) \geq \frac{K_\nu}{2} W_2^2(\rho,\nu) \quad \forall\, \rho \in \mathscr{P}_2(\mathbb{R}^d).$$

Thus,

$$\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma|\mu,\nu) \geq \Lambda_1\left(\frac{K_\mu}{2} W_2^2(\pi_{1\sharp}\gamma,\mu) + \frac{K_\nu}{2} W_2^2(\pi_{1\sharp}\gamma,\nu)\right).$$

For any $\alpha \geq 0$, we set $d_0 = \max\{\sqrt{\frac{2\alpha}{K_\mu \Lambda_1}}, \sqrt{\frac{2\alpha}{K_\nu \Lambda_1}}\}$, then

$$\left\{\gamma \in \mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \mid \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma|\mu,\nu) \leq \alpha\right\} \subset \underbrace{\left\{\gamma \;\middle|\; W_2(\pi_{1\sharp}\gamma, \mu) \leq d_0, W_2(\pi_{2\sharp}\gamma, \nu) \leq d_0\right\}}_{\text{denote as } C_\alpha}.$$

By Lemma B.2.3, C_α is compact in $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ for any α (for $\alpha < 0$, we simply get empty set and thus is also compact set). Thus the sequence of functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ is equi-coercive. □

Now our proof mainly rely on the following fundamental theorem of Γ-convergence [81] [82]:

Theorem B.2.5. *Let (X, d) be a metric space, let $\{F_{\theta_n}\}$ with $\theta_n \to +\infty$ be an equi-coercive sequence of functionals on X, assume $\{F_{\theta_n}\}$ Γ-converge to the functional F defined on X; Then*

$$\exists \min_X F = \lim_{n\to\infty} \inf_X F_{\theta_n}.$$

Moreover, if $\{x_n\}$ is a precompact sequence such that x_n is the minimizer of F_{θ_n}: $F_{\theta_n}(x_n) = \inf_X F_{\theta_n}$, then every limit of a subsequence of $\{x_n\}$ is a minimum point for F.

We can now prove Theorem 3.2.3.

Proof. We apply Theorem B.2.5 to the sequence of functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}_n$ defined on probability space equipped with 2-Wasserstein metric $(\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d), W_2)$, by Lemma B.2.4 , we know that $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}_n$ is equi-coercive. And by Theorem 3.2.2, $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}\}(\cdot|\mu,\nu)\}_n$ Γ-converge to $\mathcal{E}_\iota(\cdot|\mu,\nu)$. Recall that γ_n is the unique minimizer of $\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)$, we are going to show that $\{\gamma_n\}$ is precompact sequence in $\mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$: We define

$$\alpha = \iint c(x,y)\, d(\mu \otimes \nu) = \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\mu \otimes \nu|\mu,\nu) \quad \forall\, n \geq 1.$$

Then we have $\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma_n|\mu,\nu) \le \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\mu\otimes\nu|\mu,\nu) = \alpha$ for all n, thus

$$\gamma_n \in \left\{\gamma \mid \gamma \in \mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d),\ \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma|\mu,\nu) \le \alpha\right\} \quad \forall\, n \ge 1.$$

Now since $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot|\mu,\nu)\}$ is equi-coercive, we can pick compact C_α such that:

$$\left\{\gamma \mid \gamma \in \mathscr{P}_2(\mathbb{R}^d\times\mathbb{R}^d),\ \mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma|\mu,\nu) \le \alpha\right\} \subset C_\alpha \quad \forall\, n \ge 1.$$

Thus all $\{\gamma_n\}$ lie in the compact set C_α and $\{\gamma_n\}$ is precompact.

Now Theorem 3.2.3 asserts that any limit point of $\{\gamma_n\}$ is a minimum point of $\mathcal{E}_\iota(\cdot|\mu,\nu)$, however, $\mathcal{E}_\iota(\cdot,\mu,\nu)$ admits unique minimizer γ_{OT}, we have proved $\lim_{n\to\infty}\gamma_n = \gamma_{OT}$. $\square$

B.2.3 Gradient flow of constrained Entropy Transport functional

We compute the gradient flow of $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot|\mu,\nu)$ on $\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$. We assume every thing is in the form of Radon-Nikodym Derivative, i.e. we assume $\rho = \frac{d\gamma}{d\mathscr{L}^{2d}}$ and $\varrho_1 = \frac{d\mu}{d\mathscr{L}^d}$, $\varrho_2 = \frac{d\nu}{d\mathscr{L}^d}$. We denote $\rho_1 = \frac{d\pi_{1\sharp}\gamma}{d\mathscr{L}^d}, \rho_2 = \frac{d\pi_{2\sharp}\gamma}{d\mathscr{L}^d}$, then $\rho_1 = \int \rho\, dy$, $\rho_2 = \int \rho\, dx$. We write the functional $\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma|\mu,\nu)$ as $E(\rho)$ for shorthand, then:

$$E(\rho) = \iint_{\mathbb{R}^d\times\mathbb{R}^d}\left(c(x,y) + \Lambda\log\left(\frac{\rho_1(x)}{\varrho_1(x)}\right) + \Lambda\log\left(\frac{\rho_2(y)}{\varrho_2(y)}\right)\right)\rho(x,y)\,dxdy.$$

To compute L^2 variation of E , suppose $\rho > 0$ and consider arbitrary $\sigma \in C_0(\mathbb{R}^d\times\mathbb{R}^d)$. We denote $\sigma_1(x) = \int \sigma(x,y)\,dy$, $\sigma_2(y) = \int \sigma(x,y)\,dx$. We compute $\frac{d}{dh}E(\rho+h\sigma)\Big|_{h=0}$ as:

$$\begin{aligned}
&\frac{d}{dh}E(\rho+h\sigma)\Big|_{h=0}\\
=&\frac{d}{dh}\left[\iint (c(x,y) + \Lambda\log(\frac{\rho_1(x)+h\sigma_1(x)}{\varrho_1(x)}) + \Lambda\log(\frac{\rho_2(y)+h\sigma_2(y)}{\varrho_2(y)}))(\rho+h\sigma)\,dxdy\right]_{h=0}\\
=&\iint\left(2\Lambda + c(x,y) + \Lambda\log(\frac{\rho_1(x)}{\varrho_1(x)}) + \Lambda\log(\frac{\rho_2(y)}{\varrho_2(y)})\right)\sigma\,dxdy.
\end{aligned}$$

Since $\frac{dE(\rho+h\sigma)}{dh}\Big|_{h=0} = \langle \frac{\delta E(\rho)}{\delta \rho}, \sigma \rangle$, we can thus identify that:

$$\frac{\delta E(\rho)}{\delta \rho} = 2\Lambda + c(x,y) + \Lambda \log\left(\frac{\rho_1(x)}{\varrho_1(x)}\right) + \Lambda \log\left(\frac{\rho_2(y)}{\varrho_2(y)}\right).$$

Thus, plugging this result in (Equation 2.62), one can derive:

$$\frac{\partial \rho(x,y,t)}{\partial t} = \nabla \cdot \left(\rho(x,y,t)\nabla\left(2\Lambda + c(x,y) + \Lambda \log\left(\frac{\rho_1(x,t)}{\varrho_1(x)}\right) + \Lambda \log\left(\frac{\rho_2(y,t)}{\varrho_2(y)}\right)\right)\right).$$

Notice that ∇ means gradient with respect to both variables x and y, i.e. $\nabla f = \begin{bmatrix} \nabla_x f \\ \nabla_y f \end{bmatrix}$ for function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and $\nabla \cdot \vec{v} = \nabla_x \cdot \vec{v}_1 + \nabla_y \cdot \vec{v}_2$ for vector field $\vec{v} = \begin{bmatrix} \vec{v}_1 \\ \vec{v}_2 \end{bmatrix}$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ with $\vec{v}_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$; $\vec{v}_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$.

Then this equation will simplify to:

$$\frac{\partial \rho(x,y,t)}{\partial t} = \nabla \cdot \left(\rho(x,y,t)\nabla\left(c(x,y) + \Lambda \log\left(\frac{\rho_1(x,t)}{\varrho_1(x)}\right) + \Lambda \log\left(\frac{\rho_2(y,t)}{\varrho_2(y)}\right)\right)\right).$$

Which is exactly equation (Equation 3.19).

B.2.4 Algorithm

Direct approximation of $\nabla \log \rho(x)$ via blobing method will be expensive. We apply the Random Batch Methods (RBM) [194] here to reduce the computational effort. We divide N total particles into m batches equally. Then we approximate $\nabla \log \rho(X_i)$ by using the particles in the same batch as X_i. Now in each time step, the computational complexity will be reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n^2/m)$.

Our method with RBM implementation is summarized in the following algorithm.

Algorithm 3 Random Batch Particle Evolution Algorithm

Input: The density functions of the marginals ϱ_1, ϱ_2, timestep Δt, total number of iterations T, parameters of the chosen kernel K
Initialize: The initial locations of all particles $X_i(0)$ and $Y_i(0)$ where $i = 1, 2, \cdots, n$,
for t = 1,2,$\cdots$,T **do**
Shuffle the particles and divide them into m batches: $\mathcal{C}_1, \cdots, \mathcal{C}_m$
for each batch $\mathcal{C}_q$ **do**
Update the location of each particle (X_i, Y_i) $(i \in \mathcal{C}_q)$ according to (Equation 3.22)
end for
end for
Output: A sample approximation of the optimal coupling: $X_i(T), Y_i(T)$ for $i = 1, 2, \cdots, n$

B.3 Learning High Dimensional Wasserstein Geodesics

B.3.1 Proposed method

Proof of Theorem 3.4.1 Let us denote vector field $F : \mathbb{R}^d \to \mathbb{R}^d$. We denote $\hat{\rho}_t = (I + tF)_\sharp \rho_a$. As formulated in (Equation 3.30), we consider the following functional $\mathcal{L}$ of F and Φ:

$$\begin{aligned}
\mathcal{L}(F, \Phi) =& \hat{\mathfrak{L}}((\mathrm{Id} + tF)_\sharp \rho_a, \Phi) \\
=& \int_0^1 \int \left(-\frac{\partial \Phi(x,t)}{\partial t} - H(\nabla \Phi(x,t)) \right) \hat{\rho}(x,t)\, dxdt \\
&+ \int \Phi(x,1)\rho_b(x) - \Phi(x,0)\rho_a(x)\, dx \\
=& \int_0^1 \int \left(-\frac{\partial \Phi(x+tF(x),t)}{\partial t} - H(\nabla \Phi(x+tF(x),t)) \right) \rho_a(x,t)\, dxdt \\
&+ \int \Phi(x,1)\rho_b(x) - \Phi(x,0)\rho_a(x)\, dx
\end{aligned} \tag{B.22}$$

As stated in Theorem 3.4.1, the optimal solution obtained from dynamical OT problem (Equation 2.20) is a critical point to the functional $\mathcal{L}(F, \Phi)$. Before we prove this result, we need the following lemmas:

Lemma B.3.1. *Given a distribution with density ρ defined on $\mathbb{R}^d$, consider vector field*

$F : \mathbb{R}^d \to \mathbb{R}^d$. *Define time-varying density* $\{\rho(\cdot,t)\}_{t\in[0,1]}$ *as* $\rho(\cdot,t) = (Id + tF)_\sharp \rho_0$. *Suppose for a given* $f \in C^1(\mathbb{R}^d)$, $f(x)\rho(x,t)$ *is integrable on* $\mathbb{R}^d$. *Then*

$$\int f(x)\frac{\partial}{\partial t}\rho(x,t) = \int \nabla f(x+tF(x))\cdot F(x)\ \rho_a(x)\ dx$$

Proof. We have

$$\begin{aligned}\int f(x)\frac{\partial}{\partial t}\rho(x,t) = \frac{d}{dt}\left(\int f(x)\rho(x,t)\ dx\right) &= \frac{d}{dt}\left(\int f(x+tF(x))\rho_a(x)\ dx\right)\\ &= \int \nabla f(x+F(x))\cdot F(x)\ \rho_a(x)\ dx\end{aligned}$$

□

Lemma B.3.2. *Suppose* $\Phi^*(x,t)$ *is solved from the geodesic equation system* (Equation 2.36) *and* (Equation 2.37). *Denote* $\Phi_0^*(\cdot) = \Phi^*(\cdot,0)$, *we further assume* $\Phi^*(\cdot,t) \in C^2(\mathbb{R}^d)$. *Then we have*

$$\nabla\Phi^*(x + t\ \nabla L^{-1}(\nabla\Phi_0^*(x)),\ t) = \nabla\Phi_0^*(x). \tag{B.23}$$

Proof. Now consider Hamilton-Jacobi equation as stated in (Equation 2.37):

$$\frac{\partial\Phi^*(y,t)}{\partial t} + H(\nabla\Phi^*(y,t)) = 0 \quad \Phi^*(\cdot,0) = \Phi_0^*.$$

We take gradient with respect to x on both sides, we have

$$\frac{\partial}{\partial t}(\nabla\Phi^*(y,t)) + \nabla^2\Phi^*(y,t)\nabla H(\nabla\Phi^*(y,t)) = 0. \tag{B.24}$$

Let us denote $T_t(x) = x + t\nabla H(\nabla\Phi_0^*(x))$ for simplicity. We now compute

$$\frac{d}{dt}\nabla\Phi^*(T_t(x),t) = \frac{\partial}{\partial t}\nabla\Phi^*(T_t(x),t) + \nabla^2\Phi^*(T_t(x),t)\nabla H(\nabla\Phi_0^*(x))$$

By plugging $y = T_t(x)$ into (Equation B.24), we are able to verify $\frac{d}{dt}\nabla\Phi^*(T_t(x),t) = 0$.

Thus

$$\nabla\Phi^*(T_t(x),t) = \nabla\Phi^*(T_0(x),0) = \nabla\Phi_0^*(x) \quad \text{for } t \in [0,1] \tag{B.25}$$

Recall H and $\nabla H = \nabla L^{-1}$ stated in Definition 2.1.2, then (Equation B.25) leads to

$$\nabla\Phi^*(x + t\,\nabla L^{-1}(\nabla\Phi_0^*(x)),\ t) = \nabla\Phi_0^*(x).$$

□

Lemma B.3.3. *We keep all the notations and assumptions about Φ^* stated in Lemma B.3.2. Now denote $\hat{\rho}(\cdot,t) = (Id + t\nabla L^{-1}(\nabla\Phi_0^*))_\sharp\rho_a$. Then $\hat{\rho}(\cdot,t)$ solves*

$$\frac{\partial\hat{\rho}(x,t)}{\partial t} + \nabla\cdot(\hat{\rho}(x,t)\nabla L^{-1}(\nabla\Phi^*(x,t))) = 0.$$

Proof. For arbitrary $f \in C_0^\infty(\mathbb{R}^d)$, we consider:

$$\begin{aligned}&\int f(x)\left(\frac{\partial\hat{\rho}(x,t)}{\partial t} + \nabla\cdot(\hat{\rho}(x,t)\nabla L^{-1}(\nabla\Phi^*(x,t)))\right)\,dx\\ =&\int f(x)\frac{\partial\hat{\rho}(x,t)}{\partial t}\,dx - \int\nabla f(x)\cdot\nabla L^{-1}(\nabla\Phi^*(x,t))\hat{\rho}(x,t)\,dx\end{aligned}$$

By Lemma B.3.1, the first term equals

$$\int\nabla f(x + t\nabla L^{-1}(\nabla\Phi_0^*(x)))\cdot\nabla L^{-1}(\nabla\Phi_0^*(x))\ \rho_a(x)\,dx \tag{B.26}$$

The second term equals

$$\int\nabla f(x + t\nabla L^{-1}(\nabla\Phi_0^*(x)))\cdot\nabla L^{-1}(\nabla\Phi^*(x + t\nabla L^{-1}(\nabla\Phi_0^*(x)),t))\ \rho_a(x)\,dx \tag{B.27}$$

Using Lemma B.3.2, we know the integrals (Equation B.26) and (Equation B.27) are the

same. Thus we have

$$\int f(x)\left(\frac{\partial\hat{\rho}(x,t)}{\partial t}+\nabla\cdot(\hat{\rho}(x,t)\nabla L^{-1}(\nabla\Phi^*(x,t)))\right)\,dx=0\quad\forall\ f\in C_0^\infty(\mathbb{R}^d).$$

This leads to our result. □

Lemma B.3.4. *We keep all the notations and assumptions about Φ^* stated in Lemma B.3.2. Recall $C_{Dym}(\rho_a,\rho_b)$ denotes the optimal value of* (Equation 2.20)*, then*

$$C_{Dym}(\rho_a,\rho_b)=\int L(\nabla L^{-1}(\nabla\Phi_0^*(x)))\rho_a(x)\,dx. \tag{B.28}$$

Proof. Consider particle dynamical OT with its optimal solution

$$v^*(x,t)=\nabla L^{-1}(\nabla\Phi^*(x,t))$$

as stated in (Equation 2.31). Recall Theorem 2.1.6 stating that the optimal plan is transporting each particle $\boldsymbol{X}_t$ along straight lines with constant velocity

$$v^*(\boldsymbol{X}_0,0)=\nabla L^{-1}(\nabla\Phi_0^*(\boldsymbol{X}_0)).$$

Combining these, we obtain

$$\begin{aligned}C_{\text{Dym}}(\rho_a,\rho_b)=\int_0^1\mathbb{E}\,L(v^*(\boldsymbol{X}_t,t))\,dt=&\mathbb{E}\left(\int_0^1 L(v^*(\boldsymbol{X}_t,t))\,dt\right)\\=&\mathbb{E}\,L(\nabla L^{-1}(\nabla\Phi_0^*(\boldsymbol{X}_0))).\end{aligned}$$

Notice that we require $\boldsymbol{X}_0\sim\rho_a$. This will lead to (Equation B.28). □

Now we are able to prove Theorem 3.4.1:

Theorem 3.3.1. *Denote the optimal solution to* (Equation 2.20) *as $(\rho^*(x,t),\Phi^*(x,t))$. Set $\Phi_0^*(\cdot)=\Phi^*(\cdot,0)$. Assume $\Phi^*(\cdot,t)\in C^2(\mathbb{R}^d)$, then $(\nabla L^{-1}(\nabla\Phi_0^*),\Phi^*)$ is the critical point*

to functional $\mathcal{L}$, i.e.

$$\frac{\partial \mathcal{L}}{\partial F}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0, \quad \frac{\partial \mathcal{L}}{\partial \rho}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0.$$

Furthermore, $\mathcal{L}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = C_{Dym}(\rho_a, \rho_b)$, where $C_{Dym}(\rho_a, \rho_b)$ is denoted as the optimal value of (Equation 2.20)*. By subsection 2.1.6, this is exactly the OT distance between ρ_a and ρ_b with cost function $c(x, y) = L(x - y)$.*

Proof. Since we have assumed $\Phi^*(\cdot, t) \in C^2(\mathbb{R}^d)$, we restrict our $\Phi \in C^2(\mathbb{R}^d)$ as well. We first rewrite $\mathcal{L}(F, \Phi)$ by using integration by parts as:

$$\int_0^1 \int \Phi(x,t) \frac{\partial \hat{\rho}(x,t)}{\partial t} - H(\nabla \Phi(x,t)) \hat{\rho}(x,t)\, dxdt + \int \Phi(x,1)(\rho_a(x) - \hat{\rho}(x,1))\, dx. \tag{B.29}$$

By Lemma B.3.1, (Equation B.29) can be written as

$$\begin{aligned} \mathcal{L}(F, \Phi) = &\int_0^1 \int_{\mathbb{R}^d} [\nabla \Phi(x + tF(x), t) \cdot F(x) - H(\nabla \Phi(x + tF(x), t))] \rho_a(x)\, dxdt \\ &+ \int \Phi(x,1) \rho_b(x)\, dx - \int \Phi(x + F(x), 1) \rho_a(x)\, dx. \end{aligned} \tag{B.30}$$

Now based on (Equation B.30) here, we are able to compute $\frac{\partial \mathcal{L}(F,\Phi)}{\partial F}(x)$ as

$$\begin{aligned} \frac{\partial \mathcal{L}(F, \Phi)}{\partial F} = &\int_0^1 t \nabla^2 \Phi(x + tF(x), t) \cdot \underbrace{[F(x) - \nabla H(\nabla \Phi(x + tF(x), t))]}_{(A)} \rho_a(x)\, dt \\ &+ \underbrace{\left(\int_0^1 \nabla \Phi(x + tF(x), t)\, dt - \nabla \Phi(x + F(x), 1) \right)}_{(B)} \rho_a(x). \end{aligned} \tag{B.31}$$

Now we plug $F = \nabla L^{-1}(\nabla \Phi_0^*)$, $\Phi = \Phi^*$ into (Equation B.31), by Lemma B.3.2, we have

$$\nabla \Phi^*(x + t\, \nabla L^{-1}(\nabla \Phi_0^*(x)),\ t) = \nabla \Phi_0^*(x). \tag{B.32}$$

Then using (Equation B.32) and recall that $\nabla H = \nabla L^{-1}$, one can verify that $(A) = 0$, similarly, for (B), we have $\nabla \Phi(x + tF(x), t) = \nabla \Phi_0^*$ for all $t \in [0, 1]$. Thus $(B) = 0$ and we are able to verify $\frac{\partial \mathcal{L}}{\partial F}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0$.

On the other hand, we can compute $\frac{\partial \mathcal{L}(F,\Phi)}{\partial \Phi}(x, t)$ as

$$\frac{\delta \mathcal{L}(F, \Phi)}{\delta \Phi}(x, t) = \underbrace{\left[\frac{\partial \hat{\rho}(x, t)}{\partial t} + \nabla \cdot (\hat{\rho}(x, t) \nabla H(\nabla \Phi(x, t)))\right]}_{(C)} + \underbrace{[\rho_b(x) - \hat{\rho}(x, 1)]}_{(D)} \delta_1(t)$$

Now by Lemma B.3.3, we know $(C) = 0$. Furthermore, since Φ^* solves dynamical OT problem associated to the optimal transport problem between ρ_a and ρ_b, by (Equation 2.46), we have $\hat{\rho}(x, 1) = (Id + \nabla L^{-1}(\nabla \Phi_0^*(\cdot)))_\sharp \rho_a = \rho_b$, this verifies $(D) = 0$. Thus, we are able to verify $\frac{\partial \mathcal{L}(F,\Phi)}{\partial \Phi}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = 0$.

At last, we plug $F = \nabla L^{-1}(\nabla \Phi_0^*), \Phi = \Phi^*$ in (Equation B.30) to obtain:

$$\begin{aligned}
\mathcal{L}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) &= \int_0^1 \int \nabla \Phi_0^*(x) \cdot \nabla L^{-1}(\nabla \Phi_0^*(x)) - H(\nabla \Phi_0^*(x))\ \rho_a(x)\ dxdt \\
&= \int L(\nabla L^{-1}(\nabla \Phi_0^*(x))) \rho_a(x)\ dx.
\end{aligned}$$

Now by Lemma B.3.4, we have verified $\mathcal{L}(\nabla L^{-1}(\nabla \Phi_0^*), \Phi^*) = C_{\text{Dym}}(\rho_a, \rho_b)$. □

B.3.2 Preconditioning technique for 2-Wasserstein case

It's worth mentioning that we can apply preconditioning technique under the 2-Wasserstein cases, i.e., $L(\cdot) = \frac{|\cdot|^2}{2}$. When the support of distributions ρ_a and ρ_b are far away from each other, the computational process might get much more sensitive with respect to vector field F. In order to deal with this situation, we consider preconditioning to our initial distribution

ρ_a. In our implementation, we treat $P : \mathbb{R}^d \to \mathbb{R}^d$ as our preconditioning map. We fix its structure as $P(x) = \sigma x + \mu$ with $\sigma \in \mathbb{R}_+,\ \mu \in \mathbb{R}^d$. Such preconditioning process can be treated as an operation aiming at relocating and rescaling the initial distribution ρ_a so that the support of $P_\sharp \rho_a$ matches with the support of ρ_b in a better way, which in turn facilitates the training process of our OT problem.

Let us denote the optimal vector field of OT problem between $P_\sharp \rho_a$ and ρ_b as $\nabla \widehat{\Phi}_0$, then for the vector field $F_*(x) = \nabla \widehat{\Phi}_0 \circ P(x) + P(x) - x$, the following theorem guarantees the optimality of F_*.

Theorem B.3.5. *Suppose* $L(\cdot) = \frac{|\cdot|^2}{2}$. *We define the map* $P(x) = \sigma x + \mu$ *with* $\sigma \in \mathbb{R}_+, \mu \in \mathbb{R}^d$. *Recall* (Equation 2.31)*, we denote* $v(x,t) = \nabla \widehat{\Phi}(x,t)$ *as the optimal solution to dynamical OT problem* (Equation 2.20) *from* $P_\sharp \rho_a$ *to* ρ_b. *We set* $\widehat{\Phi}_0 = \widehat{\Phi}(\cdot, 0)$. *Furthermore, we denote* $v(x,t) = \nabla \Phi^*(x,t)$ *as the optimal solution to dynamical OT problem* (Equation 2.20) *from* ρ_a *to* ρ_b*, and set* $\Phi_0^* = \Phi^*(\cdot, 0)$. *Then we have*

$$\nabla \Phi_0^*(x) = \nabla \widehat{\Phi}_0 \circ P(x) + P(x) - x, \ \ \Phi_0^*(x) = \frac{1}{\sigma} \widehat{\Phi}_0(\sigma x + \mu) + \frac{\sigma - 1}{2} |x|^2 + \mu^T x + \textit{Const}.$$

This theorem indicates that our constructed F_* is exactly the optimal transport field $\nabla \Phi_0^*$ for the original OT problem from ρ_a to ρ_b.

Proof. According to (Equation 2.48), we have

$$(\mathrm{Id} + \nabla \widehat{\Phi}_0)_\sharp (P_\sharp \rho_a) = \rho_b$$

This yields

$$(P + \nabla \Phi_0 \circ P)_\sharp \rho_a = \rho_b$$

We rewrite this as

$$(\mathrm{Id} + \nabla \widehat{\Phi}_0 \circ P + P - \mathrm{Id})_\sharp \rho_a = \rho_b \tag{B.33}$$

We denote $u(x) = \frac{1}{\sigma}\widehat{\Phi}_0(\sigma x + \mu) + \frac{\sigma-1}{2}|x|^2 + \mu^T x$.

Then we can directly verify that

$$\nabla u(x) = \nabla\widehat{\Phi}_0(\sigma x + \mu) + (\sigma x + \mu) - x = \nabla\widehat{\Phi}_0 \circ P(x) + P(x) - x$$

Plug this into (Equation B.33) above we get:

$$(\mathrm{Id} + \nabla u)_\sharp \rho_a = \rho_b$$

Using the uniqueness of the solution to Monge-Ampere equation, we have $\Phi_0^* = u + \text{Const}$, or equivalently, $\nabla\Phi_0^*(x) = \nabla u(x) = \nabla\widehat{\Phi}_0 \circ P(x) + P(x) - x$ □

B.3.3 Algorithm

Our computation procedure is summarized in Algorithm 4. We set $F_{\theta_1}, G_{\theta_2}$ and $\Phi^F_{\omega_1}, \Phi^G_{\omega_2}$ as fully connected neural networks and optimize over their parameters.

Remark 33. *In Algorithm 4, we need to sample points $\{z_k^a\}$ from the distribution $\hat{\rho}_a = P_\sharp \rho_a$. To achieve this, we first sample $\{u_k\}$ from ρ_a. Then $\{P(u_k)\}$ are our desired samples from $\hat{\rho}_a$.*

In Algorithm 4 we define:

$$\mathcal{L}^{ab}(\Phi^F_{\omega_1}) = -\frac{1}{N}\sum_{k=1}^{N}\left[\frac{\partial}{\partial t}\Phi^F_{\omega_1}(x_k^a, t_k^a) + H(\nabla\Phi^F_{\omega_1}(x_k^a, t_k^a))\right] + \frac{1}{M}\sum_{k=1}^{M}(\Phi^F_{\omega_1}(w_k^b, 1) - \Phi^F_{\omega_1}(w_k^a, 0)),$$

$$\mathcal{L}^{ba}(\Phi^G_{\omega_2}) = -\frac{1}{N}\sum_{k=1}^{N}\left[\frac{\partial}{\partial t}\Phi^G_{\omega_2}(x_k^b, t_k^b) + H(\nabla\Phi^G_{\omega_2}(x_k^b, t_k^b))\right] + \frac{1}{M}\sum_{k=1}^{M}(\Phi^G_{\omega_2}(w_k^b, 1) - \Phi^G_{\omega_2}(w_k^a, 0)),$$

$$\mathcal{K}(F_{\theta_1}, G_{\theta_2}) = \frac{\lambda}{K}\sum_{k=1}^{K}|G_{\theta_2}(\xi_k^a + F_{\theta_1}(\xi_k^a)) + F_{\theta_1}(\xi_k^a)|^2 + \frac{\lambda}{K}\sum_{k=1}^{K}|F_{\theta_1}(\xi_k^b + G_{\theta_2}(\xi_k^b)) + G_{\theta_2}(\xi_k^b)|^2,$$

$$\widehat{W}^{ab} = \frac{1}{M}\sum_{k=1}^{M}L(F_{\theta_1}(w_k^a)), \quad \widehat{W}^{ba} = \frac{1}{M}\sum_{k=1}^{M}L(G_{\theta_2}(w_k^b)).$$

B.3.4 Experiment details

In our numerical implementation, for low dimensional cases, i.e., 2 and 10 dimensional cases, we set Φ_F, Φ_G and F, G as fully connected neural networks, where Φ_F, Φ_G have 6 hidden layers and F and G have 5 hidden layers. Each layer has 48 nodes, the activation function is chosen as Tanh. For high dimensional cases, where we deal with MNIST handwritten digits data set, we adopt similar structures of neural networks, the only difference is that in each layer we extend the number of nodes from 48 to 512. In terms of training process, for all synthetic and realistic cases we use the Adam optimizer [149] with learning rate 10^{-4}.

Algorithm 4 Computing Wasserstein geodesic from ρ_a to ρ_b via bidirectional scheme (Equation 3.31) and preconditioning

Choose our preconditioning map $P(x) = \sigma x + \mu$. Denote $\hat{\rho_a} = P_\sharp \rho_a$ (This step is only applicable for 2-Wasserstein case. If we do not need preconditioning, we treat $P = \text{Id}$.)
Set up the threshold $\epsilon > 0$ as the stopping criteria
Initialize $F_{\theta_1}, G_{\theta_2}$, $\Phi^F_{\omega_1}, \Phi^G_{\omega_2}$
for $F_{\theta_1}, G_{\theta_2}$ steps **do**
    Sample $\{(z^a_k, t^a_k)\}_{k=1}^N$ from $\hat{\rho}_a \otimes U(0,1)$ and $\{(z^b_k, t^b_k)\}_{k=1}^N$ from $\rho_b \otimes U(0,1)$;
    Set $x^a_k = z^a_k + t^a_k F_{\theta_1}(z^a_k)$, $x^b_k = z^b_k + t^b_k G_{\theta_2}(z^b_k)$;
    Sample $\{w^a_k\}_{k=1}^M$ from $\hat{\rho}_a$ and $\{w^b_k\}$ from ρ_b;
    for $\Phi^F_{\omega_1}, \Phi^G_{\omega_2}$ steps **do**
        Update (via gradient ascent) $\Phi^F_{\omega_1}, \Phi^G_{\omega_2}$ by:

$$\nabla_{\omega_1,\omega_2}(\mathcal{L}^{ab}(\Phi^F_{\omega_1}) + \mathcal{L}^{ba}(\Phi^G_{\omega_2}))$$

    end for
    Sample $\{\xi^a_k\}_{k=1}^K$ from $\hat{\rho}_a$ and $\{\xi^b_k\}_{k=1}^K$ from ρ_b
    Update (grad descent) $F_{\theta_1}, G_{\theta_2}$ by:

$$\nabla_{\theta_1,\theta_2}(\mathcal{L}^{ab}(\Phi^F_{\omega_1}) + \mathcal{L}^{ba}(\Phi^G_{\omega_2}) + \mathcal{K}(F_{\theta_1}, G_{\theta_2}))$$

    Whenever $|\widehat{W}^{ab} - \widehat{W}^{ba}| < \epsilon$, skip out of the loop.
end for
Set $F_* = F_{\theta_1} \circ P + P - \text{Id}$ and $G^* = G_{\theta_2} \circ P + P - \text{Id}$.
Wasserstein geodesic from ρ_a to ρ_b is given by $\{(\text{Id} + tF_{\theta_1})_\sharp \rho_a\}$; Wasserstein geodesic from ρ_b to ρ_a is given by $\{(\text{Id} + tG_{\theta_2})_\sharp \rho_b\}$.

APPENDIX C

APPENDIX FOR PART 4

C.1 Proof of Lemma 4.3.1

Lemma 3.3. *Suppose $\vec{u}, \vec{v}$ are two vector fields defined on $\mathbb{R}^d$, suppose φ, ψ solves $-\nabla \cdot (\rho \nabla \varphi) = -\nabla \cdot (\rho \vec{u})$ and $-\nabla \cdot (\rho \nabla \psi) = -\nabla \cdot (\rho \vec{v})$, or equivalently, $\text{Proj}_\rho[\vec{u}] = \nabla \varphi$ and $\text{Proj}_\rho[\vec{v}] = \nabla \psi$ (c.f. Definition 4.5.1). Then:*

$$\int \vec{u}(x) \cdot \nabla \psi(x) \rho(x)\, dx = \int \nabla \varphi(x) \cdot \nabla \psi(x) \rho(x)\, dx; \qquad \text{(Equation 4.11)}$$

$$\int |\nabla \psi(x)|^2 \rho(x)\, dx \leq \int |\vec{v}(x)|^2 \rho(x)\, dx. \qquad \text{(Equation 4.12)}$$

Proof of Lemma 4.3.1. For (Equation 4.11):

$$\begin{aligned}\int \vec{u}(x) \cdot \nabla \psi(x) \rho(x)\, dx &= \int -\nabla \cdot (\rho(x) \vec{u}(x)) \psi(x)\, dx = \int -\nabla \cdot (\rho(x) \nabla \varphi(x)) \psi(x)\, dx \\ &= \int \nabla \varphi(x) \cdot \nabla \psi(x) \rho(x)\, dx.\end{aligned}$$

For (Equation 4.12):

$$\begin{aligned}\int |\vec{v}(x)|^2 \rho(x)\, dx &= \int (|\nabla \psi(x)|^2 + 2(\vec{v}(x) - \nabla \psi(x)) \cdot \nabla \psi(x) + |\vec{v}(x) - \nabla \psi(x)|^2) \rho dx \\ &= \int |\nabla \psi(x)|^2 + |\vec{v}(x) - \nabla \psi(x)|^2) \rho(x)\, dx \geq \int |\nabla \psi(x)|^2 \rho(x)\, dx.\end{aligned}$$

The second equality is due to (Equation 4.11). □

C.2 Proof of Theorem 4.3.4

Theorem 3.7. *Suppose $\{\theta_t\}_{t\geq 0}$ solves (Equation 4.28). Then $\{\rho_{\theta_t}\}$ is the gradient flow of $\mathcal{H}$ on probability submanifold $\mathcal{P}_\Theta$. Furthermore, at any time t, $\dot{\rho}_{\theta_t} = \frac{d}{dt}\rho_{\theta_t} \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ is the orthogonal projection of $-grad_W \mathcal{H}(\rho_{\theta_t}) \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}$ onto the subspace $\mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ with respect to the Wasserstein metric g^W.*

Theorem 4.3.4 easily follows from the following two general results about manifold gradient.

Theorem C.2.1. *Suppose $(N, g^N), (M, g^M)$ are Riemannian Manifolds. Suppose $\varphi : N \to M$ is isometry. Consider $\mathcal{F} \in \mathcal{C}^\infty(M)$, define $F = \mathcal{F} \circ \varphi \in \mathcal{C}^\infty(N)$. Suppose $\{x_t\}_{t\geq 0}$ is the gradient flow of F on N:*

$$\dot{x} = -grad_N F(x).$$

Then $\{y_t = \varphi(x_t)\}_{t\geq 0}$ is the gradient flow of $\mathcal{F}$ on M. That is, $\{y_t\}$ satisfies $\dot{y} = -grad_M \mathcal{F}(y)$.

Proof. Since we always have $\dot{y}_t = \varphi_* \dot{x}_t = -\varphi_* \text{grad}_N F(x_t)$, we only need to show that $\varphi_* \text{grad}_N F(x_t) = \text{grad}_M \mathcal{F}(\varphi(x_t))$. Fix the time t, consider any curve $\{\xi_\tau\}$ on N passing through x_t at $\tau = 0$, since φ is isometry, we have $g^N = \varphi^* g^M$, thus:

$$\frac{d}{d\tau} F(\xi_\tau)\Big|_{\tau=0} = g^N(\text{grad}_N F(x_t), \dot{\xi}_0) = \varphi^* g^M(\text{grad}_N F(x_t), \dot{\xi}_0) = g^M(\varphi_* \text{grad}_N F(x_t), \varphi_* \dot{\xi}_0).$$

On the other hand, denote $\eta_\tau = \varphi(\xi_\tau)$, we have:

$$\frac{d}{d\tau} F(\xi_\tau)\Big|_{\tau=0} = \frac{d}{d\tau} \mathcal{F}(\eta_\tau)\Big|_{\tau=0} = g^M(\text{grad}_M \mathcal{F}(y_t), \dot{\eta}_0) = g^M(\text{grad}_M \mathcal{F}(y_t), \varphi_* \dot{\xi}_0).$$

As a result, $g^M(\varphi_* \text{grad}_N F(x_t) - \text{grad}_M \mathcal{F}(y_t), \varphi_* \dot{\xi}_0) = 0$ for all $\dot{\xi}_0 \in T_{x_t} N$. Since φ_* is surjective, we have $\varphi_* \text{grad}_N F(x_t) = \text{grad}_M \mathcal{F}(\varphi(x_t))$.

□

Theorem C.2.2. *Suppose (M, g^M) is Riemannian manifold, $M_{sub} \subset M$ is the submanifold of M. Assume M_{sub} inherits metric g^M, i.e. define $\iota : M_{sub} \to M$ as the inclusion map, which induces a metric tensor on M_{sub} as $g^{M_{sub}} = \iota^* g^M$. For any $\mathcal{F} \in \mathcal{C}^\infty(M)$, denote the restriction of $\mathcal{F}$ on M_{sub} as $\mathcal{F}^{sub}$. Then the gradient $\text{grad}_{M_{sub}} \mathcal{F}^{sub}(x) \in T_x M_{sub}$ is the orthogonal projection of $\text{grad}_M \mathcal{F}(x) \in T_x M$ onto subspace $T_x M_{sub}$ with respect to the metric g^M for any $x \in M_{sub}$.*

Proof. For any $x \in M_{\text{sub}}$, consider any curve $\{\gamma_\tau\}$ on M_{sub} passing through x at $\tau = 0$. We have

$$\begin{aligned}\frac{d}{d\tau}\mathcal{F}^{\text{sub}}(\gamma_\tau)\Big|_{\tau=0} = g^{M_{\text{sub}}}(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \dot{\gamma}_0) &= g^M(\iota_* \text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \iota_* \dot{\gamma}_0) \\ &= g^M(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \dot{\gamma}_0).\end{aligned}$$

The last equality is because ι_* restricted on TM_{sub} is identity. On the other hand, $\mathcal{F}^{\text{sub}}(\gamma_\tau) = \mathcal{F}(\gamma_\tau)$ for all τ. We also have:

$$\frac{d}{d\tau}\mathcal{F}^{\text{sub}}(\gamma_\tau)\Big|_{\tau=0} = g^M(\text{grad}_M \mathcal{F}(x), \dot{\gamma}_0).$$

Combining them we know

$$\begin{aligned} & g^M(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x) - \text{grad}_M \mathcal{F}(x), v) = 0 \quad \forall\, v \in T_x M_{\text{sub}} \\ \Rightarrow \quad & \text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x) - \text{grad}_M \mathcal{F}(x) \perp_{g^M} T_x M_{\text{sub}}, \end{aligned}$$

which proves this result. □

Proof. (Theorem 4.3.4) To prove the first part of Theorem 4.3.4, we apply Theorem C.2.1 with $(N, g^N) = (\Theta, G)$, $M = \mathcal{P}_\Theta$ with its metric inherited from $(\mathcal{P}, g^W)$ and $\varphi = T_{(\cdot)\sharp}$. To prove the second part, we apply Theorem C.2.2 with $(M, g^M) = (\mathcal{P}, g^W)$, $M_{\text{sub}} = \mathcal{P}_\Theta$. □

C.3 Proof of Lemma 4.5.2 Lemma 4.5.3 and Lemma 4.5.4

Lemma 4.6. *Suppose we fix $\theta_0 \in \Theta$, for arbitrary $\theta \in \Theta$ and $\nabla\phi \in L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_{\theta_0})$, we consider*

$$F(\theta, \nabla\phi \mid \theta_0) = \left(\int (2\nabla\phi(x) \cdot (T_\theta - T_{\theta_0}) \circ T_{\theta_0}^{-1}(x) - |\nabla\phi(x)|^2)\, \rho_{\theta_0}(x)\, dx \right) + 2hH(\theta). \tag{Equation 4.54}$$

Then $F(\theta, \nabla\phi \mid \theta_0) < \infty$, furthermore, $F(\cdot, \nabla\phi \mid \theta_0) \in C^1(\Theta)$. We can compute

$$\partial_\theta F(\theta, \nabla\phi \mid \theta_0) = 2 \left(\int \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T\, \nabla\phi(x)\, \rho_{\theta_0}(x)\, dx + h\, \nabla_\theta H(\theta) \right). \tag{Equation 4.55}$$

Proof. To show $F(\theta, \nabla\phi\,|\theta_0) < \infty$, we write

$$F(\theta, \nabla\,|\theta_0) = \underbrace{\int 2\nabla\phi \cdot T_\theta(T_{\theta_0}^{-1}(x))\rho_{\theta_0}\, dx}_{A} - \underbrace{\int 2\nabla\phi(T_{\theta_0}(x)) \cdot x dp(x)}_{B} - \underbrace{\int |\nabla\phi(x)|^2 \rho_{\theta_0}(x)\, dx}_{C} + 2hH(\theta).$$

By Cauchy–Schwarz inequality, the first two terms can be estimated as

$$|A - B| \le 2\|\nabla\phi\|_{L^2(\rho_{\theta_0})} \left(\int |T_\theta(x)|^2 dp(x) + \int x^2 dp(x) \right).$$

Recall (Equation 4.9) and p having finite second order moment, we know the first two terms are finite. In addition $C = \|\nabla\phi\|^2_{L^2(\rho_{\theta_0})} < \infty$. We thus have shown $F(\theta, \nabla\phi\,|\theta_0) < \infty$.

To show $F(\cdot, \nabla\phi\,|\theta_0) \in C^1(\Theta)$, recall $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$ as mentioned in subsection 4.3.1, we know the relative entropy $H(\cdot) \in C^1(\Theta)$, thus we only need to prove for $\tilde{F}(\cdot, \nabla\phi\,|\theta_0) = F(\cdot, \nabla\phi\,|\theta_0) - 2hH(\theta)$. We consider $\xi \in \mathbb{R}^m$ with $|\xi|$ small enough and

$\theta + \xi \in \Theta$. Then the difference

$$\tilde{F}(\theta+\xi, \nabla\phi\,|\theta_0) - \tilde{F}(\theta, \nabla\phi\,|\theta_0) = \int 2\nabla\phi(x)\cdot(T_{\theta+\xi} - T_\theta)\circ T_{\theta_0}^{-1}(x)\;\rho_{\theta_0}(x)\;dx \quad \text{(C.1)}$$

We denote the i-th component of T_θ as $T_\theta^{(i)}$, $1 \le i \le d$. By Taylor expansion (w.r.t. θ), we have $T_{\theta+\xi}^{(i)}(x) - T_\theta^{(i)}(x) = \partial_\theta T_\theta^{(i)}(x)^{\mathrm{T}}\xi + \frac{1}{2}\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}(x)\xi$ with $\lambda_i(x) \in [0,1]$, then the right hand side of (Equation C.1) is

$$\underbrace{\left(\int 2\partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^{\mathrm{T}}\nabla\phi(x)\rho_{\theta_0}\;dx\right)^{\mathrm{T}}}_{\text{Denote as } \mathcal{J}(\theta)^{\mathrm{T}}\xi}\xi + \int\left(\sum_{i=1}^d \partial_{x_i}\phi\cdot(\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\xi)\right)\rho_{\theta_0}\;dx \quad \text{(C.2)}$$

By Cauchy-Schwarz inequality, the sum in the second term of (Equation C.2) can be estimated as

$$\left(\sum_{i=1}^d |\partial_{x_i}\phi|^2\right)^{\frac{1}{2}}\cdot\left(\sum_{i=1}^d |\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\xi|^2\right)^{\frac{1}{2}} \le |\nabla\phi|\cdot\left(\sum_{i=1}^d \|\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\|_2^2\right)^{\frac{1}{2}}|\xi|^2$$

Let us recall (Equation 4.53) and the absolute value of the second term in (Equation C.2) can be upper bounded by

$$\left(\int|\nabla\phi|^2\rho_{\theta_0}\;dx\right)^{\frac{1}{2}}\cdot\left(\int\sum_{i=1}^d\|\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(x)\|_2^2 dp(x)\right)^{\frac{1}{2}}|\xi|^2 \le \|\nabla\phi\|_{L^2(\rho_{\theta_0})}^2\cdot\sqrt{H(\theta_0,|\xi|)}|\xi|^2.$$

This inequality is due to (Equation 4.53). As a result, we have

$$\frac{|\tilde{F}(\theta+\xi,\nabla\phi\,|\theta_0) - \tilde{F}(\theta,\nabla\,|\theta_0) - \mathcal{J}(\theta)^{\mathrm{T}}\xi|}{|\xi|} \le \|\nabla\phi\|_{L^2(\rho_{\theta_0})}^2\cdot\sqrt{H(\theta_0,|\xi|)}\;|\xi|. \quad \text{(C.3)}$$

Since $H(\theta_0,\epsilon)$ is increasing w.r.t. ϵ, send $|\xi| \to 0$, the upper bound in (Equation C.3) approaches to 0. This verifies the differentiability of $\tilde{F}(\cdot,\nabla\phi\,|\theta_0)$.Thus $F(\cdot,\nabla\phi\,|\theta_0)$ is also differentiable and $\partial_\theta F(\theta,\nabla\phi\,|\theta_0) = \mathcal{J}(\theta) + 2h\nabla_\theta H(\theta)$. At last, to show that $F(\cdot,\nabla\phi\,|\theta_0) \in$

$C^1(\Theta)$, we only need to prove the continuity of $\mathcal{J}(\theta)$. One only need to notice that

$$\begin{aligned} 2\partial_\theta T_{\theta'}^{(i)}(T_{\theta_0}^{-1}(x))^{\mathrm{T}}\nabla\phi(x) &\le |\partial_{\theta'} T_\theta^{(i)}(T_{\theta_0}^{-1}(x))|^2 + |\nabla\phi(x)|^2 \\ &\le L_2(T_{\theta_0}^{-1}(x)|\theta) + |\nabla\phi(x)|^2 \quad \forall\, \theta',\ |\theta'-\theta| < r(\theta). \end{aligned}$$

The last inequality is due to condition (Equation 4.10). Since $L_2(T_{\theta_0}^{-1}(x)|\theta) + |\nabla\phi(x)|^2 \in L^1(\rho_{\theta_0})$, then by Dominated Convergence Theorem, we are able to prove the continuity of $\partial_\theta F(\theta, \nabla\phi\,|\theta_0)$. □

Lemma 4.7. *Suppose we fix $\theta_0 \in \Theta$ and define $J(\theta) = \sup_{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_0})} F(\theta, \nabla\phi \mid \theta_0)$. Then J is differentiable. If we denote $\hat{\psi}_\theta = \operatorname{argmax}_\phi \{F(\theta, \nabla\phi \mid \theta_0)\}$, then*

$$\nabla_\theta J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) = 2\left(\int \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T\, \nabla\hat{\psi}_\theta(x)\, \rho_{\theta_0}(x)\, dx + h\, \nabla_\theta H(\theta)\right).$$

Proof. Let us denote $\Xi_\theta = (T_\theta - T_{\theta_0}) \circ T_{\theta_0}^{-1}$. Then for any $\xi \in \mathbb{R}^m$ such that $\theta + \xi \in \Theta$, we set $\hat{\psi}_{\theta+\xi} = \operatorname{argmax}_\phi\{F(\theta+\xi, \nabla\phi \mid \theta_0)\}$. Then according to Definition 4.5.1, $\hat{\psi}_\theta, \hat{\psi}_{\theta+\xi}$ solves

$$-\nabla\cdot(\rho_{\theta_0}\nabla\hat{\psi}_\theta) = -\nabla\cdot(\rho_{\theta_0}\Xi_\theta) \quad -\nabla\cdot(\rho_{\theta_0}\nabla\hat{\psi}_{\theta+\xi}) = -\nabla\cdot(\rho_{\theta_0}\Xi_{\theta+\xi}). \tag{C.4}$$

Subtracting the two equations, then multiply $\hat{\psi}_{\theta+\xi} - \hat{\psi}_\theta$ on both sides and integrate yields

$$\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2 \rho_{\theta_0}\, dx = \int (\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta)\cdot(\Xi_{\theta+\xi} - \Xi_\theta)\rho_{\theta_0}\, dx.$$

Then by Cauchy–Schwarz inequality, we derive

$$\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2 \rho_{\theta_0}\, dx \le \int |\Xi_{\theta+\xi} - \Xi_\theta|^2 \rho_{\theta_0}\, dx.$$

Now since $\Xi_{\theta_\xi}(x) - \Xi_\theta(x) = (T_{\theta+\xi} - T_\theta) \circ T_{\theta_0}^{-1}(x)$, by mean value theorem, the i-th

component of $\Xi_{\theta+\xi}(x)-\Xi_\theta(x)$ can be written as $\partial_\theta T^{(i)}_{\theta+\lambda_i(x)\xi}(T^{-1}_{\theta_0}(x))^{\mathrm{T}}\xi$ with $\lambda_i(x)\in[0,1]$. Then recall the definition of $L(\theta,\epsilon)$ in (Equation 4.53), we can verify

$$\int |\Xi_{\theta+\xi}-\Xi_\theta|^2\rho_{\theta_0}\,dx=\int |T_{\theta+\xi}(x)-T_\theta(x)|dp(x)\le L(\theta,|\xi|)|\xi|^2.$$

Thus we have the following estimation

$$\int |\nabla\hat{\psi}_{\theta+\xi}-\nabla\hat{\psi}_\theta|^2\rho_{\theta_0}\,dx\le L(\theta,|\xi|)|\xi|^2 \tag{C.5}$$

Now let us consider $J(\theta+\xi)-J(\theta)$

$$\begin{aligned}
&J(\theta+\xi)-J(\theta)\\
=&F(\theta+\xi,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)-F(\theta,\nabla\hat{\psi}_\theta\mid\theta_0)\\
=&\underbrace{F(\theta+\xi,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)-F(\theta,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)}_{A}+\underbrace{F(\theta,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)-F(\theta,\nabla\hat{\psi}_\theta\mid\theta_0)}_{B}.
\end{aligned}\tag{C.6}$$

Now according to Lemma 4.5.2, $F(\cdot,\nabla\phi\mid\theta_k)\in C^1(\Theta)$. By mean value theorem, term A can be written as

$$\begin{aligned}
A=&F(\theta+\xi,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)-F(\theta,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)=\partial_\theta F(\theta+\tau\xi,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)\xi\quad\text{with }\tau\in[0,1]\\
=&\partial_\theta F(\theta,\nabla\hat{\psi}_\theta\mid\theta_0)^{\mathrm{T}}\xi+(\underbrace{\partial_\theta F(\theta+\tau\xi,\nabla\hat{\psi}_\theta\mid\theta_0)-\partial_\theta F(\theta,\nabla\hat{\psi}_\theta\mid\theta_0)}_{r_1(\theta,\xi)})^{\mathrm{T}}\xi\\
&+(\underbrace{\partial_\theta F(\theta+\tau\xi,\nabla\hat{\psi}_{\theta+\xi}\mid\theta_0)-\partial_\theta F(\theta+\tau\xi,\nabla\hat{\psi}_\theta\mid\theta_0)}_{r_2(\theta,\xi)})^{\mathrm{T}}\xi.
\end{aligned}$$

Term B can be computed as

$$\begin{aligned}
B =& F(\theta, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_{\theta} \mid \theta_0) \\
=& \int (2(\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_{\theta}) \cdot \Xi_\theta - (|\nabla\hat{\psi}_{\theta+\xi}|^2 - |\nabla\hat{\psi}_{\theta}|^2))\rho_{\theta_0}\, dx \\
=& 2\int (\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_{\theta}) \cdot (\Xi_\theta - \nabla\hat{\psi}_{\theta})\rho_{\theta_0}\, dx - \int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_{\theta}|^2\rho_{\theta_0}\, dx \\
=& -\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_{\theta}|^2\rho_{\theta_0}\, dx.
\end{aligned}$$

The last equality is due to integration by parts and (Equation C.4).

Now substituting A and B in (Equation C.6) yields

$$J(\theta+\xi) - J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) + r_1(\theta,\xi)^{\mathrm{T}}\xi + r_2(\theta,\xi)^{\mathrm{T}}\xi - \|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})}$$

We can estimate

$$\frac{\left|J(\theta+\xi) - J(\theta) - \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)^{\mathrm{T}}\xi\right|}{|\xi|} \leq |r_1(\theta,\xi)| + |r_2(\theta,\xi)| + \frac{\|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})}}{|\xi|} \tag{C.7}$$

Now we prove the right hand side of (Equation C.7) approaches to 0 as $\xi \to 0$. Since $\partial_\theta F(\cdot, \nabla\hat{\psi}_\theta \mid \theta_0) \in C^1(\Theta)$, using continuity, we know $\lim_{\xi\to 0} r_1(\theta,\xi) = 0$. For $r_2(\theta,\xi)$, when $|\xi|$ is sufficiently small, we have

$$\begin{aligned}
|r_2(\theta,\xi)| =& \left|\int \partial_\theta T_{\theta+\tau\xi}(T_{\theta_0}^{-1}(x))^{\mathrm{T}}(\nabla\hat{\psi}_{\theta+\xi}(x) - \nabla\hat{\psi}_\theta(x))\rho_{\theta_0}(x)\, dx\right| \\
\leq& \left(\int \|\partial_\theta T_{\theta+\tau\xi}(x)\|_F^2 dp(x)\right)^{\frac{1}{2}} \left(\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2\rho_{\theta_0}\, dx\right)^{\frac{1}{2}} \\
\leq& \sqrt{\|L_2(\cdot|\theta)\|_{L^1(p)}}\sqrt{L(\theta,|\xi|)}|\xi|
\end{aligned}$$

The last inequality is due to (Equation 4.10) (when $|\xi|$ is small enough so that $|\xi| < r(\theta)$) and (Equation C.5). Using this we are able to show $\lim_{\xi\to 0} r_2(\theta,\xi) = 0$. Using (Equation C.5) again, we can verify $\frac{1}{|\xi|}\|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})} \leq L(\theta,|\xi|)|\xi| \to 0$ as $\xi \to 0$.

Thus J is differentiable at θ and we know $\nabla_\theta J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)$. We complete the proof by applying (Equation 4.55) of Lemma (Lemma 4.5.2).

□

Lemma 4.8. *Under assumption*(Equation 4.51)*, the optimal solution of* (Equation 4.48) θ_{k+1} *satisfies,*

$$|\theta_{k+1} - \theta_k| \sim o(1) \quad \textit{i.e.} \lim_{h\to 0^+} |\theta_{k+1} - \theta_k| = 0.$$

Proof of Lemma 4.5.4. Recall the function to be minimized in (Equation 4.48) is $J(\theta) = \widehat{W}_2^2(\theta, \theta_k) + 2hH(\theta)$. If choosing $\theta = \theta_k$ in (Equation 4.48), we have $J(\theta_k) = 2hH(\theta_k)$. Thus $J(\theta_{k+1}) \leq J(\theta_k) = 2hH(\theta_k)$. Since $H(\theta_k) \geq 0$, this will lead to $\widehat{W}_2^2(\theta_{k+1}, \theta_k) \leq 2hH(\theta_k)$. When h is small enough, $|\theta_{k+1} - \theta_k| \leq l^{-1}(2hH(\theta_k))$, here l^{-1} is the inverse function of l defined on $[0, l(r_0)]$. We know $l^{-1}(0) = 0$ and l^{-1} is also continuous and increasing function. This leads to $\lim_{h\to 0^+} |\theta_{k+1} - \theta_k| \leq \lim_{h\to 0^+} l^{-1}(2hH(\theta_k)) = 0$. □

C.4 Proofs for Lemma 4.6.6 and Lemma 4.6.7

Lemma 5.7. *The geodesic connecting $\rho_0, \rho_1 \in \mathcal{P}(M)$ is described by,*

$$\begin{cases} \frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t \nabla \psi_t) = 0 \\ \frac{\partial \psi_t}{\partial t} + \frac{1}{2}|\nabla \psi_t|^2 = 0 \end{cases} \qquad \rho_t|_{t=0} = \rho_0,\ \rho_t|_{t=1} = \rho_1. \qquad \text{(Equation 4.85)}$$

Using the notation $\dot{\rho}_t = \partial_t \rho_t = -\nabla \cdot (\rho_t \nabla \psi_t) \in \mathcal{T}_{\rho_t}\mathcal{P}(M)$, $g^W(\dot{\rho}_t, \dot{\rho}_t)$ is constant for $0 \leq t \leq 1$ and $g^W(\dot{\rho}_t, \dot{\rho}_t) = W_2^2(\rho_0, \rho_1)$ for $0 \leq t \leq 1$.

Proof. Recall the definition (Equation 2.58) of Wasserstein metric g^W, we have $g^W(\dot{\rho}_t, \dot{\rho}_t) = \int |\nabla \psi_t|^2 \rho_t \, dx$. Since $\{\rho_t\}$ is the geodesic on $(\mathcal{P}(M), g^W)$, the speed $g^W(\sigma_t, \sigma_t)$ remains constant. To directly verify this, we compute the time derivative:

$$\frac{d}{dt} g^W(\dot{\rho}_t, \dot{\rho}_t) = \frac{d}{dt}\left(\int |\nabla \psi_t|^2 \rho_t \, dx \right) = \int \frac{\partial}{\partial t} |\nabla \psi_t|^2 \rho_t \, dx + \int |\nabla \psi_t|^2 \partial_t \rho_t \, dx.$$

Using the first equation in (Equation 4.85), we obtain

$$\int |\nabla\psi_t|^2 \partial_t \rho_t \, dx = \int |\nabla\psi_t|^2 \cdot (-\nabla \cdot (\rho_t \nabla\psi_t)) \, dx = \int \nabla(|\nabla\psi_t|^2) \cdot \nabla\psi_t \rho_t \, dx,$$

Taking the spatial gradient of the second equation in (Equation 4.85), we have

$$\partial_t(\nabla\psi_t) = -\nabla(\frac{1}{2}|\nabla\psi_t|^2).$$

Then

$$\int \frac{\partial}{\partial t}|\nabla\psi_t|^2 \rho_t \, dx = \int 2\partial_t(\nabla\psi_t) \cdot \nabla\psi_t \rho_t \, dx = \int -\nabla(|\nabla\psi_t|^2) \cdot \nabla\psi_t \rho_t \, dx.$$

Adding them together, we verify $\frac{d}{dt} g^W(\dot\rho_t, \dot\rho_t) = 0$, hence $\int_0^1 g^W(\dot\rho_t, \dot\rho_t) \, dt = W_2^2(\rho_0, \rho_1)$. Thus we know $g^W(\dot\rho_t, \dot\rho_t) = W_2^2(\rho_0, \rho_1)$ for any $0 \le t \le 1$. □

Lemma 5.8. *Suppose* $\{\rho_t\}$ *solves* (Equation 4.85)*, the relative entropy* $\mathcal{H}$ *in* (Equation 4.8) *has potential* V *satisfying* $\nabla^2 V \succeq \lambda I$*, then we have* $\frac{d}{dt} g^W(\mathit{grad}_W \mathcal{H}(\rho_t), \dot\rho_t) \ge \lambda W_2^2(\rho_0, \rho_1)$*. Or equivalently,* $\frac{d^2}{dt^2} \mathcal{H}(\rho_t) \ge \lambda W_2^2(\rho_0, \rho_1)$.

Proof. Let us write:

$$g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot\rho_t) = \int \nabla(V + D\log\rho_t) \cdot \nabla\psi_t \, \rho_t \, dx.$$

Then:

$$\begin{aligned}\frac{d}{dt} g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot\rho_t) &= \frac{d}{dt}\left(\int \nabla(V + D\log\rho_t) \cdot \nabla\psi_t \, \rho_t \, dx\right) \\ &= \int (\nabla\psi_t^{\mathrm{T}} \nabla^2 V \nabla\psi_t + \mathrm{Tr}(\nabla^2\psi_t \nabla^2\psi_t)) \, \rho_t \, dx.\end{aligned}$$

The second equality can be carried out by direct calculations. One can check [29] or [7] for

its complete derivation. Using $\nabla^2 V \succeq \lambda I$, we get

$$\frac{d}{dt} g^W(\text{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) \geq \int \lambda |\nabla \psi_t|^2 \rho_t \, dx = \lambda \, g^W(\dot{\rho}_t, \dot{\rho}_t) = \lambda W_2^2(\rho_0, \rho_1).$$

The last equality is due to Lemma 4.6.6. By the definition of Wasserstein gradient stated in (Equation 2.61), we have $\frac{d}{dt}\mathcal{H}(\rho_t) = g^W(\text{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t)$, we also proved $\frac{d^2}{dt^2}\mathcal{H}(\rho_t) \geq \lambda W_2^2(\rho_0, \rho_1)$. □

APPENDIX D

APPENDIX FOR PART 5

D.1 The background of Schrödinger Bridge Problem (SBP)

Denote $\Omega = C([0,1], \mathbb{R}^d)$. Given $R \in \mathcal{M}^+(\Omega)$ the law of the reversible Brownian motion (here we consider the Brownian motion with the volume Lebesgue measure, denoted by Leb, as the initial distribution). Consider the relative entropy of any probability measure with respect to R,

$$\mathcal{H}(P|R) = \int_\Omega \log(\frac{dP}{dR})dP.$$

The SBP can be formulated as

$$\min \mathcal{H}(P|R), P \in \mathcal{P}(\Omega) : P_0 = \mu_0, P_1 = \mu_1. \tag{D.1}$$

Here $P_0 := P(X_0 \in \cdot)$, $P_1 := P(X_1 \in \cdot)$ and $X_t(\omega) := \omega(t)$ is the canonical process with $\omega \in \Omega$. It is proven (see e.g. [174]) that if $\mathcal{H}(\widetilde{\mu}_0|Leb) < \infty$ and $\mathcal{H}(\widetilde{\mu}_1|Leb) < \infty$, the SBP has a unique solution $\widehat{P}$ which enjoys the following decomposition

$$\widehat{P} = f_0(X_0)g_1(X_1)R \in \mathcal{P}(\Omega),$$

where f_0, g_1 are nonnegative measurable functions such that

$$\mathbb{E}_R[f_0(X_0)g_1(X_1)] = 1.$$

Introduce the function f_t, g_t defined by

$$f_t(z) := \mathbb{E}_R[f_0(X_0)|X_t = z],$$
$$g_t(z) := \mathbb{E}_R[g_1(X_1)|X_t = z],\ P_t\text{-}a.e.,\ z \in \mathbb{R}^d,$$

and the constraint

$$\widetilde{\mu}_0 = f_0 g_0 Leb,\ \widetilde{\mu}_1 = f_1 g_1 Leb.$$

Then the SBP (Equation 5.2) with $\hbar = 1$ is equivalent to the following minimal action problem, i.e.,

$$\begin{aligned}
&\inf\{\mathcal{H}(P|R) : P_0 = \widetilde{\mu}_0, P_1 = \widetilde{\mu}_1\} - \mathcal{H}(\mu_0|Leb) \qquad \text{(D.2)}\\
&= \inf\Big\{ \int_0^1 \int_{\mathbb{R}^d} \frac{|v_t|^2}{2} \mu_t\, dxdt : (\partial_t - \frac{\Delta}{2})\mu + \nabla\cdot(v\mu) = 0,\\
&\quad P_0 = \mu_0, P_1 = \mu_1 \Big\}
\end{aligned}$$

We denote ρ_t the density of μ_t with respect to the Lebesgue measure. In addition, with the assumption that μ_0, μ_1 have finite second moments, the critical point of the minimal action problem satisfies the following system

$$(\partial_t - \frac{\Delta}{2})\rho + \nabla\cdot(\nabla\phi\rho) = 0,\ \rho(0) = \rho_0,$$
$$(\partial_t + \frac{\Delta}{2})\phi + \frac{1}{2}|\nabla\phi|^2 = 0,\ \phi(1) = \log(g_1)$$

with $v_t = \nabla\phi_t$. There is also a backward version of this PDE system, namely

$$(-\partial_t - \frac{\Delta}{2})\rho + \nabla\cdot(\nabla\psi\rho) = 0,\ \rho(1) = \rho_1,$$
$$(-\partial_t + \frac{\Delta}{2})\psi + \frac{1}{2}|\nabla\psi|^2 = 0,\ \psi(0) = \log(f_0).$$

Here we have the relation $\nabla\psi_t + \nabla\phi_t = \nabla\log(\rho_t)$.

Applying the transformation

$$S_t = \phi_t - \frac{1}{2}\log(\rho_t) \tag{D.3}$$

as being done in [169], we arrive at the Hamiltonian system on the density space,

$$\begin{aligned}\frac{\partial}{\partial t}\rho + \nabla\cdot(\rho(t,x)\nabla S) = 0,\\ \frac{\partial}{\partial t}S + \frac{1}{2}|\nabla S|^2 - \frac{1}{8}\frac{\delta}{\delta\rho_t}I(\rho_t) = 0.\end{aligned}$$

The corresponding Hamiltonian is $\mathscr{H}(\rho, S) = \frac{1}{2}\int_{\mathbb{R}^d}|\nabla S|^2\rho dx - \frac{1}{8}I(\rho)$ where $I(\rho) = \int_{\mathbb{R}^d}|\nabla\log(\rho)|^2\rho dx$ is the Fisher information. Meanwhile, the action minimizing problem (Equation D.2) can be rewritten as

$$\begin{aligned}\inf_{v_t}\Big\{\int_0^1 \mathbb{E}[\frac{1}{2}v(t,X(t))^2] + \frac{1}{8}I(\rho(t))dt + \frac{1}{2}\int(\rho^1\log(\rho^1) - \rho^0\log(\rho^0))dx \\ |\ dX_t = v(t,X_t)dt,\ X(0)\sim\rho^0,\ X(1)\sim\rho^1\Big\}.\end{aligned} \tag{D.4}$$

Here $\rho(t)$ is the density of the marginal distribution of X_t.

Next, by introducing the conjugate Madelung transformation $f = \sqrt{\rho}e^S, g = \sqrt{\rho}e^{-S}$ (also known as "Hopf-Cole" transformation), f and g satisfy so-called "Schrödinger system" (see e.g. [179, 186, 187]),

$$\begin{aligned}(\partial_t - \frac{\Delta}{2})g = 0,\ g(0) = g_0,\\ (\partial_t + \frac{\Delta}{2})f = 0,\ f(1) = f_1.\end{aligned} \tag{D.5}$$

This also implies the following relationships

$$\phi = \log(f) = S + \frac{1}{2}\log(\rho), \psi = \log(g) = -S + \frac{1}{2}\log(\rho).$$

REFERENCES

[1] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[2] L. Kantorovich, "On translation of mass (in russian), c r," in *Doklady. Acad. Sci. USSR*, vol. 37, 1942, pp. 199–201.

[3] J. D. Lafferty, "The Density Manifold and Configuration Space Quantization," *Transactions of the American Mathematical Society*, vol. 305, no. 2, pp. 699–741, 1988.

[4] F. Otto, "The Geometry of Dissipative Evolution Equations: The Porous Medium Equation," *Communications in Partial Differential Equations*, vol. 26, no. 1-2, pp. 101–174, 2001.

[5] 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.

[6] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the monge-kantorovich mass transfer problem," *Numerische Mathematik*, vol. 84, no. 3, pp. 375–393, 2000.

[7] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," in *arXiv preprint arXiv:1701.07875*, 2017.

[9] A. Galichon, *Optimal transport methods in economics*. Princeton University Press, 2016.

[10] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[11] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt, "Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion," *Geophysics*, vol. 83, no. 1, R43–R62, 2018.

[12] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in neural information processing systems*, 2013, pp. 2292–2300.

[13] V. Seguy, B. Damodaran, R. Flamary, R. Courty N., A., and M. Blondel, "Large-scale optimal transport and mapping estimation," *arXiv preprint arXiv:1711.02283*, 2017.

[14] A. Tong Lin, W. Li, S. Osher, and G. Montúfar, "Wasserstein proximal of gans," 2018.

[15] Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha, "On scalable and efficient computation of large scale optimal transport," in *International Conference on Machine Learning*, 2019, pp. 6882–6892.

[16] G. Lu, Z. Zhou, J. Shen, C. Chen, W. Zhang, and Y. Yu, "Large-scale optimal transport via adversarial training with cycle-consistency," in *arXiv preprint arXiv:2003.06635*, 2019.

[17] L. Ruthotto, S. J. Osher, W. Li, L. Nurbekyan, and S. W. Fung, "A machine learning framework for solving high-dimensional mean field game and mean field control problems," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9183–9193, 2020.

[18] A. Makkuva A.and Taghvaei, S. Oh, and J. Lee, "Optimal transport mapping via input convex neural networks," in *International Conference on Machine Learning*, 2020, pp. 6672–6681.

[19] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev, "Wasserstein-2 generative networks," *arXiv preprint arXiv:1909.13082*, 2019.

[20] J. Fan, S. Liu, S. Ma, Y. Chen, and H. Zhou, "Scalable computation of monge maps with general costs," *arXiv preprint arXiv:2106.03812*, 2021.

[21] S. Liu, H. Sun, and H. Zha, "A particle-evolving method for approximating the optimal transport plan," in *Geometric Science of Information*, F. Nielsen and F. Barbaresco, Eds., Cham: Springer International Publishing, 2021, pp. 878–887, ISBN: 978-3-030-80209-7.

[22] ——, *Approximating the optimal transport plan via particle-evolving method*, 2021. arXiv: 2105.06088 [math.OC].

[23] S. Liu, S. Ma, Y. Chen, H. Zha, and H. Zhou, "Learning high dimensional wasserstein geodesics," *arXiv preprint arXiv:2102.02992*, 2021.

[24] W. Li, S. Liu, H. Zha, and H. Zhou, "Parametric fokker-planck equation," in *Geometric Science of Information*, F. Nielsen and F. Barbaresco, Eds., Cham: Springer International Publishing, 2019, pp. 715–724, ISBN: 978-3-030-26980-7.

[25] S. Liu, W. Li, H. Zha, and H. Zhou, "Neural parametric fokker-planck equations," *arXiv preprint arXiv:2002.11309, accepted by SIAM Journal on Numerical Analysis*, 2020.

[26] J. Cui, S. Liu, and H. Zhou, "What is a stochastic hamiltonian process on finite graph? an optimal transport answer," *Journal of Differential Equations*, vol. 305, pp. 428–457, 2021.

[27] S. Chow, W. Huang, Y. Li, and H. Zhou, "Fokker-Planck equations for a free energy functional or Markov process on a graph," *Arch. Ration. Mech. Anal.*, vol. 203, no. 3, pp. 969–1008, 2012.

[28] J. Maas, "Gradient flows of the entropy for finite markov chains," *Journal of Functional Analysis*, vol. 261, no. 8, pp. 2250–2292, 2011.

[29] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2003.

[30] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[31] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, Calif.: University of California Press, 1951, pp. 481–492.

[32] L. A. Caffarelli, "The monge-ampere equation and optimal transportation, an elementary review," in *Optimal transportation and applications*, Springer, 2003, pp. 1–10.

[33] L. A. Caffarelli and R. J. McCann, "Free boundaries in optimal transport and monge-ampere obstacle problems," *Annals of mathematics*, pp. 673–730, 2010.

[34] L. C. Evans, *Partial differential equations*. American Mathematical Soc., 2010, vol. 19.

[35] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[36] A. D. Fokker, "Die mittlere energie rotierender elektrischer dipole im strahlungsfeld," *Annalen der Physik*, vol. 348, no. 5, pp. 810–820, 1914.

[37] H. Risken, *The Fokker-Planck Equation*, ser. Springer Series in Synergetics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989, vol. 18.

[38] J. L. Vázquez, *The porous medium equation: mathematical theory*. Oxford University Press on Demand, 2007.

[39] J. Fan, S. Liu, S. Ma, Y. Chen, and H.-M. Zhou, "Scalable computation of monge maps with general costs," in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[40] S. Liu, S. Ma, Y. Chen, H. Zha, and H. Zhou, "Learning high dimensional wasserstein geodesics," in *arXiv preprint arXiv:2102.02992*, 2021.

[41] J. Fan, A. Taghvaei, and Y. Chen, "Scalable computations of wasserstein barycenter via input convex neural networks," *arXiv preprint arXiv:2007.04462*, 2020.

[42] B. Amos, L. Xu, and J. Kolter, "Input convex neural networks," in *International Conference on Machine Learning*, 2017, pp. 146–155.

[43] J.-D. Benamou, B. D. Froese, and A. M. Oberman, "Two numerical methods for the elliptic monge-ampere equation," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 44, no. 4, pp. 737–758, 2010.

[44] B. D. Froese, "A numerical method for the elliptic monge–ampère equation with transport boundary conditions," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, A1432–A1459, 2012.

[45] R. Glowinski, H. Liu, S. Leung, and J. Qian, "A finite element/operator-splitting method for the numerical solution of the two dimensional elliptic monge–ampère equation," *Journal of Scientific Computing*, vol. 79, no. 1, pp. 1–47, 2019.

[46] H. Liu, R. Glowinski, S. Leung, and J. Qian, "A finite element/operator-splitting method for the numerical solution of the three dimensional monge–ampère equation," *Journal of Scientific Computing*, vol. 81, no. 3, pp. 2271–2302, 2019.

[47] A. M. Oberman and Y. Ruan, "An efficient linear programming method for optimal transportation," *arXiv preprint arXiv:1509.03668*, 2015.

[48] J. D. Walsh III and L. Dieci, "General auction method for real-valued optimal transport," *arXiv preprint arXiv:1705.06379*, 2017.

[49] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013.

[50] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via sinkhorn iteration," *Advances in neural information processing systems*, pp. 1964–1974, 2017.

[51] A. Genevay, G. Peyré, and M. Cuturi, "Learning generative models with sinkhorn divergences," *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

[52] R. Li, X. Ye, H. Zhou, and H. Zha, "Learning to match via inverse optimal transport," in *J. Mach. Learn. Res.*, 2019, 20, pp.80–1.

[53] Y. Xie, X. Wang, R. Wang, and H. Zha, "A fast proximal point method for computing exact wasserstein distance," in *Uncertainty in Artificial Intelligence*, 2020, pp. 433–453.

[54] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Neural Information Processing Systems*, 2017.

[55] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[56] R. Flamary, N. Courty, A. Rakotomamonjy, and D. Tuia, "Optimal transport with laplacian regularization," in *Advances in Neural Information Processing Systems, Workshop on Optimal Transport and Machine Learning*, 2014.

[57] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1853–1865, 2016.

[58] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen, "Tsallis regularized optimal transport and ecological inference," *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[59] A. Dessein, N. Papadakis, and J. Rouas, "Regularized optimal transport and the rot mover's distance," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 590–642, 2018.

[60] Y. Mroueh, C. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev gan," in *arXiv preprint arXiv:1711.04894*, 2020.

[61] J. Benamou and Y. Brenier, "A computational fluid mechanics solution to the monge-kantorovich mass transfer problem," *Numerische Mathematik*, vol. 84, no. 3, pp. 375–393, 2000.

[62] W. Li, E. Ryu, S. Osher, W. Yin, and W. Gangbo, "A parallel method for earth mover's distance," *Journal of Scientific Computing*, vol. 75, no. 1, pp. 182–197, 2018.

[63] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "An interpolating distance between optimal transport and fisher–rao metrics," *Foundations of Computational Mathematics*, vol. 18, no. 1, pp. 1–44, 2018.

[64] W. Gangbo, W. Li, S. Osher, and M. Puthawala, "Unnormalized optimal transport," *Journal of Computational Physics*, vol. 399, p. 108 940, 2019.

[65] V. Krishnan and S. Martínez, "Distributed optimal transport for the deployment of swarms," *IEEE Conference on Decision and Control (CDC)*, pp. 4583–4588, 2018.

[66] D. Inoue, Y. Ito, and H. Yoshida, "Optimal transport-based coverage control for swarm robot systems: Generalization of the voronoi tessellation-based method," *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1483–1488, 2020.

[67] S. Ma, S. Liu, H. Zha, and H. Zhou, "Learning stochastic behaviour from aggregate data," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 7258–7267.

[68] I. Haasler, R. Singh, Q. Zhang, J. Karlsson, and Y. Chen, "Multi-marginal optimal transport and probabilistic graphical models," in *arXiv preprint arXiv:2006.14113*, 2020.

[69] A. Korotin, D. Selikhanovych, and E. Burnaev, "Neural optimal transport," *arXiv preprint arXiv:2201.12220*, 2022.

[70] L. Rout, A. Korotin, and E. Burnaev, "Generative modeling with optimal transport maps," in *International Conference on Learning Representations*, 2022.

[71] M. Gazdieva, L. Rout, A. Korotin, A. Filippov, and E. Burnaev, "Unpaired image super-resolution with optimal transport maps," *arXiv preprint arXiv:2202.01116*, 2022.

[72] J.-C. Hütter and P. Rigollet, *Minimax estimation of smooth optimal transport maps*, 2020. arXiv: 1905.05828 [math.ST].

[73] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Berlin/Heidelberg: Springer, 2003, ISBN: 0387952845.

[74] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," 2010.

[75] M. Liero, A. Mielke, and G. Savaré, "Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures," *Inventiones mathematicae*, vol. 211, no. 3, pp. 969–1117, 2018.

[76] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[77] W. Li, E. K. Ryu, S. Osher, W. Yin, and W. Gangbo, "A parallel method for earth mover's distance," *Journal of Scientific Computing*, vol. 75, no. 1, pp. 182–197, 2018.

[78] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev, "Wasserstein-2 generative networks," *arXiv preprint arXiv:1909.13082*, 2019.

[79] C. Daaloul, T. L. Gouic, J. Liandrat, and M. Tournus, "Sampling from the wasserstein barycenter," *arXiv preprint arXiv:2105.01706*, 2021.

[80] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[81] G. Dal Maso, *An introduction to* Γ*-convergence*. Springer Science & Business Media, 2012, vol. 8.

[82] A. Braides, "A handbook of Γ-convergence," in *Handbook of Differential Equations: stationary partial differential equations*, vol. 3, Elsevier, 2006, pp. 101–213.

[83] 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.

[84] J. A. Carrillo, K. Craig, and Y. Yao, "Aggregation-diffusion equations: Dynamics, asymptotics, and singular limits," in *Active Particles, Volume 2*, Springer, 2019, pp. 65–108.

[85] J. A. Carrillo, K. Craig, and F. S. Patacchini, "A blob method for diffusion," *Calculus of Variations and Partial Differential Equations*, vol. 58, no. 2, p. 53, 2019.

[86] C. Chen, R. Zhang, W. Wang, B. Li, and L. Chen, "A unified particle-optimization framework for scalable bayesian sampling," *arXiv preprint arXiv:1805.11659*, 2018.

[87] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. M. Oberman, "How to train your neural ode: The world of jacobian and kinetic regularization," in *arXiv preprint arXiv:2002.02798*, 2020.

[88] L. Yang and G. Karniadakis, "Potential flow generator with L_2 optimal transport regularity for generative models," in *arXiv preprint arXiv:1908.11462*, 2019.

[89] Y. Chen, T. Georgiou, and M. Pavon, "On the relation between optimal transport and schrödinger bridges: A stochastic control viewpoint," in *Journal of Optimization Theory and Applications*, 2016.

[90] ——, "Optimal transport in systems and control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, 2020.

[91] A. Lin, S. Fung, W. Li, L. Nurbekyan, and S. Osher, "Apac-net: Alternating the population and agent control via two neural networks to solve high-dimensional stochastic mean field games," in *arXiv preprint arXiv:2002.10113*, 2020.

[92] R. Flamary *et al.*, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.

[93] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Sep. 2001.

[94] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[95] E. Nelson, *Quantum Fluctuations*, ser. Princeton Series in Physics. Princeton, N.J: Princeton University Press, 1985.

[96] D. Qi and A. J. Majda, "Low-dimensional reduced-order models for statistical response and uncertainty quantification: Barotropic turbulence with topography," *Physica D: Nonlinear Phenomena*, vol. 343, pp. 7–27, 2017.

[97] Q. Liu and D. Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm," *arXiv:1608.04471 [cs, stat]*, 2016. arXiv: 1608.04471 [cs, stat].

[98] M. Pavon, E. G. Tabak, and G. Trigila, "The data-driven Schroedinger bridge," *arXiv:1806.01364 [math]*, 2018. arXiv: 1806.01364 [math].

[99] J. Sirignano and K. Spiliopoulos, "Mean field analysis of neural networks," *arXiv preprint arXiv:1805.01053*, 2018.

[100] L. Pichler, A. Masud, and L. A. Bergman, "Numerical solution of the fokker–planck equation by finite difference and finite element methods—a comparative study," in *Computational Methods in Stochastic Dynamics*, Springer, 2013, pp. 69–85.

[101] J. Chang and G. Cooper, "A practical difference scheme for fokker-planck equations," *Journal of Computational Physics*, vol. 6, no. 1, pp. 1–16, 1970.

[102] P. Kumar and S. Narayanan, "Solution of fokker-planck equation by finite element and finite difference methods for nonlinear systems," *Sadhana*, vol. 31, no. 4, pp. 445–461, 2006.

[103] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[104] S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[105] ——, *Information Geometry and Its Applications*, ser. Applied Mathematical Sciences volume 194. Japan: Springer, 2016.

[106] N. Ay, J. Jost, H. V. Lê, and L. J. Schwachhöfer, *Information Geometry*, ser. Ergebnisse Der Mathematik Und Ihrer Grenzgebiete A @series of Modern Surveys in Mathematics$l3. Folge, Volume 64. Cham: Springer, 2017.

[107] W. Li, "Geometry of probability simplex via optimal transport," *arXiv:1803.06360 [math]*, 2018. arXiv: 1803.06360 [math].

[108] A. T. Lin, W. Li, S. Osher, and G. Montufar, *Wasserstein proximal of GANs*, 2019.

[109] W. Li and G. Montufar, "Natural gradient via optimal transport," *arXiv:1803.07033 [cs, math]*, 2018. arXiv: 1803.07033 [cs, math].

[110] ——, "Ricci curvature for parametric statistics via optimal transport," *arXiv preprint arXiv:1807.07095*, 2018.

[111] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep relu neural networks," *Neural Networks*, vol. 108, pp. 296–330, 2018.

[112] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[113] U. Grenander and M. I. Miller, "Representations of knowledge in complex systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 4, pp. 549–581, 1994.

[114] G. O. Roberts, R. L. Tweedie, *et al.*, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.

[115] J. A. Carrillo, K. Craig, L. Wang, and C. Wei, "Primal dual methods for wasserstein gradient flows," *arXiv preprint arXiv:1901.08081*, 2019.

[116] A. J. Leverentz, C. M. Topaz, and A. J. Bernoff, "Asymptotic dynamics of attractive-repulsive swarms," *SIAM Journal on Applied Dynamical Systems*, vol. 8, no. 3, pp. 880–908, 2009.

[117] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, D. Slepčev, *et al.*, "Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations," *Duke Mathematical Journal*, vol. 156, no. 2, pp. 229–271, 2011.

[118] A. Klar and S. Tiwari, "A multiscale meshfree method for macroscopic approximations of interacting particle systems," *Multiscale Modeling & Simulation*, vol. 12, no. 3, pp. 1167–1192, 2014.

[119] P.-E. Jabin, "A review of the mean field limits for vlasov equations," *Kinetic & Related Models*, vol. 7, no. 4, p. 661, 2014.

[120] J. A. Carrillo, Y.-P. Choi, and M. Hauray, "The derivation of swarming models: Mean-field limit and wasserstein distances," in *Collective dynamics from bacteria to crowds*, Springer, 2014, pp. 1–46.

[121] D. Maoutsa, S. Reich, and M. Opper, "Interacting particle solutions of fokker-planck equations through gradient-log-density estimation," *arXiv preprint arXiv:2006.00702*, 2020.

[122] S. Pathiraja and S. Reich, "Discrete gradients for computational bayesian inference," *arXiv preprint arXiv:1903.00186*, 2019.

[123] S. Reich and S. Weissmann, *Fokker-planck particle systems for bayesian inference: Computational approaches*, 2020. arXiv: 1911.10832 [math.NA].

[124] E. Weinan, J. Han, and A. Jentzen, "Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations," *Communications in Mathematics and Statistics*, vol. 5, no. 4, pp. 349–380, 2017.

[125] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[126] Y. Khoo, J. Lu, and L. Ying, "Solving parametric pde problems with artificial neural networks," *arXiv preprint arXiv:1707.03351*, 2017.

[127] ——, "Solving for high-dimensional committor functions using artificial neural networks," *Research in the Mathematical Sciences*, vol. 6, no. 1, p. 1, 2019.

[128] Y. Zang, G. Bao, X. Ye, and H. Zhou, "Weak adversarial networks for high-dimensional partial differential equations," *arXiv preprint arXiv:1907.08272*, 2019.

[129] N. Nüsken and L. Richter, *Solving high-dimensional hamilton-jacobi-bellman pdes using neural networks: Perspectives from the theory of controlled diffusions and measures on path space*, 2020. arXiv: 2005.05409 [math.OC].

[130] W. Li, S. Liu, H. Zha, and H. Zhou, "Parametric fokker-planck equation," in *Geometric Science of Information*, F. Nielsen and F. Barbaresco, Eds., Cham: Springer International Publishing, 2019, pp. 715–724, ISBN: 978-3-030-26980-7.

[131] G. A. Pavliotis, *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*. Springer, 2014, vol. 60.

[132] T. Lelièvre and G. Stoltz, "Partial differential equations and stochastic methods in molecular dynamics," *Acta Numerica*, vol. 25, pp. 681–880, 2016.

[133] T. Schlick, *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*. Springer Science & Business Media, 2010, vol. 21.

[134] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 681–688.

[135] I. Goodfellow *et al.*, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014.

[136] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[137] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[138] E. Pardoux and A. Y. Veretennikov, "On the poisson equation and diffusion approximation. i," *Annals of probability*, pp. 1061–1085, 2001.

[139] V. A. Volpert, *Elliptic partial differential equations*. Springer, 2011, vol. 1.

[140] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[141] J. L. Doob, "The brownian movement and stochastic equations," *Annals of Mathematics*, pp. 351–369, 1942.

[142] M. Essid, D. Laefer, and E. G. Tabak, "Adaptive Optimal Transport," *arXiv:1807.00393 [math]*, 2018. arXiv: 1807.00393 [math].

[143] S. Afriat, "Theory of maxima and the method of lagrange," *SIAM Journal on Applied Mathematics*, vol. 20, no. 3, pp. 343–357, 1971.

[144] J. Martens and R. Grosse, "Optimizing neural networks with kronecker-factored approximate curvature," in *International conference on machine learning*, 2015, pp. 2408–2417.

[145] W. Li, A. T. Lin, and G. Montúfar, "Affine natural proximal learning," *unknown*, 2019.

[146] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[147] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[148] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011.

[149] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[150] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[151] A. T. Patera, "A spectral element method for fluid dynamics: Laminar flow in a channel expansion," *Journal of computational Physics*, vol. 54, no. 3, pp. 468–488, 1984.

[152] B. Szabó and I. Babuška, *Finite element analysis*. John Wiley & Sons, 1991.

[153] D. A. C. Cabrera, P. Gonzalez-Casanova, C. Gout, L. H. Juárez, and L. R. Reséndiz, "Vector field approximation using radial basis functions," *Journal of Computational and Applied Mathematics*, vol. 240, pp. 163–173, 2013.

[154] D. Yarotsky, "Error bounds for approximations with deep relu networks," *Neural Networks*, vol. 94, pp. 103–114, 2017.

[155] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, "Nonlinear approximation and (deep) relu networks," *arXiv preprint arXiv:1905.02199*, 2019.

[156] R. Holley and D. Stroock, "Logarithmic sobolev inequalities and stochastic ising models," *Journal of statistical physics*, vol. 46, no. 5, pp. 1159–1194, 1987.

[157] D. Bakry and M. Émery, "Diffusions hypercontractives," in *Séminaire de Probabilités XIX 1983/84*, Springer, 1985, pp. 177–206.

[158] J. Lott, "Some geometric calculations on wasserstein space," *Communications in Mathematical Physics*, vol. 277, no. 2, pp. 423–437, 2008.

[159] F. Otto and C. Villani, "Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality," *Journal of Functional Analysis*, vol. 173, no. 2, pp. 361–400, 2000.

[160] S. Surjanovic and D. Bingham, *Virtual library of simulation experiments: Test functions and datasets*, Retrieved February 8, 2020, from http://www.sfu.ca/~ssurjano, 2020.

[161] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*. Springer Science & Business Media, 2013, vol. 23.

[162] H. Rosenbrock, "An automatic method for finding the greatest or least value of a function," *The Computer Journal*, vol. 3, no. 3, pp. 175–184, 1960.

[163] P. H. Rabinowitz, "Periodic solutions of Hamiltonian systems," *Comm. Pure Appl. Math.*, vol. 31, no. 2, pp. 157–184, 1978.

[164] V. I. Arnold, *Mathematical methods of classical mechanics*, Second, ser. Graduate Texts in Mathematics. Springer-Verlag, New York, 1989, vol. 60, pp. xvi+508, Translated from the Russian by K. Vogtmann and A. Weinstein, ISBN: 0-387-96890-3.

[165] J. Mawhin and M. Willem, *Critical point theory and Hamiltonian systems*, ser. Applied Mathematical Sciences. Springer-Verlag, New York, 1989, vol. 74, pp. xiv+277, ISBN: 0-387-96908-X.

[166] J. Maas, "Gradient flows of the entropy for finite Markov chains," *J. Funct. Anal.*, vol. 261, no. 8, pp. 2250–2292, 2011.

[167] C. Villani, *Optimal transport*, ser. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, vol. 338, pp. xxii+973, Old and new, ISBN: 978-3-540-71049-3.

[168] E. Schrödinger, *Über die Umkehrung der Naturgesetze*, ser. Sitzungsberichte der Preussischen Akademie der Wissenschaften. Physikalisch-mathematische Klasse. Akad. d. Wissenschaften, 1931.

[169] E. Nelson, "Derivation of the Schrödinger equation from Newtonian mechanics," *Physical Review*, vol. 150, no. 4, pp. 1079–1085, 1966.

[170] E. Madelung, "Quanten theorie in hydrodynamischer form," *Zeitschrift für Physik*, vol. 40, no. 3-4, pp. 322–326, 1927.

[171] S. Chow, W. Li, and H. Zhou, "A discrete Schrödinger equation via optimal transport on graphs," *J. Funct. Anal.*, vol. 276, no. 8, pp. 2440–2469, 2019.

[172] J. Benamou and Y. Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numer. Math.*, vol. 84, no. 3, pp. 375–393, 2000.

[173] M. Pavon, "Quantum Schrödinger bridges," in *Directions in mathematical systems theory and optimization*, ser. Lect. Notes Control Inf. Sci. Vol. 286, Springer, Berlin, 2003, pp. 227–238.

[174] C. Léonard, "A survey of the Schrödinger problem and some of its connections with optimal transport," *Discrete Contin. Dyn. Syst.*, vol. 34, no. 4, pp. 1533–1574, 2014.

[175] W. Gangbo, W. Li, and C. Mou, "Geodesics of minimal length in the set of probability measures on graphs," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 25, p. 78, 2019.

[176] C. Léonard, "Lazy random walks and optimal transport on graphs," *Ann. Probab.*, vol. 44, no. 3, pp. 1864–1915, 2016.

[177] S. Chow, W. Li, C. Mou, and H. Zhou, "A discrete schrodinger bridge problem via optimal transport on graphs," *Journal of Dynamics and Differential Equations*, vol. 20, no. 33, p. 34, 2020.

[178] E. A. Carlen, "Conservative diffusions," *Communications in Mathematical Physics*, vol. 94, no. 3, pp. 293–315, 1984.

[179] S. Chow, W. Li, and H. Zhou, "Wasserstein Hamiltonian flows," *J. Differential Equations*, vol. 268, no. 3, pp. 1205–1219, 2020.

[180] A. J. van der Schaft, "Port-Hamiltonian systems: An introductory survey," in *International Congress of Mathematicians. Vol. III*, Eur. Math. Soc., Zürich, 2006, pp. 1339–1365.

[181] A. J. van der Schaft and B. M. Maschke, "Port-Hamiltonian systems on graphs," *SIAM J. Control Optim.*, vol. 51, no. 2, pp. 906–937, 2013.

[182] J. Cui, L. Dieci, and H. Zhou, "Time discretizations of Wasserstein-Hamiltonian flows," *arXiv:2006.09187*, 2020.

[183] V. N. Kolokoltsov, *Nonlinear Markov processes and kinetic equations*, ser. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2010, vol. 182, pp. xviii+375, ISBN: 978-0-521-11184-3.

[184] Y. Chen, T. Georgiou, M. Pavon, and A. Tannenbaum, "Robust transport over networks," *IEEE Trans. Automat. Control*, vol. 62, no. 9, pp. 4675–4682, 2017.

[185] C. Léonard, "Girsanov theory under a finite entropy condition," in *Séminaire de Probabilités XLIV*, ser. Lecture Notes in Math. Vol. 2046, Springer, Heidelberg, 2012, pp. 429–465.

[186] G. Conforti and M. Pavon, "Extremal flows in Wasserstein space," *J. Math. Phys.*, vol. 59, no. 6, pp. 063502, 15, 2018.

[187] A. Blaquiére, "Controllability of a Fokker-Planck equation, the schrödinger system, and a related stochastic optimal control," *Dynamics and Control*, vol. 2, pp. 235–253, 1992.

[188] T. M. Cherry, "On periodic solutions of Hamiltonian systems of differential equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 227, pp. 137–221, 1928.

[189] G. Teschl, *Ordinary differential equations and dynamical systems*, ser. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2012, vol. 140, pp. xii+356, ISBN: 978-0-8218-8328-0.

[190] J. Dolbeault, B. Nazaret, and G. Savaré, "A new class of transport distances between measures," *Calc. Var. Partial Differential Equations*, vol. 34, no. 2, pp. 193–231, 2009.

[191] B.-R. Nawaf and V.-E. Eric, "Continuous-time random walks for the numerical solution of stochastic differential equations," *Mem. Amer. Math. Soc.*, vol. 256, no. 1228, pp. v+124, 2018.

[192] A. Braides *et al.*, *Gamma-convergence for Beginners*. Clarendon Press, 2002, vol. 22.

[193] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[194] S. Jin, L. Li, and J.-G. Liu, "Random batch methods (rbm) for interacting particle systems," *Journal of Computational Physics*, vol. 400, p. 108 877, 2020.

VITA

Shu Liu was born in Funing, Yancheng, China on January, 31, 1993. He was raised up in Nanjing. He attended Chu Kochen Honors College of Zhejiang University in Hangzhou in the summer of 2012. He graduated from Zhejiang University with Bachelor degree in science majored in mathematics and applied mathematics in 2016. He started pursuing Ph.D. degree in Computational Science and Engineering with home unit in mathematics under the supervision of Prof. Haomin Zhou at Georgia Tech in the summer of 2016.