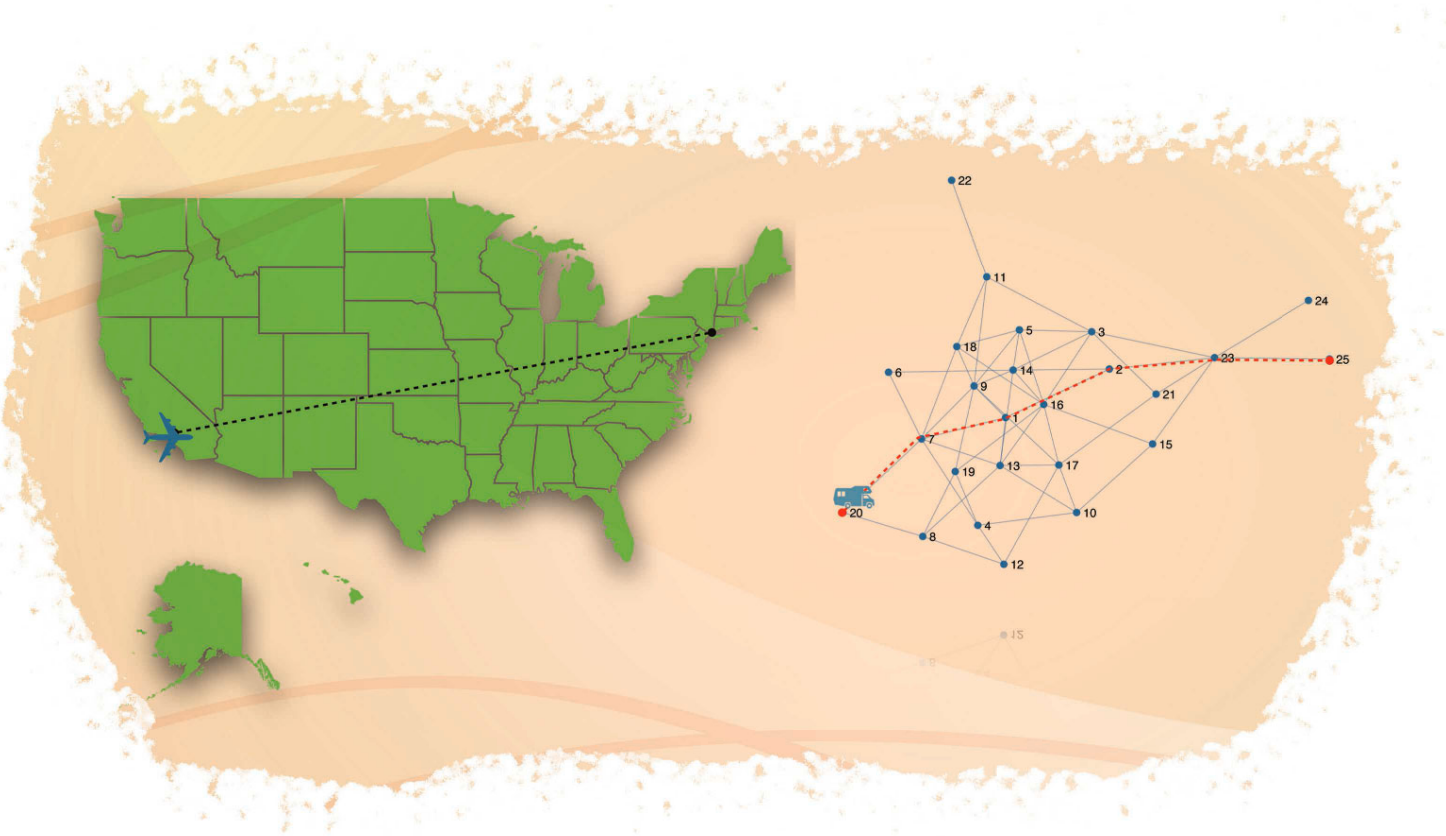
Optimal Transport on Networks

AN INTRODUCTION TO THE DIFFERENCES BETWEEN DISCRETE AND CONTINUOUS SPACES WITH AN APPLICATION IN PATH PLANNING



HAOMIN ZHOU

Have you ever thought about how to ship a collection of packages from their original locations to final destinations, how to construct a wall using a pile of bricks, or how to pair customers with providers in an industrial supply chain? If your answer is yes to any of these questions, you thought about optimal transport. If your answer is no, you still must have encountered problems like how to drive from one location to another in the shortest time or distance or how to walk from one place to another quickly. We all have thought about optimal transport, either consciously or subconsciously. The problem can be as simple as moving a box from one point to another in a straight line (which obviously is the optimal solution), or it can be more complicated (when the number of packages becomes large and the optimal solution can't be identified easily, even with the help of modern supercomputers).

It is undoubtable that optimal transport has been practiced throughout human history. See "Summary" for more information. Mathematicians formulated those problems as abstractions, as they always do, which reads as *finding the optimal way to change, by means of transport, one nonnegative function $f(x)$ to another $g(x)$ with equal mass*. In this statement, the nonnegative function $f(x)$ may refer to the number of packages at their original locations, and $g(x)$ represents their final destinations. Alternately, $f(x)$ and $g(x)$ are piles of

Digital Object Identifier 10.1109/MCS.2021.3076541
Date of current version: 19 July 2021

1066-033X/21©2021IEEE

bricks in different shapes, before and after, respectively, the construction of a wall. Optimality considers the lowest cost or highest reward. The first recorded description was found in an article written by the French mathematician/engineer Gaspard Monge in 1781 [29], and a major breakthrough was obtained by Nobel laureate Leonid Kantorovich during World War II [19]. In light of their contributions, optimal transport problems are often referred to as Monge–Kantorovich problems in mathematics. See "Monge and Kantorovich Optimal Transport" for more precise formulations.

Up to a constant scaling [namely dividing the two nonnegative functions by their mass; $\mu(x) = f(x)/\| f(x)\|$ and $\nu(x) = g(x)/\| g(x)\|$, where $\|\cdot\|$ is the L_1-norm], the problem can be equivalently posed as transporting a probability density function μ, a nonnegative function with unit mass, to another ν with optimal cost. This seemingly simply operation is a vital step allowing the study to be carried out in a unified framework written in the language of probability, which facilitates the modern theory on optimal transport. At the core of this theory is a mathematical concept called the Wasserstein distance between two probability distributions, defined as the lowest cost of transporting from one distribution to the other. This canonical distance enables a mechanism transforming the probability space into a Riemannian manifold (known as a Wasserstein manifold), so that geometric structures and partial differential equation (PDE) techniques can be established and analyzed (which provides powerful tools in many applications).

In the past few decades, optimal transport has become an active research area, attracting mathematicians, scientists, and engineers. Its theory and applications have been

Summary

Optimal transport, also known as the Monge–Kantorovich problem, is a mathematical theory that has received considerable attention in recent decades because of its applications across different disciplines. It has a close tie to control theory because it can be formulated as an optimal control with partial differential equation constraints (known as the Benamou–Brenier formula). The existing theory is mainly developed for continuous underlying spaces (like the Euclidean space $\mathbb{R}^d$ or a smooth manifold, such as a sphere). Its extension to networks was initiated more recently. The discrete structure of a network creates extra difficulties, making the extension surprisingly subtle. Among many contributing factors, the main challenges are that a network is not a length space (in which one curve can be continuously moved to another), a network may not be properly embedded in a continuous space, and it is nontrivial to define white noise on a network. These difficulties fundamentally change the optimal transport theory on networks. This article discusses those challenges and explains how to define the Wasserstein distance on networks and how it helps to construct an algorithm for path planning in an unknown environment.

Monge and Kantorovich Optimal Transport

The first recorded mathematical formulation of optimal transport (the Monge problem, when paraphrased in modern mathematical notation) can be posed as finding a map T between two sets X and Y with equal mass, such that an associated cost function $c(x, y)$ is minimized [29]. More precisely, assume that the mass on X, denoted as $\mu(X)$, equals the mass on Y, $\nu(Y)$, and the cost to transport by T a unit mass located at $x \in X$ to another location $y = T(x) \in Y$ is $c(x, T(x))$. The total cost of transporting X to Y is

$$C(T) = \int_X c(x, T(x))\, d\mu(x). \tag{S1}$$

The notation of $\mu(x)$ is also used as the Radon measure defined on X. The optimal transport map T^* is the one that minimizes $C(T)$; that is,

$$C(T^*) = \min_T C(T). \tag{S2}$$

The Monge problem is a difficult mathematical problem, attracting researchers to study it for generations. Whether or not there exists a solution (and especially how to compute the solution, if it exists) remain challenging. The fact that T is a 1:1 map in (S2) restricts the mass at one point to being transported only to another point (which prohibits mass splitting from one point to multiple targets). Kantorovich's relaxation removes this restriction [19]. Instead of considering the map T, Kantorovich proposed to study the total cost function

$$\mathcal{W}(\kappa) = \int_{X\times Y} c(x, y)\, d\kappa(x, y), \tag{S3}$$

where $\kappa(x, y)$ is a joint distribution defined on $X \times Y$. Let M be the set of joint distributions with two marginals given by μ and ν; that is,

$$M = \Big\{\kappa(x, y) \text{ is a Radon measure on } X\times Y \,\Big|\, \int_Y d\kappa(x, y) = \mu(x),\ \int_X d\kappa(x, y) = \nu(y)\Big\}.$$

The optimal joint distribution $\kappa^*(x, y)$ is the one minimizing $\mathcal{W}(\kappa)$,

$$\mathcal{W}(\kappa^*) = \min_{\kappa\in M} \mathcal{W}(\kappa). \tag{S4}$$

Each map T induces a joint distribution in M, while the reverse is not necessarily true. This is the reason that the solution of (S4) is considered a "weak" or "relaxed" solution of (S2). The linearity of Kantorovich relaxation (which is widely known as linear programming in the discrete case [13] and its dual formulation) makes it popular in applications [15], [32], [39].

Three main ingredients are used to define the Wasserstein distance using an optimal control viewpoint: the probability space, the cost function, and a vector field used for transportation.

undergoing rapid developments with many remarkable contributions, such as the relation between optimal transport maps and the Monge–Ampere equation [4]; the existence, regularity, and structure of solutions [17], [25], [27]; gradient flows [1], [18]; and the Ricci curvature bounds on Wasserstein manifolds [23], [35]. The list can go on. There is no attempt to survey the vast literature on the subject in this article. Readers are directed to an introductory note [15]; several books for rigorous, yet comprehensive, descriptions of its mathematical theory and applications [32], [34], [39]; and a more recent book for its computational aspect and applications in machine learning [31]. The discussion of this article is based on the Benamou–Brenier formula, which recasts the problem in the form of optimal control, giving an explicit dynamical description of the optimal transport process [3].

Currently, optimal transport has become a rich theory, sitting at the intersection of several branches in mathematics, such as PDEs, probability, geometry, dynamical systems, and numerical analysis. It also finds connections to other disciplines, including physics, chemistry, computer science, engineering, and economics. The majority of the existing theory assumes that the underlying space for transportation is continuous (such as $\mathbb{R}^n$, the sphere, or other smooth manifolds). Its extension to discrete spaces, such as networks or graphs, emerged more recently. Rigorous mathematical investigations were initiated independently [9], [26], [28] from three different angles at around the same time, providing complementary understandings of the theory. Similar to the story in continuous space, the introduction of the Wasserstein distance in a discrete space transforms the probability simplex into a Riemannian manifold, on which we can establish and perform PDEs and geometric analysis. This leads to various advancements in the theory, such as gradient flows and entropy inequalities [5], [10], Ricci curvatures in discrete space [14], [16], [30], and many applications, for example, in information geometry [22], biological networks [33], opinion dynamics in social networks [21], and game theory [12].

The goal of this article is to provide a quick introductory tour, without getting into the technical details, of the benefits and challenges regarding the fascinating topic of optimal transport on networks. More attention is paid to the differences between the discrete and continuous cases and their incurred consequences through three questions: 1) How does one define optimal transport and Wasserstein distance on networks using an optimal control framework? 2) What is the gradient flow induced by the Wasserstein distance on networks? 3) What is white noise in networks? Addressing these questions helps to explain several conceptual and technical difficulties that fundamentally change the optimal transport theory on networks. As an application example, a recently proposed algorithm for path exploration in unknown environments is briefly described at the end of the article. It is noted that there are other extensions, such as optimal transport for vector- and matrix-valued functions [7], [8], that are not discussed here.

DEFINING OPTIMAL TRANSPORT ON NETWORKS

The consideration of optimal transport on networks is motivated by real-world applications. Some are from traditional lines of research, such as shipping products using limited transport networks, in which the goal is to deliver goods using the lowest cost. Others are from emergent phenomena, such as information propagation on social media. The network can be Facebook or Twitter, with each user being a node. The objective is to determine the probability of each node receiving the information at a given time. In addition, the optimal solutions (also called optimal transport maps in the literature) can only be found analytically for special cases, such as in 1D problems. Computer simulations become mandatory to calculate their solutions. In this case, it is necessary to discretize the space and calculate the solution on a grid, which can be viewed as a network. Ultimately, addressing optimal transport on networks becomes inevitable for certain practical problems.

There are several possible approaches to address the problem on networks. Discretization is at the top of the candidate list. If the network is a grid, the problem can be studied by discretizing its continuous counterpart, and one hopes that the theory and conclusions may be transferred naturally. However, this approach is surprisingly subtle because of the structural differences between continuous and discrete spaces. Of course, if the network in consideration cannot be regarded as a discretization of a continuous space, this approach is not applicable. This marks the first major difference between discrete and continuous problems.

Three main ingredients are used to define the Wasserstein distance using an optimal control viewpoint: the probability space, the cost function, and a vector field used for transportation. To better explain the problem, denote $G = (V, E)$ as the network, with V being the node set and E

the edge set. A weight and/or direction may be associated to each edge. However, for simplicity, those considerations are not adopted here. It is also assumed that the network is finite and connected without self-looping edges. The probability space is a collection of normalized nonnegative functions defined on nodes,

$$\mathcal{P}(G)=\left\{\rho\in\mathbb{R}^n\,\middle|\,\sum_{i=1}^{n}\rho_i=1,\quad \rho_i\geq 0,\quad \text{for any } i\in V\right\},$$

where ρ_i is the probability at node i, and n is the number of nodes of the network G. In other words, $\rho\in\mathcal{P}(G)$ means that ρ is an n-dimensional vector whose entries are between zero and one, and the sum of all entries is one. The optimal transport problem is finding the vector field v to transport one probability function μ to another ν within the probability space $\mathcal{P}(G)$ so that the total cost is minimized. The optimal cost is called the Wasserstein distance between μ and ν.

In existing studies, the cost may be related to distance. For example, the square of the Euclidean distance between two points in $\mathbb{R}^n$ is among the most commonly used cost functions in continuous space, resulting in the famous 2-Wasserstein distance between two distributions μ and ν, given by the Benamou–Brenier formula,

$$W_2(\mu,\nu)=\inf_{v}\left\{\int\frac{1}{2}\|v(x,t)\|^2\rho(x,t)dxdt\right\}^{\frac{1}{2}},$$

$$\text{subject to}\quad \frac{\partial\rho}{\partial t}+\nabla\cdot(v\rho)=0,\qquad \rho(0)=\mu,\quad \rho(1)=\nu. \tag{1}$$

This is an optimal control problem with a PDE constraint, known as the transport equation, whose purpose is to transfer the initial density μ to the target density ν by vector field v (see Figure 1). It is known that not all vector fields can transport μ to ν. However, among those that can, the Benamou–Brenier formula seeks the one that minimizes the total kinetic energy used in the transport process. The minimizer vector field v^* induces the optimal transport solution for Monge–Kantorovich problems.

The generalization of Euclidean distance between two nodes on a network becomes ambiguous. If the network is obtained by discretizing a finite region in $\mathbb{R}^n$, its inherited Euclidean distance can be taken. Otherwise, the Euclidean distance cannot be an option. A popular choice in practice is the smallest number of edges connecting the two nodes, which can be applied to any network. In this case, the distance only takes nonnegative integer values, which may not necessarily agree with the inherited Euclidean distance. Thus, which choice of distance to use becomes a relevant question.

Each definition of distance between two nodes can endow its own formulation for the Wasserstein distance in $\mathcal{P}(G)$. In applications, the preferred one is consistent with the 2-Wasserstein distance in the continuous case. In this case, consistency means that the definition of the 2-Wasserstein distance on a network should be the same as or close to the 2-Wasserstein distance in the continuous space if the network is from a discretization of a continuous region. This seemingly simple question does not have a simple answer, and the reason for this can be illustrated with a 2D lattice grid, as shown in Figure 2. Consider four connected nodes A, B, C, and D, forming a square with unit length for each edge. The Euclidean distance between the two diagonal nodes A and C is $\sqrt{2}$. The challenge is that transportation cannot be done in the diagonal direction because there is no edge. When going along the edges, via either B or D, the travel distance becomes two. This discrepancy creates problems. It is worth emphasizing that consistency is vital since a consistent definition may allow one to transfer desirable properties established in the continuous space to the discrete case and to empower its usage in applications.

The fact that transportation is allowed only along edges on networks causes intrinsic differences between the continuous and discrete cases. Let us still consider $\mathbb{R}^2$ and a 2D lattice grid G as an example. Transportation can happen in all directions in the continuous case, while only four directions—up, down, left, and right—are allowed in the discrete situation. In other words, the workspace is homogeneous in the continuous case, while it is inhomogeneous on a network. This becomes more problematic if the network is heterogeneous, meaning that the number of edges connected

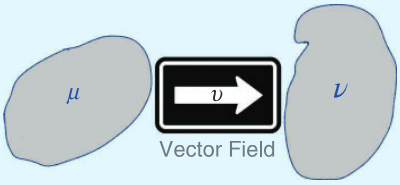


FIGURE 1 The vector field v in the transport equation is selected so that the initial density μ can be transported to the target density ν through v.

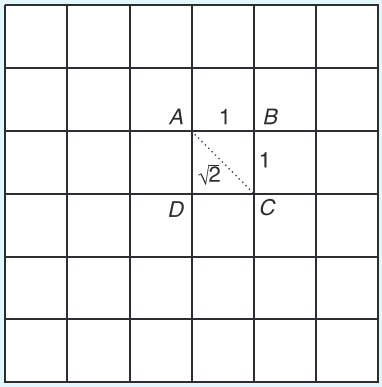


FIGURE 2 On a lattice network, transport can only be done along the edges. The travel distance between A and C is two, not the Euclidean distance $\sqrt{2}$.

The fact that transportation is allowed only along edges on networks causes intrinsic differences between the continuous and discrete cases.

to a node may be different for each node. Such differences cause several technical difficulties in defining optimal transport on networks. The first issue is that a vector field v on a network must be defined on the edges, not on the nodes. Meanwhile, the probability function stays on the nodes, which is a mismatch. In contrast, v and ρ are defined at every point in the continuous space. Thus, the integration in (1) can be completed without any problem.

The second challenge is that $\mathbb{R}^2$ is a so-called length space, in which infinitely small adjustments in the transport direction can be made if needed, while adjustments must jump from one edge to another in the discrete case. In other words, there exists no small change in discrete directions. The properties of a length space play crucial roles in the development of the existing theory for continuous problems. Not having those properties prevents one from adopting many well-established continuous techniques and conclusions. Much of the recent theoretical research in mathematics on this topic is devoted to overcome those difficulties. For example, it has been proven that the optimal solution of (1) must be achieved by using a constant-speed geodesic, a curve $\rho(x,t)$ in the probability space satisfying the scaling relation

$$W_2(\rho(s),\rho(t)) = |t-s|\ W_2(\mu,\nu), \quad \text{for any } s,t \text{ in } [0,1].$$

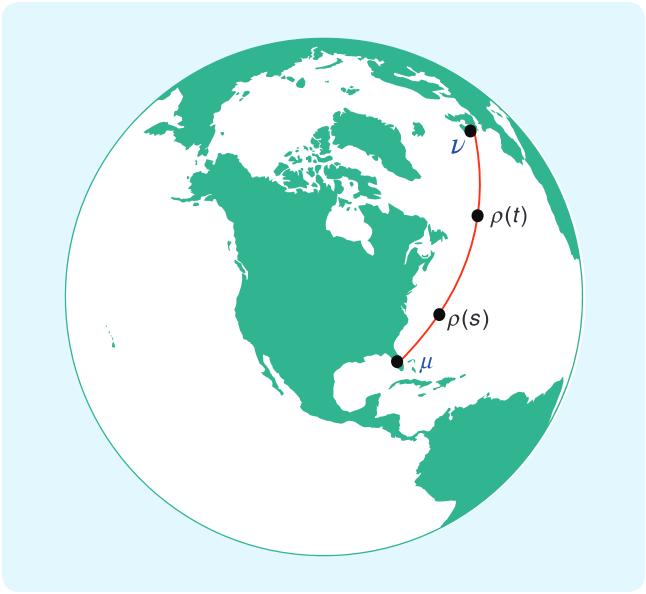


FIGURE 3 In the continuous case, the geodesic between μ and ν must be a curve on the Wasserstein manifold with constant speed. The Wasserstein distance between any two points $\rho(s)$ and $\rho(t)$ must be $|t-s|\,W_s(\mu,\nu)$ for any s, t in the interval $[0,1]$.

In other words, as shown in Figure 3, the geodesic must be a curve on the Wasserstein manifold with constant speed. The Wasserstein distance between any two points, $\rho(s)$ and $\rho(t)$, on the geodesic connecting μ and ν is proportional to the time difference $|t-s|$. However, this is not possible for curves in the discrete setting. These difficulties fundamentally change the optimal transport theory on networks.

Despite the challenges, the Wasserstein distance on G can still be defined by an optimal control formulation,

$$W_2(\mu,\nu) = \inf_v \Big\{ \int (v,v)_\rho dt \Big\}^{\frac{1}{2}},$$
$$\text{subject to} \quad \frac{\partial \rho}{\partial t} + \operatorname{div}_G(v\rho) = 0, \qquad \rho(0) = \mu, \quad \rho(1) = \nu, \tag{2}$$

in which the discrete inner product and divergence operator are given, respectively, by

$$(v,v)_\rho = \frac{1}{2}\sum_{(i,j)\in E} v_{ij}^2 \theta_{ij}(\rho), \qquad \operatorname{div}_G(\rho v)_i = -\sum_{j\in N(i)} v_{ij}\theta_{ij}(\rho).$$

The weight function θ_{ij} is introduced to compensate the mismatch between v and ρ so that the influence of the node-defined ρ can be extended to the edges, and vice versa. It is noted that (2) is an optimization with ordinary differential equation constraints. Its goal is to find a time-dependent vector v^*, defined on the edges of the graph, so that the kinetic energy is minimized while also transporting the node-defined function μ to ν. Obviously, (2) can be viewed as an analog of (1). The caveat is the choice of θ, which must be selected carefully. The conditions on how to select θ have been studied in the articles by Chow et al., 2018 [10] and Maas [26]. Among many possible candidates, a commonly studied one is the nonlinear logarithmic mean $\theta_{ij} = (\rho_i - \rho_j)/(\log\rho_i - \log\rho_j)$ [9], [26]. This nonlinearity has deep implications for the properties of optimal transport on networks.

The introduction of the discrete Wasserstein distance W_2 transforms the probability space $\mathcal{P}(G)$ into a Riemannian manifold, a function space equipped with a smooth metric (the term metric is used interchangeably with distance in this article, although their mathematical definitions are different) on which geometric operations can be conducted. For convenience, the probability space, coupled with the optimal transport distance, is also called the Wasserstein manifold, denoted by $(\mathcal{P}(G), W_2)$, in mathematics. Several recent advancements in this direction have been reported. For example, Ricci curvatures as well as gradient and Hamiltonian flows are studied on $(\mathcal{P}(G), W_2)$ [11], [14].

THE GRADIENT FLOW ON A WASSERSTEIN MANIFOLD

Symbolically, the gradient flow of a functional $\mathcal{F}(\rho)$ on a Riemannian manifold is

$$\frac{\partial \rho}{\partial t} = -\text{grad}_W \mathcal{F}(\rho),$$

where grad_W is the gradient operator with respect to the metric on the manifold. The negative gradient direction is the steepest descent direction locally. Thus, the gradient flow describes the dynamics of $\rho(t)$ reducing $\mathcal{F}(\rho)$ on the manifold without an external force (such as a ball rolling downhill on a smooth surface). The gradient operator can be derived explicitly on the Wasserstein manifold. For example, in the continuous case, if $\mathcal{F}(\rho)$ is the free energy functional given by a combination of potential energy and entropy,

$$\mathcal{F}(\rho) = \int \Phi \rho dx + \beta \int \rho \log \rho dx,$$

(where Φ is a given potential function and β is a constant), the gradient flow becomes

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla \Phi) + \beta \Delta \rho, \tag{3}$$

which is the well-known Fokker–Planck equation. If the potential Φ is a constant, the equation is reduced to the standard heat equation. This routine calculation reveals the surprising fact that the Fokker–Planck equation is the gradient flow of free energy $\mathcal{F}$ on the Wasserstein manifold. The heat equation is a special case, indicating that the Laplace term $\Delta\rho$ is the gradient of the entropy $\int \rho \log \rho$ with respect to the Wasserstein distance. This result was first reported in the seminal work of Jordan et al. [18]. It has profound impacts on modern optimal transport theory and its related applications. An immediate consequence, when viewing the Fokker–Planck equation as a gradient flow, is that the free energy $\mathcal{F}(\rho)$ is a Lyapunov function [meaning that the free energy value decreases monotonically along the solution of (3)]. This further implies that there exists a unique asymptotic limit of the solution, regardless of where the initial distribution is. Furthermore, a direct calculation verifies that the asymptotic solution is actually the famous Gibbs distribution:

$$\rho^* = \left(\int e^{-\frac{\Phi(x)}{\beta}} dx \right)^{-1} e^{-\frac{\Phi(x)}{\beta}}.$$

The aforementioned gradient flow derivation can be mimicked, although not trivially, for Wasserstein manifolds on networks (leading to similar conclusions). However, the structure of the gradient flow shows significant departures from the existing knowledge in the continuous setting. A toy example can be used to illustrate the problem. Take G as a 1D lattice network with only five nodes. A potential value Φ_i is assigned on each node, and they are plotted in Figure 4(a). The potential function determines a discrete Gibbs distribution,

$$\rho_i^* = \left(\sum_{i=1}^{5} e^{-\frac{\Phi_i}{\beta}} \right)^{-1} e^{-\frac{\Phi_i}{\beta}},$$

which is plotted in Figure 4(b). According to the theory, the Gibbs distribution is the unique minimizer for the free energy and the asymptotic solution of the Fokker–Planck equation. However, checking the minimizer is straightforward. Verifying the asymptotic solution is not. When discretizing the Fokker–Planck equation by the center-difference method (a commonly used numerical scheme for parabolic equations), its asymptotic solution is depicted in Figure 4(c) (which is different from the Gibbs distribution). By monitoring the free energy along the solution, the curve shown in Figure 5(a) is observed. It does not decay in time. Instead, it increases along the solution. Those disagreements indicate that the discrete Fokker–Planck equation obtained by using the center-difference scheme cannot be the gradient flow of the free energy with respect to the discrete Wasserstein metric. From classical textbook material [38], the commonly used schemes for (3) are all

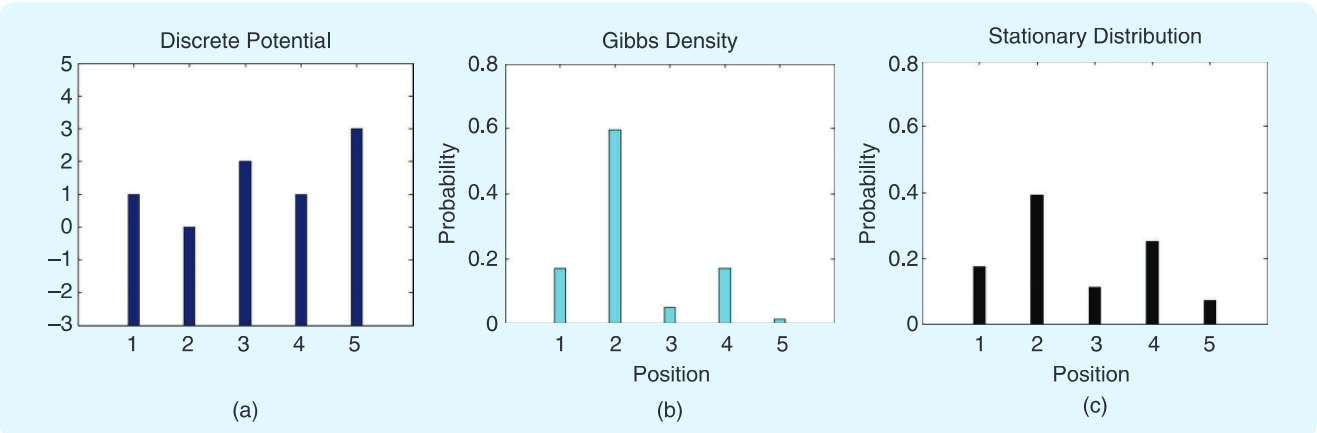


FIGURE 4 (a) The potential function defined on a five-point lattice network. (b) Its corresponding Gibbs distribution. (c) An asymptotic solution of discrete Fokker–Planck equation (3) using the center-difference scheme.

This routine calculation reveals the surprising fact that the Fokker–Planck equation is the gradient flow of free energy $\mathcal{F}$ on the Wasserstein manifold.

linear because it is a linear equation. Numerical experiments show that they also cannot reach the Gibbs distribution. Hence, there must be something fundamentally different for the Wasserstein manifold on networks. In fact, it is rigorously proven [9] that the gradient flow of free energy on the discrete Wasserstein manifold *must* be expressed by nonlinear equations, for example,

$$\frac{d\rho_i}{dt} = \sum_{j\in N(i)} ((\Phi_j + \beta\log\rho_j) - (\Phi_i + \beta\log\rho_i))\theta_{ij}(\rho). \tag{4}$$

With different choices of θ_{ij}, (4) may result in different nonlinear Fokker–Planck equations. This also echoes the nonlinearity in the selection of θ_{ij} on networks and contrasts sharply with the linearity of the Fokker–Planck equation of (3). Regardless of the choice of θ_{ij}, the corresponding nonlinear Fokker–Planck equation is always the gradient flow of the free energy with respect to the Wasserstein distance given by (2), and the Gibbs distribution is its asymptotic solution. For instance, in Figure 5(b), the free energy curve along the solution of (4) decays as expected, and its asymptotic solution is exactly the discrete Gibbs distribution shown in Figure 4(b). The theory and experiments agree perfectly in this example. It is interesting to note that the Fokker–Planck equation (3) can be written in a conservative form,

$$\frac{\partial\rho}{\partial t} = \nabla\cdot(\rho\nabla(\Phi + \beta\log\rho)),$$

whose discretization is consistent with the form used in (4). It is also worth noting that, if the weight function $\theta_{ij}(\rho)$ is chosen as the logarithmic mean, $(\rho_i - \rho_j)/(\log\rho_i - \log\rho_j)$, it cancels the logarithmic terms in (4), resulting in $(\rho_i - \rho_j)$ (which becomes linear). This is one reason why the logarithmic mean is a favorite choice. However, its impact on Φ terms remains nonlinear, which marks a fundamental difference between discrete and continuous spaces.

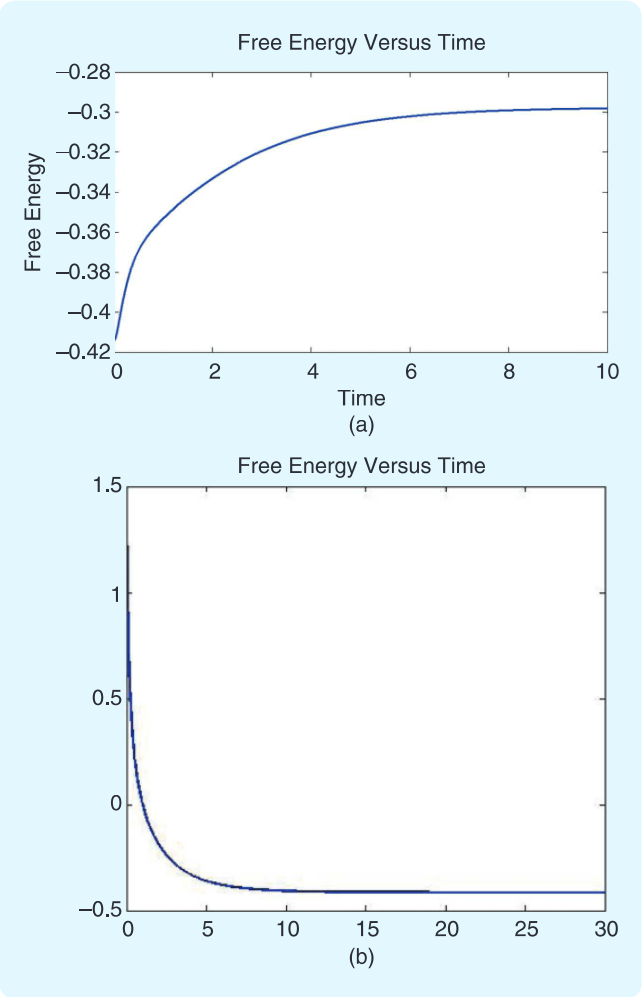


FIGURE 5 Free energy versus time along the solution. (a) The Fokker–Planck equation (3) using center-difference scheme. It doesn't decay. (b) The nonlinear discrete Fokker–Planck equation (4); it decays as the theory predicted.

WHITE NOISE IN NETWORKS

White noise, a seemingly remote concept, plays an essential role in the development of optimal transport theory. This can be seen through the connections between the Fokker–Planck equation and stochastic differential equations (SDEs). The classical diffusion theory, which was developed decades earlier than the modern optimal transport theory, states that the solution ρ of Fokker–Planck equation (3) is the probability density function for a random variable $X(t)$ satisfying the well-known Langevin dynamics,

$$dX(t) = -\nabla\Phi(X(t))dt + \sqrt{2\beta}\,dW(t), \tag{5}$$

where $W(t)$ is the Brownian motion, and $dW(t)$ is the white noise (which is added to model the randomness or uncertainties of the vector field). The intrinsic links between Fokker–Planck equation (3) and SDE (5) as well as between (3) and the free energy $\mathcal{F}(\rho)$ reveal the deep connections between classical diffusion theory and optimal transport (thanks to the gradient flow structure on the Wasserstein manifold).

Mathematically, the white noise $dW(t)$ is defined as an independent identically distributed random variable following a standard Gaussian distribution. It is homogeneous in all directions. The word *white* is selected to indicate that

White noise, a seemingly remote concept, plays an essential role in the development of optimal transport theory.

there is no preference in frequency or direction, just as white color is a mix of all other colors with equal contributions. Otherwise, a random perturbation is called colored noise. The property of being homogeneous in all directions is precisely the challenging point in extending the concept of white noise to networks because (as explained before) any perturbation to a dynamical process on a network must be conducted along its edges. It is hard, if not impossible, to be homogeneous on a heterogeneous network.

The challenge of properly defining white noise on networks can be demonstrated from another angle, through a concept called a random walk, which is regarded as an analog or approximation of Brownian motion in the literature. A random walk on a network is defined as a stochastic process that has equal probability of jumping from a node to one of its neighboring nodes. As explained in "Random Walks on Networks," randomly walking on a network asymptotically leads to an equilibrium distribution ρ^* whose value at each node is linearly proportional to the number of edges connected to the node. In other words, the equilibrium ρ^* is not a uniform distribution on a heterogeneous network, which contrasts sharply with a well-known property of Brownian motion, namely, that its probability becomes uniform in an asymptotic time limit. This disagreement further suggests that a random walk on a network cannot be used directly to construct white noise on the network, at least not in the same way that the Brownian motion is used to build white noise in the continuous case.

The connection between Fokker–Planck equation (3) and SDE (5) may shed light on how to approach the challenge. Examining the correspondence between these two equations, it can be found that $-\nabla\Phi(X(t))dt$ and $\nabla\cdot(\rho\nabla(\Phi))$ form a pair, and $dW(t)$ is solely responsible for the Laplace term $\Delta\rho$. The latter pair is the reason for a commonly known statement that "adding white noise is equivalent to adding

Random Walks on Networks

As its name suggests, a random walk on a network is a stochastic process with equal probability of jumping from a node to one of its neighboring nodes. Considering the network depicted in Figure S1 as an example, the probability from node *A* to *B* or *C* is 1/2 because *A* has two edges connected to it. The probability from *A* to *D* is zero since they are not directly connected. The probability from *B* to one of its neighbors is 1/3.The probability is assigned similarly for *C* and *D*. If $\rho(k)$, a four-dimensional column vector, is the probability on the network at step *k*, following the random walk, the probability at the next step $\rho(k+1)$ is calculated by

$$\rho(k+1)=P\rho(k),$$

where *P* is the transition probability matrix

$$P=\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}.$$

Performing a random walk on this network for an infinitely long time (meaning $k\to\infty$), the probability $\rho(k)$ approaches an equilibrium distribution $\rho^*=(1/8)[2,3,2,1]^T$, where *T* is the transpose operator. It is verifiable that ρ^* is the eigenvector of

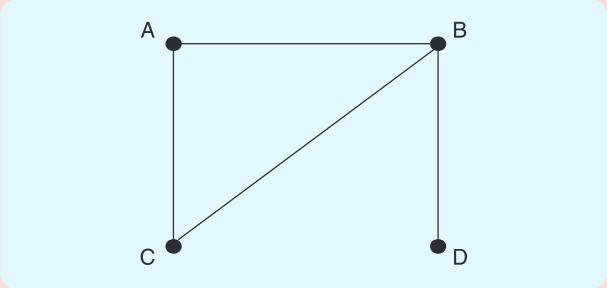


FIGURE S1 A random walk on this network is defined as a stochastic process having equal probability, inversely proportional to the number of edges connected to a node, moving from the node to one of its neighbors.

P corresponding to eigenvalue one, and it is the asymptotic distribution of the random walk [regardless of the initial probability $\rho(0)$]. When the probability reaches the equilibrium, the intake at each node must be the same as its output, so that the net change remains zero. Alternately, the rate of change from *A* to *B* is 1/2, while the rate is 1/3 from *B* to *A*. To maintain the balance, it is necessary to maintain a higher probability at *B* than that at *A*. This explains why the equilibrium probability at each node is linearly proportional to the number of edges connected to the node. Therefore, the equilibrium ρ^* cannot be a uniform distribution for a heterogeneous network.

The task of motion planning is to find a feasible path for the team to move from given initial positions to target configurations while avoiding any collisions with obstacles or violations of constraints.

diffusion" to a dynamical system. Following this rationale, it is desirable to find the corresponding Laplace operator Δ on a network. Considering the Laplace term as the gradient flow of entropy on the Wasserstein manifold, the similarity between (3) and (4) is compared. It is not hard to identify that the Laplace term $\Delta\rho$, which can also be rewritten as $\nabla \cdot (\rho \nabla \log \rho)$, corresponds to all of the terms involving logarithmic functions in (4). This observation leads to the following definition for white noise on networks given in [9].

A Markov process on a network G induced by a potential function Φ is defined by a stochastic process whose transition rate from j and i is given by $(\Phi_j - \Phi_i)$ if $\Phi_j > \Phi_i$ and there exists an edge between i and j. Otherwise, the rate is zero. Adding white noise to this Markov process perturbs the transition rate by modifying Φ_i to $(\Phi_i + \beta \log \rho_i)$, meaning that the transition rate from j and i is changed to $((\Phi_j + \beta \log \rho_j) - (\Phi_i + \beta \log \rho_i))$ if it is positive. When the potential function is a constant, the white noise becomes a stochastic process whose transition rate from j to i is given by $(\log \rho_j - \log \rho_i)$ if $\rho_j > \rho_i$. Otherwise, the white noise reverses its transition direction. This logarithmic-based random process approaches the uniform distribution on the network, regardless of the initial probability (which meets the expectation that the asymptotic limit for the distribution of white noise is uniform). It is widely accepted in the modern optimal transport theory that the heat equation is the gradient flow of entropy with respect to the Wasserstein metric. The heat equation is also the governing equation for the density evolution of white noise in the continuous space. From this viewpoint, the logarithmic definition of white noise on networks is the gradient flow of entropy with respect to the discrete Wasserstein metric, which is consistent with the conclusion in the continuous setting. However, such a definition of white noise is nonlinear in formula, and it can reverse direction depending on the values of ρ_i and ρ_j. Furthermore, it requires the knowledge of the probability function ρ on the network, which may not be easy to obtain unless the available data set is sufficiently large. At the completion of this article, this still remains as a challenge, and further investigation is needed.

A PATH-PLANNING ALGORITHM IN AN UNKNOWN ENVIRONMENT

Optimal transport has been used in different disciplines. As an application example, an algorithm in robotics that was inspired by the gradient flow on the Wasserstein manifold on networks is briefly discussed. The algorithm, which was reported recently by Zhai et al. [42], is used for motion planning of a team of agents in unknown environments in a high-dimensional configuration space. Instead of presenting the technical details of the method, the focus is on the similarities between the design ideas and properties of optimal transport on networks.

Before describing the algorithm, imagine a scenario where water flows from a source to a sink on an uneven landscape in a bounded domain. The terrain is unknown to water, yet water flows in the descent direction, if possible. When trapped at a local minimizer, water accumulates to form a reservoir and eventually overflows at the lowest point of the barrier. In any situation, water eventually reaches the sink. The design principle of the algorithm mimics the motion of water. However, instead of real water, the algorithm follows the Wasserstein gradient flow of free energy [Fokker–Planck equation (4)] on an existing but not explicitly constructed potential tree, which is a special graph structure designed to guide the flow from the starting point (source) to the target location (sink). The task of motion planning is to find a feasible path for the team to move from given initial positions to target configurations while avoiding any collisions with obstacles or violations of constraints. In this consideration, the number of agents is fixed, and some of the obstacles or constraints stay unknown unless one of the team members traverses close enough to them. It is also assumed that the obstacle information, once available, is shared among the team members.

Compared to path planning in known environments, there are several significant challenges when the problem is posed in unknown environments. For example, replanning becomes inevitable because the prior planned path may become infeasible when a newly detected obstacle blocks it. There may exist traps of local minima. When there are multiple agents, one cannot plan their paths individually if there are constraints imposed among team members (for example, they can't be too close to cause collisions or too far away from each other to lose communications). If the number of agents grows, the dimension of the configuration space increases, and the computation cost may increase exponentially (which quickly becomes intractable). This is called the curse of dimensionality in computational mathematics. Note that there is an extensive literature on path-planning methods. Many standard methods (such as the family of bug algorithms [24], the

probabilistic road map [36], and the rapidly exploring random tree [20], [37]) have been adopted for motion planning in unknown environments. They may be efficient in low-dimensional situations. Optimal transport is also used for motion planning in different setups, such as robot swarming by linear programming [2] and robust transport over networks [6]. See Zhai et al. [42] for more references on related work.

Unlike existing methods, the method proposed in [42] explores the idea of Wasserstein gradient flow on a potential tree. The nodes are connected according to the distance to the target in the configuration space. The resulting algorithm is a deterministic strategy with a provable convergence guarantee, meaning that the algorithm stops in a finite number of steps, either returning a locally optimal path or concluding that a feasible one connecting the initial and target positions does not exist. A major advantage is that, using the potential tree, a flexible discrete structure allows the algorithm to be scalable for higher dimensional problems.

The algorithm contains three steps: generate a tree in configuration space, find a path on the tree, and update environment information along the motion. These steps are repeated if planned paths are blocked. To illustrate the main ideas while keeping the discussion simple, let us consider one point robot moving in a 2D domain populated with obstacles of different shapes and sizes. As shown in the plots of Figures 6 and 7, the obstacles are in light gray when they are unknown to the robot. They become dark gray after being detected. Figure 6(a) shows the first generated tree from the starting position near the upper-right corner to the target at the bottom left. Initially, all obstacles are not yet known. Thus, the tree is expanded from the starting point toward the target through some obstacles and then stops at the target. There is a unique path on the tree connecting the initial and final points [displayed in Figure 6(b)]. Obviously, when the robot moves along this path, it encounters obstacles, and the path needs to be replanned. Figure 7 depicts another iteration of tree generation and path finding. In Figure 7(a), the current position is near a newly detected obstacle, shown in dark gray. The tree is expanded toward the target. However, it excludes the red nodes inside the known obstacles. The red nodes near the target indicate the new nodes added to the tree. Once reaching the target, a new path on the tree can be identified, as shown in

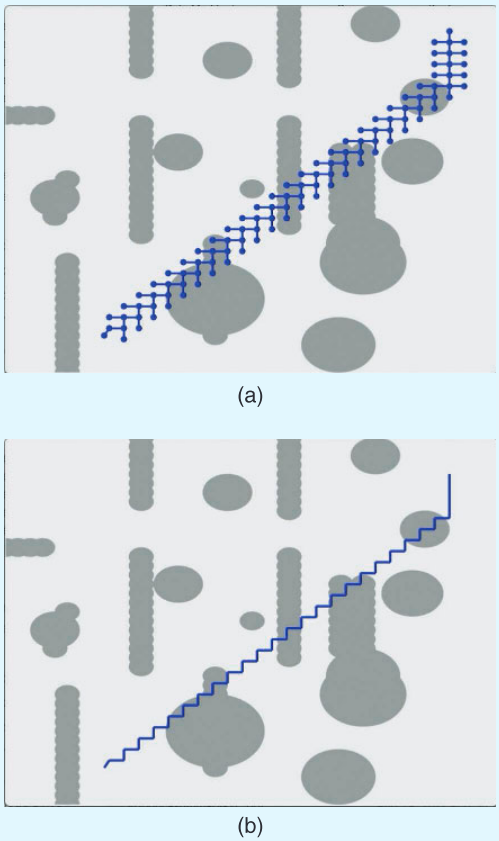


FIGURE 6 (a) The generated tree from the start point to the target. (b) The path on the tree. Since the obstacles are not yet known to the robot, the tree and path go through some obstacles.

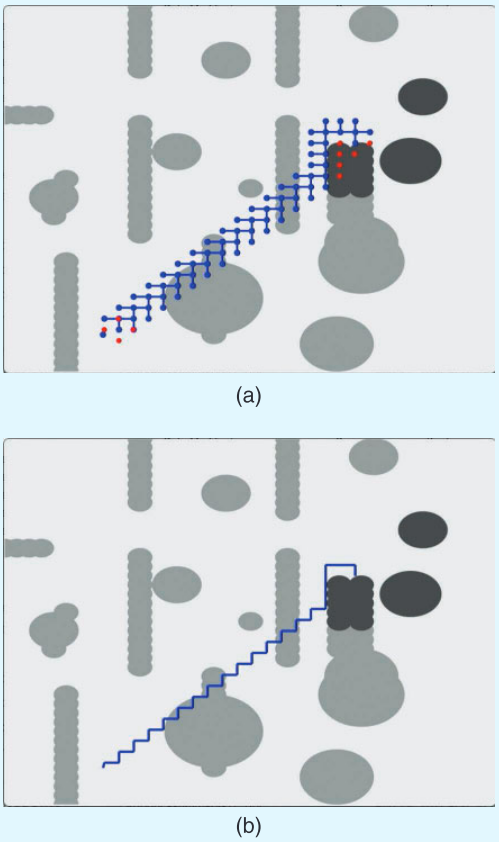


FIGURE 7 (a) The tree from the current position. (b) The path from the current position. The dark gray obstacles are known to the robot; the generated tree and path no longer go through them.

In this case, the algorithm introduces diffusion to grow the tree (meaning new nodes along the boundary of obstacles are added to the tree until a new feasible descent direction can be identified).

Figure 7(b). A YouTube video [40] demonstrates the complete process of finding the path for this example.

Another video [41] shows the complete motion of a team of five agents (in a 10D configuration space) in the same environment. It is noticed from the example that the tree grows by adding new nodes only along coordinate directions with a fixed step size. No diagonal node is used in the tree. This keeps the computational complexity low. If the diagonal nodes were added, there could possibly be up to 2^d new nodes added to the tree at each step, where d is the dimension of the configuration space. There are at most $2d$ new nodes if only the coordinate directions are allowed.

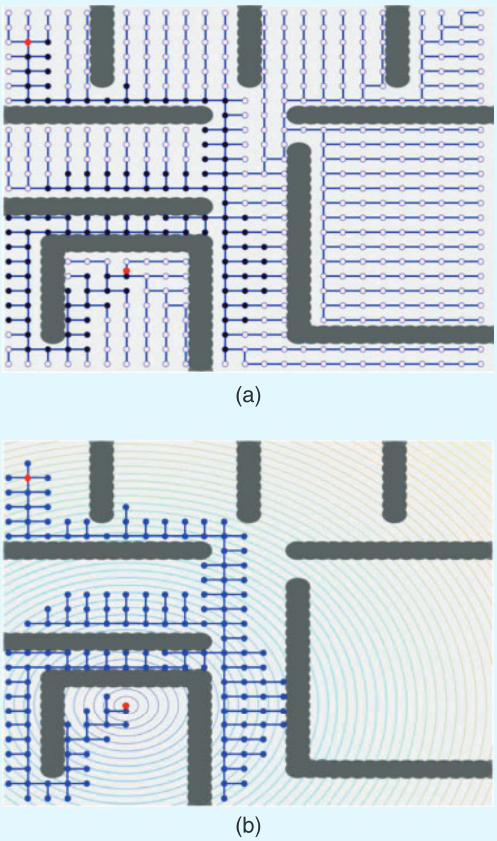


FIGURE 8 (a) The whole potential tree. (b) The region containing the nodes generated by the algorithm. The selected nodes are the ones for which the solution of (4) takes nonzero values following the intermittent diffusion process and also their immediate neighbors.

With this choice, the complexity of the algorithm increases linearly with respect to the dimension, which can be scaled in the high-dimensional space. The tree is extended toward the target by introducing a potential function Φ, such as the distance to the target position. The node with the lowest potential value on the tree is selected, and its neighbors are added to the tree if they are not in a known obstacle or already on the tree. This procedure ensures that the tree grows to reduce the potential value, mimicking gradient descent actions. Since only the nodes along the coordinate directions are used, the tree does not expand according to the exact gradient descent direction. Instead, it grows in an approximated potential decreasing direction. If the agents move along the path without encountering an obstacle, the task is finished. When the path is blocked, replanning is needed. The agents may be trapped at a local minimum, where there is no new feasible node with a lower potential value to be added to the tree. In this case, the algorithm introduces diffusion to grow the tree (meaning new nodes along the boundary of obstacles are added to the tree until a new feasible descent direction can be identified). It is also observed from the video that the algorithm only explores a subregion without knowing the environment a priori.

Recall the analogy of water flowing from a source to a sink. This algorithm is designed by following not the flow of real water, but the solution of Fokker–Planck equation (4) on the potential tree. As illustrated in Figure 8, nodes of the potential tree are the circles, and the edges are solid lines connecting them. The tree is constructed according to the distance to the target [which is the potential Φ, whose level lines are plotted in Figure 8(b)]. The start and end points are marked in red. The initial value for (4) is a point mass distribution at the start point. The evolution is completed with β being zero when the solution is not trapped at a local minimum (which corresponds to the gradient descent situation). Otherwise, β is taken as a positive value that corresponds to the diffusion scenario. Since the gradient descent and diffusion processes take place alternately, the process is called intermittent diffusion. This evolution determines the subregion within which the algorithm searches for a path. The solid nodes on the tree in Figure 8(a) are the ones where the solution takes nonzero values. The tree generated by the algorithm is within the region indicated by the solid nodes in Figure 8(b). These nodes are the nonzero-valued nodes and their feasible neighbors. Note that this path-planning algorithm does not require one to explicitly construct the

entire potential tree in advance or to solve (4) numerically. The tree and (4) are only used in guiding the algorithmic design and proving its theoretical properties, such as convergence. As the gradient flow on the Wasserstein manifold, the solution of (4) converges to the Gibbs distribution (which takes the maximum value at the target configuration). This property provides the mathematical foundation to prove the completeness of the algorithm in a finite number of steps. This differs from existing path-planning algorithms using randomness, in which the completeness (if one exists) is proven to be achievable only asymptotically.

AUTHOR INFORMATION

Haomin Zhou (hmzhou@math.gatech.edu) received the B.S. in pure mathematics from Peking University, the M. Phil. in applied mathematics from the Chinese University of Hong Kong, and the Ph.D. in computational mathematics from the University of California, Los Angeles. He was a postdoctoral scholar at the California Institute of Technology for three years. He has been a faculty member in the School of Mathematics at the Georgia Institute of Technology, Atlanta, Georgia, 30332, USA, since 2003. His research interests include optimal transport, inverse problems, and stochastic differential equations.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savare, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures (Lectures in Mathematics)*. Basel: ETH Zürich Birkhäuser Verlag, 2008.
[2] S. Bandyopadhyay, S.-J. Chung, and F. Y. Hadaegh, "Probabilistic swarm guidance using optimal transport," in *Proc. IEEE Conf. Control Appl. (CCA)*, 2014, pp. 498–505.
[3] 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. doi: 10.1007/s002110050002.
[4] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Comm. Pure Appl. Math.*, vol. 44, no. 4, pp. 375–417, 1991. doi: 10.1002/cpa.3160440402.
[5] E. Carlen and J. Maas, "Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance," *J. Funct. Anal.*, vol. 273, no. 5, pp. 1810–1869. doi: 10.1016/j.jfa.2017.05.003.
[6] Y. Chen, T. Georgiou, M. Pavon, and A. Tannenbaum, "Robust transport over networks," *IEEE Trans. Autom. Control*, vol. 62, no. 9, pp. 4675–4682, Sept. 2017. doi: 10.1109/TAC.2016.2626796.
[7] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Vector-valued optimal mass transport," *SIAM J. Appl. Math.*, vol. 78, no. 3, pp. 1682–1696, doi: 10.1137/17M1130897.
[8] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Matrix optimal mass transport: A quantum mechanical approach," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2612–2619. doi: 10.1109/TAC.2017.2767707.
[9] S.-N. Chow, W. Huang, Y. Li, and H. M. 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. doi: 10.1007/s00205-011-0471-6.
[10] S.-N. Chow, W. Li, and H. Zhou, "Entropy dissipation of Fokker–Planck equations on graphs," *Discrete Cont. Dynam. Syst. A*, vol. 38, no. 10, pp. 4929–4950, 2018. doi: 10.3934/dcds.2018215.
[11] S.-N. Chow, W. Li, and H. M. Zhou, "Wasserstein Hamiltonian flow," *J. Diff. Equation*, vol. 268, no. 3, pp. 1205–1219, 2020. doi: 10.1016/j.jde.2019.08.046.
[12] S.-N. Chow, W. Li, J. Lu, and H. M. Zhou, "Equilibrium selection via optimal transport," *SIAM J. Appl. Math.*, vol. 80, no. 1, pp. 142–159, 2020. doi: 10.1137/18M1163828.
[13] G. B. Dantzig, "Application of the simplex method to a transportation problem," *Activity Anal. Prod. Allocation*, vol. 13, pp. 359–373, 1951.
[14] M. Erbar and J. Maas, "Ricci curvature of finite Markov chains via convexity of the entropy," *Arch. Ration. Mech. Anal.*, vol. 206, no. 3, pp. 997–1038, 2012. doi: 10.1007/s00205-012-0554-z.
[15] L. C. Evans, "Partial differential equations and Monge–Kantorovich Mass Transfer," in *Current Developments in Mathematics*, S. T. Yau, Ed. 1997.
[16] M. Fathi and J. Maas, "Entropic Ricci curvature bounds for discrete interacting systems," *Ann. Appl. Prob.*, vol. 26, no. 3, pp. 1774–1806.
[17] W. Gangbo and R. J. McCann, "Optimal maps in Monge's mass transport problem," *C. R. Acad. Sci. Paris Ser. I Math*, vol. 321, no. 12, pp. 1653–1658, 1995.
[18] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker–Planck equation," *SIAM J. Math. Anal.*, vol. 29, no. 1, pp. 1–17, 1998. doi: 10.1137/S0036141096303359.
[19] L. Kantorovich, "On the transfer of masses," *Dokl. Akad. Nauk. SSSR*, vol. 37, pp. 227–229, 1942.
[20] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," 1998.
[21] H. Lavenant and B. Maury, "Opinion propagation on social networks: A mathematical standpoint," *ESAIM: Procs*, vol. 67, pp. 285–335, 2020. doi: 10.1051/proc/202067016.
[22] W. Li and G. Montúfar, "Natural gradient via optimal transport," *Info. Geo*, vol. 1, no. 2, pp. 181–214, 2018. doi: 10.1007/s41884-018-0015-3.
[23] J. Lott and C. Villani, "Ricci curvature for metric-measure spaces via optimal transport," *Ann. Math*, vol. 169, no. 3, pp. 903–991, doi: 10.4007/annals.2009.169.903.
[24] V. Lumelsky and A. Stepanov, "Dynamic path planning for a mobile automaton with limited information on the environment," *IEEE Trans. Autom. Control*, vol. 31, no. 11, pp. 1058–1063, 1986. doi: 10.1109/TAC.1986.1104175.
[25] X.-N. Ma, N. S. Trudinger, and X.-J. Wang, "Regularity of potential functions of the optimal transportation problem," *Arch. Ration. Mech. Anal.*, vol. 177, no. 2, pp. 151–183, 2005. doi: 10.1007/s00205-005-0362-9.
[26] J. Maas, "Gradient flows of the entropy for finite Markov chains," *J. Funct. Anal.*, vol. 261, no. 8, pp. 2250–2292, 2011. doi: 10.1016/j.jfa.2011.06.009.
[27] R. J. McCann, "Existence and uniqueness of monotone measure-preserving maps," *Duke Math. J.*, vol. 80, no. 2, pp. 309–323, 1995. doi: 10.1215/S0012-7094-95-08013-2.
[28] A. Mielke, "A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems," *Nonlinearity*, vol. 24, no. 4, pp. 1329–1346, 2011. doi: 10.1088/0951-7715/24/4/016.
[29] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. (Open Library): de l'Imprimerie Royale, 1781.
[30] Y. Ollivier abd C. Villani, "A curved Brunn–Minkowski inequality on the discrete hypercube, or: What is the Ricci curvature of the discrete hypercube?" *SIAM J. Discrete Math.*, vol. 26, no. 3, pp. 983–996, 2012. doi: 10.1137/11085966X.
[31] G. Peyré and M. Cuturi, "Computational optimal transport," *Found. Trends Mach. Learn.*, 2019.
[32] S. Rachev and L. Rüschendorf, *Mass Transportation Problems*. New York: Springer-Verlag, 1998.
[33] R. Sandhu, S. Tannenbaum, T. Georgiou, and A. Tannenbaum, "Geometry of correlation networks for studying the biology of cancer," in *Proc. IEEE 55th Conf. Decision Control (CDC)*, Las Vegas, NV, 2016, pp. 2501–2506.
[34] F. Santambrogio, *Optimal Transport for Applied Mathematicians, Calculus of Variations, PDEs, and Modeling*. Berlin: Springer-Verlag, 2015.
[35] K.-T. Sturm, "On the geometry of metric measure spaces. I and II," *Acta Math.*, vol. 196, no. 1, pp. 65–177, 2006. doi: 10.1007/s11511-006-0002-8.
[36] P. Svestka, "Robot motion planning using probabilistic roadmaps," Ph.D. thesis, Universiteit Utrecht, 1997.
[37] Y. Tian et al., "Application of RRT-based local path planning algorithm in unknown environment," in *Proc. Int. Symp. Comput. Intell. Robot. Automat. (CIRA 2007)*, 2007, pp. 456–460.
[38] J. W. Thomas, *Numerical Partial Differential Equations: Finite Difference Methods*. New York: Springer-Verlag, 1995.
[39] C. Villani, "Optimal transport," in *Old and New* (*Grundlehren der Mathematischen Wissenschaften*, vol. 338). Berlin: Springer-Verlag, 2009.
[40] MLSS Africa. *Optimal Transport Path Planning in Unknown Environment*. (Jan. 10, 2019). [Online Video]. Available: youtu.be/g8MWRv1wk0A
[41] H. Zhai, *Five Robots Move in Unknown Environment*. (Oct. 25, 2018). [Online Video]. youtu.be/H5lfzAYbfRA
[42] H. Zhai, M. Egerstedt, and H. M. Zhou, "Path planning in unknown environments using optimal transport theory," 2019. [Online]. Available: arxiv.org/abs/1909.11235