Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

INTRODUCTION TO OPTIMAL TRANSPORT

Ilya O. Ryzhov[1], Raghu Pasupathy[2], and Harsha Honnappa[2]

[1]Robert H. Smith School of Business, College Park, MD, USA
[2]Department of Statistics, Purdue University, West Lafayette, IN, USA
[3]School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

ABSTRACT

We review optimal transport (OT) which can be informally described as "deforming" a source probability distribution into a target probability distribution with minimal cost. OT was formulated more than two centuries ago and became famous for its relevance to economics and logistics especially after the seminal work on linear programming by Kantorovich, Koopmans, and Dantzig. OT has since seen multiple resurgences, first due to the rise of computer vision in the 1990s, and more recently due to the maturing of large-scale optimization solvers alongside the rise of machine learning and artificial intelligence. This tutorial formally introduces OT to a simulation audience that is steeped in the concepts of operations research (OR). The early parts of the tutorial focus on application contexts. This is followed by the Monge and Kantorovich OT formulations along with key structural results and examples. The tutorial ends with a short section on semidiscrete OT.

1 INTRODUCTION AND PERSPECTIVE

Every undergraduate student majoring in OR eventually sees the "transportation problem" (Ford and Fulkerson 1956) as an example application of linear programming. This problem is formulated as

$$\min_{h_{m,n}} \sum_{m=1}^{M} \sum_{n=1}^{N} c_{m,n} h_{m,n} \tag{1}$$

subject to $h_{m,n} \geq 0$ and

$$\sum_{n=1}^{N} h_{m,n} = p_m, \quad m = 1,...,M, \tag{2}$$

$$\sum_{m=1}^{M} h_{m,n} = q_n, \quad n = 1,...,N. \tag{3}$$

The typical intepretation of this model is that goods are being shipped from M "supply nodes" (such as warehouses) to N "demand nodes" (such as distribution centers). The decision variables $h_{m,n}$ determine the number of units to ship from supply node m to demand node n. The coefficients $c_{m,n}$ represent unit cost of shipment from m to n. The right-hand side values p_m (respectively, q_n) represent the total units supplied by node m (demanded by node n). Without loss of generality, these values may be normalized, i.e., $\sum_m p_m = \sum_n q_n = 1$, thus representing the proportion of overall supply at node m or demand at node n. In that case, the decision variable can be viewed as a "joint probability," or the proportion of total units to be assigned to a particular supply/demand combination. The objective function can then be interpreted as the expected cost of shipping one unit; our goal is to find a transportation policy h that minimizes this expected value.

Problem (1)-(3) is, of course, trivial to solve using linear programming. We may, however, generalize it by making one or both of the supply/demand distributions continuous. Suppose, for example, that we

still have M supply facilities, but demand can now arise anywhere in some geographical region $\mathscr{Y} \subseteq \mathbb{R}^2$. Then, (1) becomes

$$\min_h \sum_{m=1}^{M} \int_{\mathscr{Y}} c(m,y)\, h(m,y)\, dy. \tag{4}$$

The decision variable is now an infinite-dimensional function h defined on $\{1,...,M\} \times \mathscr{Y}$. Constraints (2)-(3) now become

$$\int_{\mathscr{Y}} h(m,y)\, dy = p_m, \quad m = 1,...,M, \tag{5}$$

$$\sum_{m=1}^{M} h(m,y) = g(y), \quad y \in \mathscr{Y}, \tag{6}$$

where g is a probability density function supported on $\mathscr{Y}$. This formulation is known as the *semidiscrete optimal transport* problem, and also has applications in logistics, particularly in geographical partitioning or districting problems (see Section 3 for examples).

In a more general form, the *optimal transport* problem (Villani 2021) transforms one continuous probability distribution into another. Formally, we may have nearly arbitrary domains $\mathscr{X}$ and $\mathscr{Y}$ and solve

$$\min_h \int_{\mathscr{X}} \int_{\mathscr{Y}} c(x,y)\, h(x,y)\, dxdy. \tag{7}$$

subject to

$$\int_{\mathscr{Y}} h(x,y)\, dy = f(x), \quad x \in \mathscr{X}, \tag{8}$$

$$\int_{\mathscr{X}} h(x,y)\, dx = g(y), \quad y \in \mathscr{Y}, \tag{9}$$

wherein our goal is now to find a joint density h coupling the two marginal densities f and g in a way that minimizes expected cost (7). This general formulation has diverse applications in economics, computer vision, machine learning, statistics, algorithmic fairness, and other contexts. Of course, it is also much harder to solve than the transportation LP! Most computational methods focus on special cases, and virtually all research on the general problem assumes the quadratic cost function $c(x,y) = \|x-y\|_2^2$ (in which case the objective is called the *Wasserstein 2-distance* between f and g).

The authors of this tutorial all have research interests in simulation. All three of us, independently, became interested in OT because of its deep connections to ideas and methods that are very well-known to the simulation community. OT itself may not be a "simulation" problem in a strict sense, but simulation researchers appear to be in an excellent position to contribute to it. Moreover, OT has a way of arising unexpectedly in many operations research problems involving optimization and uncertainty. A good example is the tutorial by Blanchet and Shapiro (2023) on distributionally robust optimization (DRO), which appeared in last year's WSC proceedings. In this paper, Kantorovich duality results, which are central to the study of OT, were used to tractably reformulate various DRO problems.

This tutorial aims to provide a comprehensive introduction to OT for operations researchers in general, and for simulation researchers in particular. No prior knowledge of OT is assumed. Unlike many overviews of this area (e.g., Villani 2021), which tend to focus on the deep theory underlying the general case, we devote considerable space to the semidiscrete problem, as it has many interesting applications and can be solved tractably using simulation optimization methods. Indeed, some of these approaches may offer useful guidance for how to think about the more general version. Also, the tutorial does not cover "control" versions of OT where, in addition to the couplings associated with the transport variables, one also introduces certain other cost variables that evolve in time. See, for instance, the treatment in (Carlier and Lachapelle 2009).

The tutorial is organized as follows. After some mathematical preliminaries in Section 2, the tutorial presents a variety of application contexts in Section 3. This is followed by Section 4, Section 5 and Section 6 which draw heavily on (Villani 2021; Peyré, Cuturi, et al. 2019; Carlier 2012) to provide an intuitive presentation of the basic OT formulations along with the key results. Section 7 treats the semidiscrete OT formulation, followed by some concluding remarks in Section 8.

2 PRELIMINARIES

In this section, we introduce key notation and definitions used throughout the paper.

2.1 Notation

(a) $\mathbb{1}_m$ is the $m \times 1$ column vector of ones. (b) For a metric space $(\mathscr{Z}, d_z)$, $\mathscr{P}(\mathscr{Z})$ refers to the Borel probability measures on $\mathscr{Z}$, that is, the set of measures μ satisfying $\mu(A) \geq 0$ for each $A \in \Sigma_z$, $\mu(\mathscr{Z}) = 1$, and $\mu(\bigcup_{j=1}^{\infty} A_j) = \sum_j \mu(A_j)$ for a countable collection $\{A_j, j \geq 1\}$ of pairwise disjoint sets in $\mathscr{Z}$. (c) $X \sim \mu$ means X is distributed according to the probability measure μ. (d) $\mathrm{C}(X)$ denotes the space of continuous real-valued functions on X, and $\mathrm{BC}(X)$ the space of bounded continuous real-valued functions on X. (e) If μ and ν are probability measures on measurable spaces $(\mathscr{X}, \Sigma_x)$ and $(\mathscr{Y}, \Sigma_y)$, respectively, then $\mu \otimes \nu$ refers to the product probability measure defined on $(\mathscr{X} \times \mathscr{Y}, \Sigma_x \otimes \Sigma_y)$ given by $\mu \otimes \nu(A_1 \times A_2) = \mu(A_1)\nu(A_2), A_1 \in \Sigma_x, A_2 \in \Sigma_y$.

2.2 Definitions

Definition 1 (atomic and non-atomic) A probability measure μ on (Z, Σ_z) is said to be *non-atomic* if for each $A \in \Sigma_z$ satisfying $\mu(A) > 0$, there exists $B \subset A$ with $B \in \Sigma_z$ such that $0 < \mu(B) < \mu(A)$. The probability measure μ is *atomic* if it is not non-atomic, that is, there exists $A \in \Sigma_z$ such that any $B \subset A$ with $B \in \Sigma_z$ implies either $\mu(B) = \mu(A)$ or $\mu(B) = 0$.

Definition 2 (support of a measure) The *support* of a probability measure $\mu \in \mathscr{P}(\mathscr{X})$ is the set consisting of points x such that every open neighborhood of x has positive probability under μ. Formally,

$$\mathrm{supp}(\mu) := \bigcup \{x \in \mathscr{X} : \mu(N_x) > 0, N_x \text{ is any open neighborhood of } x\}. \tag{10}$$

Equivalently, $\mathrm{supp}(\mu)$ is the largest set C such that any open set having a non-empty intersection with C has positive measure assigned to it by μ. The support should be loosely understood as the smallest set such that the measure assigned to the set is one.

3 APPLICATION CONTEXTS

We describe several applications of the optimal transport framework. Section 3.1 describes geographical partitioning problems. Section 3.2 describes the use of OT to remove algorithmic bias. Section 3.3 discusses applications to color transfer in image processing, while Section 3.4 discusses text mining. Section 3.5 briefly mentions some applications in statistics.

3.1 Geographical Partitioning

The partitioning, or districting, problem is an instance of the semidiscrete formulation (4)-(6). Again, recall that $\mathscr{Y} \subseteq \mathbb{R}^2$ is a geographical region. There are M supply facilities with known, fixed locations $y_1, ..., y_M \in \mathscr{Y}$. Customers appear randomly (for example, following a spatio-temporal Poisson process), and the location of each new customer is distributed according to a density g supported on $\mathscr{Y}$.

Consider a customer appearing at some fixed location $y \in \mathscr{Y}$. The customer must choose (or be assigned to) one of the supply facilities. The cost of assigning the customer to the mth facility is proportional to the distance that the customer has to travel, i.e., $c(m, y) = \|y - y_m\|_2$. The mth facility has enough resources to serve a proportion p_m of all customers.

The goal in geographical partitioning (Carlsson et al. 2016) is to find disjoint sets $A_1,...,A_M \subseteq \mathscr{Y}$ such that $y_m \in A_m$ for all m, $\bigcup_{m=1}^{M} A_m = \mathscr{Y}$, and $\int_{A_m} g(y)\,dy = p_m$. Essentially, each A_m represents a zone served exclusively by the mth facility, thus giving us a very simple assignment rule: the location y of a new customer must belong to exactly one of the sets A_m, and the customer must be assigned to the facility serving that set.

Strictly speaking, this problem is more restrictive than (4)-(6), because the latter allows us to make assignments probabilistically: a customer appearing at location y is assigned to the mth facility with probability $\frac{h(m,y)}{g(y)}$. As we will see in Section 7, however, there is no loss of generality in requiring assignments to be deterministic, because the optimal solution to (4)-(6) will produce a geographical partition under some mild regularity conditions.

3.2 Algorithmic Fairness

Chzhen et al. (2020) considers the following situation. Let X be a scalar-valued random variable representing the output of some predictive statistical model. The simplest possible case is the linear model $Y = \beta^\top V$ for some $\beta \in \mathbb{R}^p$ and some random vector V of covariates. However, there is no loss of generality in considering more complicated statistical models; for example, Y could even be a prediction calculated by a deep neural network.

Suppose, however, that there is another discrete random variable Z which is correlated with Y in some way, and that this correlation is undesirable. To give a specific context, Zhu and Ryzhov (2024) considers an algorithmic hiring problem where V is a vector describing attributes of a job applicant (such as undergraduate GPA, years of work experience, test scores, etc.), Y is a prediction of the applicant's performance, and Z is a "sensitive attribute" indicating whether the applicant belongs to an underrepresented demographic or socioeconomic group. Due to many systemic factors not observable to the hiring manager, there may be statistical correlations between Y and Z. Then, using Y to make a hiring decision may have the effect of discriminating against a disadvantaged group even if the decision-maker does not explicitly consider Z.

Our goal is to remove this dependence on Z from Y. We construct some estimate X of Y that is as close to Y as possible while remaining probabilistically independent of Z. Formally, we write the objective

$$\min_{X \perp Z} \mathbb{E}\left[(X-Y)^2\right]. \tag{11}$$

This has some of the flavor of the more general OT problem (7-9), except that we do not know what the marginal distribution of X should be. Chzhen et al. (2020) shows that (11) can be rewritten as

$$\min_f \sum_z P(Z=z)\,\Gamma_z(f), \tag{12}$$

where $\Gamma_z(f)$ is the optimal value of an OT problem with quadratic cost $c(x,y) = (x-y)^2$, the given f as the marginal density of X, and $g(y) = P(Y \in dy \mid Z = z)$ as the marginal density of Y. Since every Γ_z is evaluated at the same f, we ensure that $X \perp Z$. We then add an outer layer of optimization over f in (12) to find the best possible X. Although this problem appears to be very complicated, it actually has a closed-form solution (see Section 5.3) based on the deep analysis by Agueh and Carlier (2011) of "Wasserstein barycenters."

3.3 Image Processing

In the realm of image processing, an important task is color transfer, when an image is modified with the color palette of a different image. Suppose, for example, that a visual designer is given a photograph of an evening scene with a dark blue color scheme, and asked to modify it in Photoshop so that the sky has vivid sunset colors. A second image of a sunset may be provided as a model.

This problem can be approached using the framework of OT (Rabin et al. 2014). An image can be represented as a joint distribution on $\mathscr{Y} \times \mathscr{C}$, where the first two dimensions $\mathscr{S} \subseteq \mathbb{R}^2$ are the spatial coordinates of the image, and the last three dimensions $\mathscr{C} \subseteq \mathbb{R}^3$ represent the RGB color space. Formally, one may define a function mapping $\mathscr{S}$ into $\mathscr{C}$ and let $\mathscr{U} \subseteq \mathscr{Y} \times \mathscr{C}$ be the graph of this function.

The visual designer is given *two* images: one to be modified, and one serving as a color model. The space $\mathscr{Y} = \mathscr{U} \times \mathscr{V}$ in the OT formulation is taken to be the concatenation of the graphs $\mathscr{U}$ and $\mathscr{V}$ of the two images. The goal is to create a *third* image with the same spatial domain as $\mathscr{U}$, but whose color distribution matches that of $\mathscr{V}$. The color map of $\mathscr{U}$ is discarded; only the color distribution of $\mathscr{V}$ is included in the cost function.

Bonneel and Digne (2023) surveys other applications of OT in image processing. The concepts sketched out above for color transfer can also be applied in generative models, where the goal is to create an image in a certain visual style. Essentially, instead of a distribution over a color space, one considers a distribution over a space of stylistic techniques or elements.

3.4 Text Analytics

Yurochkin et al. (2019) studies a text mining problem in which $\mathscr{X}$ and $\mathscr{Y}$ are semantic spaces, and f and g represent the frequency of certain phrases, topics, or other textual elements in two different documents. Each textual element is first converted into a mathematical representation (for example, a word is turned into a word embedding) and the similarity between two elements can be measured using the Euclidean distance between their embeddings (Kusner et al. 2015).

In a manner of speaking, the OT problem shows how one document can be "translated" into another. We are matching one set of textual elements with another in a way that preserves contextual similarity between them as much as possible. The optimal value of the OT problem may be viewed as a measure of similarity between the two documents, but, just as in the image processing example, the optimal solution may also be useful in generative contexts, where the goal may be to change the literary style of a document in some prespecified way while keeping the original subject matter.

3.5 Statistics

Statistics provides a fertile ground for the application of OT, particularly in defining metrics between probability distributions (Klatt et al. 2020). For example, suppose that we have i.i.d. data $Y_1, ..., Y_n$ and we wish to test the null hypothesis that they have density f. Then, we can formulate an OT problem matching the empirical (discrete) distribution of the data to the hypothesized distribution. This literature focuses on the asymptotic behavior of the optimal value of the OT problem, which enables the design of tractable test statistics. See, e.g., Del Barrio et al. (1999) for a treatment of the setting where the hypothesized distribution is normal, and Hallin et al. (2021) for a more general multivariate setting. Other modern statistical problems where OT has been used include uncoupled isotonic regression (Rigollet and Weed 2019), causal inference (Torous, Gunsilius, and Rigollet 2021) and multivariate ranking and quantile estimation (Carlier, Chernozhukov, and Galichon 2014).

4 MONGE & KANTOROVICH FORMULATIONS

OT originated in a restrictive formulation that is attributed to Gaspard Monge (1746–1818), and referred to as the *Monge* problem. To state the Monge problem in its most general form, let's recall the notion of a *push-forward measure*. Let $(\mathscr{X}, d_x)$ and $(\mathscr{Y}, d_y)$ be metric spaces equipped with sigma-algebras Σ_x, Σ_y, respectively, and let $T : \mathscr{X} \to \mathscr{Y}$ be a measurable map. Then, the push-forward of a probability measure μ by T is the probability measure

$$T_{\#}\mu(B) := \mu(T^{-1}(B)) = \mu(\{x : T(x) \in B\}), \quad B \in \Sigma_y. \tag{13}$$

A map $T : \mathscr{X} \to \mathscr{Y}$ such that $T_{\#}\mu = \nu$ is called a *transport map* between probability measures μ and ν. Intuitively, think of the transport map T as "transporting" $x \in \mathscr{X}$ to $T(x) \in \mathscr{Y}$, and hence of $T_{\#}\mu$ as the distribution of $T(X)$, where $X \sim \mu$. An alternate and equivalent way of expressing (13) is that the following "change of variable" formula holds for any continuous real-valued function $\varphi : \mathscr{Y} \to \mathbb{R}$:

$$\int_{\mathscr{Y}} \varphi(y)\, T_{\#}\mu(\mathrm{d}y) = \int \varphi(T(x))\, \mu(\mathrm{d}x), \quad \varphi \in \mathrm{C}(\mathscr{Y}), \tag{14}$$

where $\mathrm{C}(\mathscr{Y})$ is the set of continuous real-valued functions on $\mathscr{Y}$.

Given a *source* probability measure μ and a *target* probability measure ν supported on $\mathscr{X}$ and $\mathscr{Y}$, respectively, the Monge formulation seeks that transport map, among those resulting in a push-forward equal to ν, which incurs least total cost as measured by a function $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R} \bigcup \{\infty\}$:

$$M_c^*(\mu,\nu) := \inf_T \left\{ M_c(T;\mu,\nu) := \int_{\mathscr{X}} c(x,T(x))\mu(\mathrm{d}x) \ : \ T_{\#}\mu = \nu \right\}. \tag{M}$$

In the discrete version of the Monge problem where μ and ν are supported on $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$, respectively, (M) becomes

$$\begin{aligned} \min_T : \quad & \sum_{i=1}^{m} c(x_i, T(x_i)) a_i \\ \text{subject to:} \quad & \sum_{i \in T^{-1}(y_j)} a_i = b_j, \quad j = 1,2,\ldots,n \end{aligned} \tag{15}$$

where $a = (a_1, a_2, \ldots, a_m), a_i := \mu(\{x_i\}), i = 1,2,\ldots,m$; and $b = (b_1, b_2, \ldots, b_n), b_j := \nu(\{y_j\}), j = 1,2,\ldots,n$.

Three observations about the Monge formulation are salient because they have traditionally formed the basis for reservations about (M). First, the optimization problem associated with the Monge formulation may not be feasible. This is easy to see — notice that in the discrete version (15), each i is "assigned" to a single j, implying that if $m < n$ and $\nu(y_j) > 0$ for all $j = 1,2,\ldots,n$ then (15) is necessarily infeasible. Second, the problem in (15) can produce a nonconvex constraint set, that is, if T_1 and T_2 satisfy $T_{1\#}\mu = \nu$ and $T_{2\#}\mu = \nu$, then it is not necessary that $(\gamma T_1 + (1-\gamma)T_2)_{\#}\mu = \nu$ for $\gamma \in (0,1)$. And, third, the Monge formulation violates symmetry. To be precise, suppose $M_c^*(\mu,\nu)$ denotes that optimal value associated with (M), with $M_c^*(\mu,\nu) = \infty$ if (M) is infeasible. Then it can be shown without too much effort that $M_c^*(\mu,\nu) \neq M_c^*(\nu,\mu)$ in general, implying that M_c^* is not symmetric in its arguments. Such asymmetry is seen as undesirable given that $M_c^*(\cdot,\cdot)$ represents a distance between two probability measures. (It can be shown, under mild conditions on the cost function c, that M_c^* satisfies the other two axioms of non-negativity and triangle-inequality required by a metric.)

The Monge problem is often described as the OT formulation that "does not allow mass splitting." This is apt because the map T implicitly stipulates that *all mass* at a support point x_i of the source distribution is assigned to a *single* support point y_j of the target distribution. Remarkably, relaxing this restriction by allowing "probabilistic mass splitting," addresses all three objections raised in the context of the Monge problem. To see this, consider the discrete setting of (15). And instead of searching over the space of transport maps T, suppose we relax to allow *transport plan* matrices, that is, $m \times n$ matrices Q of probabilities such that $Q\,\mathbb{1}_m = a$ and $Q^T\,\mathbb{1}_n = b$:

$$\begin{aligned} \min_Q : \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c(x_i, y_j)) Q(i,j) \\ \text{subject to:} \quad & Q\,\mathbb{1}_m = a; \quad Q^T\,\mathbb{1}_n = b; \\ & Q(i,j) \in [0,1] \text{ for } (i,j) \in \{1,2,\ldots,m\} \times \{1,2,\ldots,n\}, \end{aligned} \tag{16}$$

where we recall that $a_i = \mu(\{x_i\}), i = 1,2,\ldots,m$ and $b_j = \nu(\{y_j\}), j = 1,2,\ldots,n$. It can be shown that the problem in (16) is always feasible. Furthermore, the feasible region is a convex polytope, that is, the convex hull of a finite set of matrices (Brualdi 2006), and it's easily seen that the optimal value $K_c^*(\mu,\nu)$ of (16) is symmetric in that $K_c^*(\mu,\nu) = K_c^*(\nu,\mu)$.

The general version of the Kantorovich problem, where the measures μ,ν are not necessarily discrete, should be somewhat evident from (16), although some care is needed in precisely formulating the problem. The *transport plan* between probability measures μ,ν on $(\mathscr{X},\Sigma_x)$ and $(\mathscr{Y},\Sigma_y)$, respectively, is a probability measure π on $\mathscr{X}\times\mathscr{Y}$ having marginals μ and ν; and the set of all such transport plans, denoted $\Pi(\mu,\nu)$, is

$$\Pi(\mu,\nu) := \left\{\pi \in \mathscr{P}(\mathscr{X}\times\mathscr{Y}) : \Pi_{\#,\mathscr{X}}\pi = \mu; \Pi_{\#,\mathscr{Y}}\pi = \nu\right\}, \tag{17}$$

where $\Pi_{\mathscr{X}} : (x,y) \in \mathscr{X}\times\mathscr{Y} \mapsto x$ and $\Pi_{\mathscr{Y}} : (x,y) \in \mathscr{X}\times\mathscr{Y} \mapsto y$ are projection maps, and $\Pi_{\#,\mathscr{X}}, \Pi_{\#,\mathscr{Y}}$ are corresponding push-forwards of π. Analogous to (14), the set $\Pi(\mu,\nu)$ appearing in (17) can also be characterized as the set of probability measures π on $\mathscr{X}\times\mathscr{Y}$ such that

$$\int_{\mathscr{X}\times\mathscr{Y}} (\varphi_1(x)+\varphi_2(y))\,\pi(\mathrm{d}(x,y)) = \int_{\mathscr{X}} \varphi_1(x)\,\mu(\mathrm{d}x) + \int_{\mathscr{Y}} \varphi_2(y)\,\nu(\mathrm{d}y), \quad (\varphi_1,\varphi_2) \in C(\mathscr{X})\times C(\mathscr{Y}).$$

Then, the Kantorovich problem searches in the space $\Pi(\mu,\nu)$ to identify a transport plan which incurs minimum cost under c:

$$K_c^*(\mu,\nu) = \inf_{\pi}\left\{K_c(\pi;\mu,\nu) := \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\pi(\mathrm{d}(x,y)), \quad \pi\in\Pi(\mu,\nu)\right\} \tag{K}$$

The problem in (K) is a relaxation of (M) in a sense to be discussed, but it also has much more structure and is well-posed. Indeed, the set $\Pi(\mu,\nu)$ can be shown to be convex. It is also always non-empty because the measure $\mu\otimes\nu \in \Pi(\mu,\nu)$. Furthermore, it can be shown that the problem in (K) is well-defined, and that $K_c : \pi \mapsto \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\pi(\mathrm{d}(x,y))$ attains its minimum on $\Pi(\mu,\nu)$. (That K_c attains its minimum on $\Pi(\mu,\nu)$ follows upon showing that $\Pi(\mu,\nu)$ is weak* compact, and that K_c is weak* continuous on $\Pi(\mu,\nu)$.)

Theorem 1 The problem in (K) admits a solution, that is, the set $\Pi^*(\mu,\nu)$ where the infimum in (K) is attained, is non-empty. □

How do transport maps relate to transport plans? Suppose $T : \mathscr{X}\to\mathscr{Y}$ is a transport map. Then, there corresponds a transport plan π_T defined through the push-forward of μ by $(\mathrm{id},T) : \mathscr{X}\to\mathscr{X}\times\mathscr{Y}$, that is, $\pi_T := (\mathrm{id},T)_{\#}\mu$. Then, as usual, π_T satifies

$$\int_{\mathscr{X}\times\mathscr{Y}} \varphi(x,y)\,\pi_T(\mathrm{d}(x,y)) = \int_{\mathscr{X}} \varphi(x,T(x))\,\mu(\mathrm{d}(x)), \quad \varphi\in \mathrm{C}(\mathscr{X}\times\mathscr{Y}). \tag{18}$$

Also notice that by the definition of π_T, we have that

$$M_c(T;\mu,\nu) := \int_{\mathscr{X}} c(x,T(x))\,\mu(\mathrm{d}(x)) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\pi_T(\mathrm{d}(x,y)),$$

implying that the minimum associated with the Kantorovich problem (K) is necessarily smaller than the infimum associated with the Monge problem (M), that is, $K_c^*(\mu,\nu) \le M_c^*(\mu,\nu)$.

Another symptom of the structure inherent to (K) is its connection with the useful p-Wasserstein metric on $\mathscr{P}(\mathscr{X})\times\mathscr{P}(\mathscr{Y})$. To see this, suppose $\mathscr{X}=\mathscr{Y}$ and d is a metric on $\mathscr{X}$, that is, $\forall x,y\in\mathscr{X}, d(x,y) = d(y,x) \ge 0$, $d(x,y)=0$ if and only if $x=y$, and $\forall x,y,z\in\mathscr{X}, d(x,y)+d(y,z)\ge d(x,z)$. Assume that the resulting metric space $(\mathscr{X},d)$ is a Polish space, that is, $(\mathscr{X},d)$ is separable and completely metrizable. Then, assuming the probability measures μ,ν on $\mathscr{P}(\mathscr{X})$ and $\mathscr{P}(\mathscr{X})$, respectively, have finite p-th moments, the p-Wasserstein distance between μ and ν is given by

$$W_p(\mu,\nu) = \left(\inf_{\pi}\int_{\mathscr{X}\times\mathscr{Y}} d^p(x,y)\,\pi(\mathrm{d}(x,y))\right)^{1/p}, \quad p\in[1,\infty). \tag{19}$$

Comparing (19) and (K), it should be clear that $W_p(\mu,\nu) = \left(K_{d^p}^*(\mu,\nu)\right)^{1/p}$.

5 EXAMPLE TRANSPORT MAPS

The quantile transportation map is useful in solidifying the simulationist's intuition of optimal transport. As we discuss, this map is optimal to (M) under certain circumstances.

5.1 Quantile Transport Map

Suppose μ and ν are probability measures on $\mathbb{R}$, with μ being *non-atomic* (see Section 2.) Define the x-quantile of μ:

$$T_q(x;\mu,\nu) = T_q(x) = \min\{t : \nu((-\infty,t]) \geq \mu((-\infty,t])\}. \tag{20}$$

It can be shown that the push-forward of μ by T_q is ν, that is, $T_{q,\#}\mu = \nu$, and that T_q is the unique monotone transport map with this property. It turns out that for a wide family of transportation costs, the optimal transport map takes the form in (20).

5.2 Knothe's Conditional Quantile Transport Map

Knothe's transport builds on the quantile map to construct a powerful transport map in d-dimensions. For simplicity, suppose μ and ν are probability measures on $\mathbb{R}^2$, and further suppose μ is absolutely continuous. Let $\mu_1 = \Pi_{\mathscr{X},\#}\mu, \nu_1 = \Pi_{\mathscr{Y},\#}\nu$ be the push-forwards of μ and ν by the projections $\Pi_{\mathscr{X}}(x,y) = x$ and $\Pi_{\mathscr{Y}}(x,y) = y$, respectively; and let μ^{x_1}, ν^{y_1} be the corresponding *conditional measures* so that

$$\mu = \mu_1 \otimes \mu^{x_1}; \quad \nu = \nu_1 \otimes \nu^{y_1}. \tag{21}$$

(In other words, if $(X_1,X_2) \sim \mu$, then $X_1 \sim \mu_1$ is the marginal measure corresponding to X_1, and $X_2|X_1 = x_1 \sim \mu^{x_1}$ is the measure corresponding to the conditional random variable $X_2|X_1 = x_1$.) With this notation, Knothe's transport map $T_K(x) = (T_{K,1}(x_1), T_{K_2}(x_1,x_2))$ can be constructed "coordinatewise" from the quantile transport map:

$$\begin{aligned} T_{K,1}(x_1) &= T_q(x_1;\mu_1,\nu_1); \\ T_{K,2}(x_1,x_2) &= T_q(x_2;\mu_2,\nu^{T_{K,1}(x_1)}). \end{aligned} \tag{22}$$

Notice now that for any $\varphi \in \mathrm{C}(\mathscr{Y})$,

$$\begin{aligned} \int_{\mathbb{R}^2} \varphi(T_K(x))\,\mu(\mathrm{d}x) &= \int_{\mathbb{R}^2} \varphi(y)\,T_{K,\#}\mu(\mathrm{d}y) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(y)\,\nu^{y_1}(\mathrm{d}y_2)\,\nu_1(\mathrm{d}y_1) \\ &= \int_{\mathbb{R}^2} \varphi(y)\,\nu(\mathrm{d}(y)), \end{aligned} \tag{23}$$

where the first equality comes from the definition of the push-forward, the second equality from (22) and (20), and the last equality from (21). The equality in (23) justifies the Knothe's transport from μ to ν, and extending it to $d \geq 2$ dimensions should be clear from the construction in (22).

5.3 When is a Quantile Transport Map Optimal?

When $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}$ and the cost function satisfies the so-called *twist condition* on the cost function c, namely, that the mapping $x \mapsto \nabla_y c(x,y)$ is injective (Gangbo and McCann 1996; Levin 1999), the continuous OT problem can be solved explicitly. A consequence of the twist condition in this setting is that the optimal Monge map T^* is the unique monotonic map (Maggi 2023) satisfying,

$$T^*(x) = T_q(x;\mu,\nu), \tag{24}$$

where T_q is defined in (20). In other words, the problem is solved by matching the quantiles of both distributions, so that the x-quantile of μ is transformed into the x-quantile of ν.

This result is used to solve the algorithmic fairness problem in Section 3.2. Returning to (12), we see that each individual OT problem Γ_z can be solved by (24), which yields the optimal transformation

$$X = \sum_z P(Z=z)\, G_z\left(G_Z^{-1}(Y)\right),$$

where $G_z(y) = P(Y \le y \mid Z = z)$. In words, we observe (Y,Z) and calculate the percentile of Y within its subgroup. We then "blind" it by turning it into a weighted average of those same percentiles across all subgroups.

6 DUALITY

Like the classical linear programming problem (Luenberger and Ye 1984), the Kantorovich problem (K) can be paired with a *dual problem* that is sometimes computationally easier, and sheds insight on OT's structure, while also helping to establish the existence of transport maps. Following the broad lines used in the classical linear programming construction, we can arrive at the following dual problem to (K):

$$\begin{aligned} \sup \quad & \int_{\mathscr{X}} \varphi(x)\,\mu(\mathrm{d}x) + \int_{\mathscr{Y}} \psi(y)\,\nu(\mathrm{d}y) \\ \text{s.t.} \quad & c(x,y) \ge \varphi(x) + \psi(y), \quad \forall (x,y) \in \mathscr{X} \times \mathscr{Y}, \\ & (\varphi,\psi) \in C(\mathscr{X}) \times C(\mathscr{Y}). \end{aligned} \tag{25}$$

The functions φ, ψ are analogues of the Lagrange variables in classical linear programming and have clear physical interpretations. Suppose a company wishes to transport goods "spatially distributed" in $\mathscr{X}$ as μ, as cheaply as possible, and to locations "spatially distributed" in $\mathscr{Y}$ as ν. Assuming the cost of transporting from x to y is $c(x,y)$, the problem (K) describes the problem that the company would have to solve to minimize the cost of transport. Now, imagine that the job is outsourced to an operator who would first collect goods from each $x \in \mathscr{X}$, and then transport the collected goods to each $y \in \mathscr{Y}$, charging a price for each step. Specifically, suppose $\varphi(x)$ represents the price charged by the vendor for collecting from x, and $\nu(y)$ the price of transporting to y. The vendor attempts to choose φ, ψ so that the total price $\int_{\mathscr{X}} \varphi(x) + \int_{\mathscr{Y}} \psi(y)$ is maximized, but while making sure that $\varphi(x) + \psi(y) \le c(x,y)$, for otherwise the job will not be outsourced.

Given the formulation in (25), we can argue that *weak duality* holds, that is, the supremum in (25) lower bounds the infimum in (K). Indeed for any π that is *primal feasible*, that is, $\pi \in \Pi(\mu,\nu)$, and for dual feasible φ, ψ, that is, $c(x,y) \ge \varphi(x) + \psi(y)$, we can write

$$\begin{aligned} \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\pi(\mathrm{d}(x,y)) &\ge \int_{\mathscr{X}\times\mathscr{Y}} \varphi(x)\,\pi(\mathrm{d}(x,y)) + \int_{\mathscr{X}\times\mathscr{Y}} \psi(y)\,\pi(\mathrm{d}(x,y)) \\ &= \int_{\mathscr{X}} \varphi(x)\,\mu(\mathrm{d}x) + \int_{\mathscr{Y}} \psi(y)\,\nu(\mathrm{d}y), \end{aligned} \tag{26}$$

implying that the supremum in (25) has to necessarily be a lower bound to the infimum in (K).

We say *strong duality* holds if the supremum in (25) equals the infimum in (K). While we do not go into details here, the Kantorovich dual formula can be used to demonstrate that strong duality indeed holds, and there exists a primal feasible π^* and a dual feasible (φ^*, ψ^*) such that

$$\int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\pi^*(\mathrm{d}(x,y)) = \int_{\mathscr{X}} \varphi^*(x)\,\mu(\mathrm{d}x) + \int_{\mathscr{Y}} \psi^*(y)\,\nu(\mathrm{d}y). \tag{27}$$

In such a case, $\pi^* \in \Pi(\mu,\nu)$ is necessarily an optimal transport plan, that is, it solves (K).

6.1 Kantorovich-Monge Equivalence

We have seen that the problem in (M) may be infeasible, that is, there may exist no transport map such that $T_{\#}\mu = \nu$. This was easily evident in the discrete case (15) where if $m < n$, the problem became infeasible. It turns out that such existence of atoms presents the only serious impediment to feasibility even for the general Monge problem. In fact, as the next theorem states, the infimum in (M) and the infimum in (K) coincide as long as μ is atomless. (See Definition 1 for a non-atomic measure.) Also, see (Pratelli 2007) for alternative conditions.

Theorem 2 Suppose μ is non-atomic, and let (id, T) be the map $(x,y) \mapsto (x, T(x))$ for $(x,y) \in \mathscr{X} \times \mathscr{Y}$. Then the set $\{\pi_T = (\mathrm{id}, T)_{\#}\mu : T_{\#}\mu = \nu\}$ of push-forward measures is weak* dense in $\Pi(\mu, \nu)$, and

$$\inf_T \left\{ \int_{\mathscr{X}} c(x, T(x)) \ : \ T_{\#}\mu = \nu \right\} = \inf_{\pi} \left\{ \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\, \pi(\mathrm{d}(x,y)) \right\}.$$

□

The equivalence in Theorem 2 does not imply the existence of an optimal transport map. To understand when such a map might exist, notice that since $c(x,y) \geq \varphi^*(x) + \psi^*(y)$ for all $(x,y) \in \mathscr{X} \times \mathscr{Y}$, the equality in (27) implies that π^*-almost surely,

$$\begin{aligned} \varphi^*(x) &= c(x,y) - \psi^*(y), \quad (x,y) \in \mathscr{X} \times \mathscr{Y} \\ &= \inf_{z \in \mathscr{Y}} \{c(x,z) - \psi^*(z)\}. \end{aligned} \tag{28}$$

Thus, if for μ-almost surely $x \in \mathscr{X}$, there exists a single y that attains the infimum in (28), this will prove the existence of a Monge transport map corresponding to the optimal transport plan π^*.

The expression in (28) can be used to show that φ^* is Lipschitz on $\mathscr{X}$. Then from Rademacher's theorem (Mattila 1995, p. 101-102), we know that φ^* is almost-everywhere differentiable on $\mathscr{X}$. Now if we further assume that μ is absolutely continuous on an open ball B of $\mathbb{R}^d$, then φ^* is μ-almost surely differentiable on B. Letting S denote the set where φ^* is discontinuous, we can write for $x \in B \setminus \{S \bigcup \partial B\}$,

$$\varphi^*(x+h) = \varphi^*(x) + \nabla\varphi^*(x)^T h + o(h). \tag{29}$$

if the cost $c(\cdot,\cdot)$ has the special structure $c(x,y) \equiv c_0(x-y)$ for some $c_0 : \mathbb{R}^d \to \mathbb{R}$, we can write for $x \in B \setminus \{S \bigcap \partial B\}$,

$$\begin{aligned} \varphi^*(x+h) &= c(x+h,y) - \psi^*(y) \\ &= c_0(x-y) + \nabla c_0(x-y)^T h + o(h) - \psi^*(y) \\ &= \varphi^*(x) + \nabla c_0(x-y)^T h. \end{aligned} \tag{30}$$

From (29) and (30), we see that

$$\nabla\varphi^*(x) = \nabla c_0(x-y), \quad x \in B \setminus \{S \bigcup \partial B\}. \tag{31}$$

If we further assume that c_0 is strictly convex, then the y satisfying (31) is unique and takes the form of a transport map

$$T(x) = x - \nabla c_0^{-1}(\nabla\varphi^*(x)). \tag{32}$$

Let's gather the assumptions and arguments leading to (32) as a theorem. Also see (Gangbo and McCann 1996; Levin 1999) for the so-called *twist condition*, leading to a variation.

Theorem 3 Suppose $c_0 : \mathbb{R}^d \to \mathbb{R}$ is a C^1 and strictly convex function. Let $B \subset \mathbb{R}^d$ be a ball, and μ, ν probability measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively, such that $\text{supp}(\mu) \subseteq B$, $\text{supp}(\nu) \subseteq B$, and μ is absolutely continuous on B. Then, the Monge problem given by

$$\inf_T \left\{ \int_{\mathscr{X}} c_0(x - T(x))\, \mu(\mathrm{d}x), \quad T_{\#}\mu = \nu \right\}, \tag{33}$$

attains its minimum at

$$T^*(x) = x - \nabla c_0^{-1}(\nabla \varphi^*(x)),$$

where φ^* is a conjugate c_0-concave function. Moreover, the transport plan given by the push-forward $T^*_{\#}\mu$ uniquely solves the Kantorovich problem

$$\inf_{\pi} \left\{ \int_{\mathscr{X} \times \mathscr{Y}} c_0(x - y)\, \pi(\mathrm{d}(x,y)), \quad \pi \in \Pi(\mu, \nu) \right\}, \tag{34}$$

that is, if R is a measure that solves (34), then $R = T^*_{\#}\mu$, μ-almost surely. □

Now let's consider the special and frequently encountered context of a quadratic cost $c_0(x) = \frac{1}{2}\|x\|^2$, and notice from (28) that

$$\varphi^*(x) = \inf_{y \in \mathscr{Y}} \left\{ \frac{1}{2}\|x - y\|^2 - \psi^*(y) \right\} \tag{35}$$

implying that

$$\frac{1}{2}\|x\|^2 - \varphi^*(x) = \sup_{y \in \mathscr{Y}} \left\{ x^T y - \left(\frac{1}{2}\|y\|^2 - \psi^*(y) \right) \right\}. \tag{36}$$

Analogously, we also have

$$\frac{1}{2}\|y\|^2 - \psi^*(y) = \sup_{x \in \mathscr{X}} \left\{ y^T x - \left(\frac{1}{2}\|x\|^2 - \varphi^*(x) \right) \right\}. \tag{37}$$

Notice from (36) and (37) that the functions $u(x) := \frac{1}{2}\|x\|^2 - \varphi^*(x)$ and $v(x) := \frac{1}{2}\|y\|^2 - \psi^*(y)$ are convex conjugates (Boyd and Vandenberghe 2004) of each other. Furthermore, since $c_0(x) = \frac{1}{2}\|x\|^2$, we see from Theorem 3 that the optimal transport plan has the form

$$T(x) = x - \nabla c_0^{-1}(\nabla \psi^*(x)) = \nabla u(x), \tag{38}$$

where $u(x) := \frac{1}{2}\|x\|^2 - \varphi^*(x)$ is convex since it is the convex conjugate of v.) The map in (38) is called Brenier's map, named after the now famous Brenier's theorem.

Theorem 4 Consider the Monge problem in (34) with the quadratic cost $c_0(x) = \frac{1}{2}\|x\|^2$. Then, the map

$$T^*(x) = \nabla u(x), \quad x \in \mathscr{X}$$

solves (34) uniquely (up to μ negligible sets), where $u : \mathscr{X} \to \mathbb{R}$ is a convex function. Moreover, the push-forward $T^*_{\#}\mu$ of μ by T^* uniquely (up to μ negligible sets) solves the corresponding Kantorovich problem (34) with quadratic cost $c_0(x) = \frac{1}{2}\|x\|^2$. □

Notice that Theorem 3 establishes equivalence between the Monge and Kantorovich problems when, among other assumptions, the cost function $c(\cdot,\cdot)$ takes the particular form $c(x,y) \equiv c_0(x-y))$. If such a structure on cost does not hold, then the equivalence between the Monge and Kantorovich formulations may not hold. This issue might be interesting because a number of applications in machine learning and statistics explicitly seek a transport map, and in such cases, solving the Kantorovich formulation may make sense only when the equivalence between the Monge and Kantorovich formulations is guaranteed.

7 SEMIDISCRETE OPTIMAL TRANSPORT

In this section, we show how (4)-(6) can be solved efficiently, using classical methods for simulation optimization. The key observation is that the semidiscrete problem can be viewed as a kind of linear program, much like the classical transportation LP (1)-(3). Since the integral is a linear operator, both (4) and (5)-(6) are *linear* in the decision variables $h(m,y)$. It should therefore be intuitive (and Theorem 1.3 in Villani 2021 rigorously proves) that (4)-(6) also has a *Kantorovich dual*. This problem is written as

$$\sup_{v,w} \sum_{m=1}^{M} p_m v_m + \int_{\mathscr{Y}} w(y)\, g(y)\, dy \tag{39}$$

subject to

$$v_m + w(y) \leq c(m,y) \quad m = 1,...,M,\ y \in \mathscr{Y}. \tag{40}$$

The decision variables v and w are functions, but since one of our distributions is supported only on $\{1,...,M\}$, we may write $v = (v_1,...,v_M)$. Since the primal problem had only equality constraints, there are no sign restrictions on v or w.

The infinite-dimensional system of inequalities (40) can be handled by taking

$$w(y) = \min_m c(m,y) - v_m.$$

Consequently, (39)-(40) admits the finite-dimensional (!) reformulation

$$\sup_{v \in \mathbb{R}^M} \mathbb{E}\left(\min_m c(m,Y) - v_m\right) + p^\top v. \tag{41}$$

For any y and any m, the function $v_m \mapsto c(m,y) - v_m$ is linear. Therefore, the minimum of these functions over $m \in \{1,...,M\}$ is concave. Because concavity is preserved under expectations, (41) is an unconstrained, finite-dimensional concave maximization problem.

This is a problem that the simulation community is in a prime position to solve. Let $H(v,y) = \min_m c(m,y) - v_m$. Then, it can be shown, under some mild conditions on the distribution of Y, that $\nabla_v \mathbb{E}(H(v,Y)) = \mathbb{E}(\nabla_v H(v,Y))$, an example of what the simulation literature calls "infinitesimal perturbation analysis" (Fu 2006; Kim 2006). It is also easy to see that

$$(\nabla_v H(v,y))_m = -1_{\{m = \arg\min_{m'} c(m',y) - v_{m'}\}}$$

at any v that yields a unique argmin on the right-hand side. Consequently, the optimal solution v^* of (41) is the solution to the system of equations

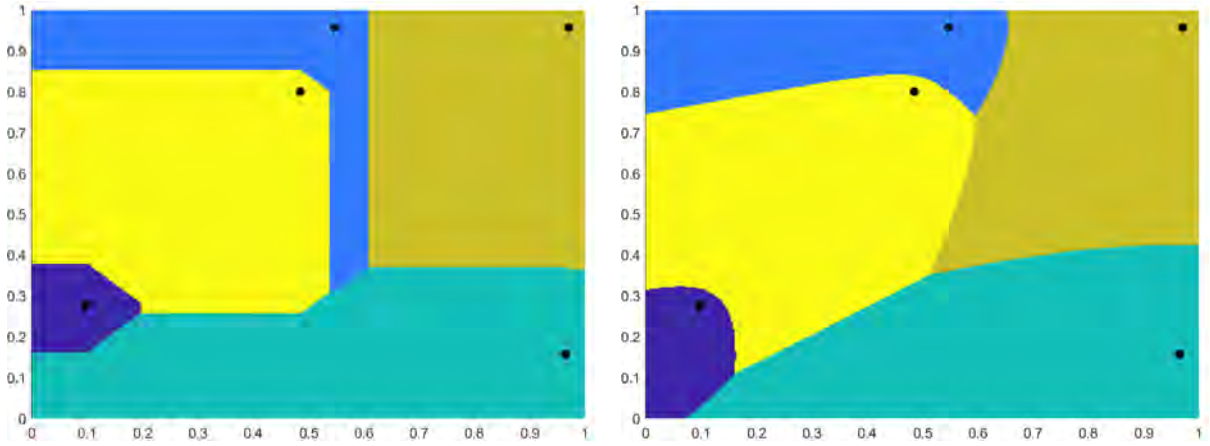
$$P\left(m = \arg\min_{m'} c\left(m',Y\right) - v_{m'}\right) = p_m. \tag{42}$$

The probability in (42) is an integral over the marginal density g of Y only. Essentially, v^* is a vector of bonuses and penalties that are used to adjust the costs $c(m,Y)$ such that the mth cost has precisely probability p_m of being the smallest.

The system (42) is none other than a stochastic root-finding problem (Pasupathy and Kim 2011). To solve it, we actually do not need to know the density g. We only need to be able to simulate independent replications $Y^1, Y^2, ...$ from that density. Then, we can compute v^* iteratively using the stochastic approximation algorithm

$$v_m^{n+1} = v_m^n + \mu_n\left(-1_{\left\{m = \arg\min_{m'} c(m',Y^{n+1}) - v_{m'}^n\right\}} + p_m\right),$$

where $\{\mu_n\}_{n=0}^{\infty}$ is a stepsize sequence satisfying the standard conditions $\sum_n \alpha_n = \infty$, $\sum_n \alpha_n^2 < \infty$. The initial estimate v^0 can be chosen arbitrarily. Since the stochastic gradient is bounded, the convergence $v^n \to v^*$



(a) $c(m,y) = \|y - y_m\|_1$. (b) $c(m,y) = \|y - y_m\|_2$.

Figure 1: Optimal partitions for two distance metrics. The pre-specified areas p_m for each facility, starting in the lower-left corner and moving clockwise, are 0.0439, 0.2970, 0.1305, 0.2452, and 0.2834.

easily follows from classical stochastic approximation theory (Kushner and Yin 2003). To our knowledge, Genevay et al. (2016) was the first to solve semidiscrete OT in this way, though there are some alternate approaches based on mathematical programming (Carlsson et al. 2016; Hartmann and Schuhmacher 2020).

We now have an optimal dual solution v^*. It is relatively straightforward to show (really it is just a consequence of strong duality) that

$$X = \arg\min_m c(m,Y) - v^*_m, \tag{43}$$

with ties broken arbitrarily, yields an optimal solution to the primal problem (4)-(6); in other words, the joint density of (X,Y) achieves the minimum expected cost when X is chosen in this way. Thus, the geographical partitioning problem of Section 3.1 is solved: we simply let

$$A_m = \left\{ y \in \mathscr{Y} \,:\, c(m,y) - v^*_m \leq \min_{m' \neq m} c\left(m', y\right) - v^*_{m'} \right\}. \tag{44}$$

In fact, partitions of the form (44) have a long history in computational geometry, where they are known as "additively weighted Voronoi diagrams" (Aurenhammer 1991). The framework described here allows us to compute such diagrams for a wide variety of cost functions; Figure 1 shows an illustration for Manhattan and Euclidean distances in an instance with five facilities where $\mathscr{Y} = [0,1]^2$ and the facility locations y_m are represented by black dots.

8 CONCLUSION

Optimal Transport provides a versatile and powerful framework for a wide range of problems that can be posed as the question of optimally transporting resources between distributions. Understanding OT is crucial for researchers and practitioners seeking solutions to complex problems involving resource allocation, distribution matching, and logistics. This tutorial has explored OT's mathematical foundations, key formulations, diverse applications across different domains, and some fundamental ideas that could form the basis of numerical algorithms. This tutorial does not treat computation, but numerical methods for solving the many flavors of OT is a fertile area of ongoing research.

ACKNOWLEDGMENTS

The authors thank the track coordinators of the Winter Simulation Conference for providing an opportunity to write this tutorial on optimal transport. Raghu Pasupathy thanks the Office of Naval Research for support provided through the grants N000141712295 and 13000991.

REFERENCES

Agueh, M. and G. Carlier. 2011. "Barycenters in the Wasserstein space". *SIAM Journal on Mathematical Analysis* 43(2):904–924.

Aurenhammer, F. 1991. "Voronoi diagrams – a survey of a fundamental geometric data structure". *ACM Computing Surveys* 23(3):345–405.

Blanchet, J. and A. Shapiro. 2023. "Statistical limit theorems in distributionally robust optimization". In *Proceedings of the 2023 Winter Simulation Conference*, 31–45. IEEE.

Bonneel, N. and J. Digne. 2023. "A survey of optimal transport for computer graphics and computer vision". *Computer Graphics Forum* 42(2):439–460.

Boyd, S. P. and L. Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Brualdi, R. A. 2006. *Combinatorial matrix classes*, Volume 13. Cambridge University Press.

Carlier, G. 2012. "Optimal transportation and economic applications". *Lecture Notes* 18.

Carlier, G., V. Chernozhukov, and A. Galichon. 2014. "Vector quantile regression: an optimal transport approach". *arXiv preprint arXiv:1406.4643*.

Carlier, G. and A. Lachapelle. 2009. "A planning problem combining optimal control and optimal transport". *submitted, available on http://hal. archives-ouvertes. fr/hal-00432785/fr*.

Carlsson, J. G., E. Carlsson, and R. Devulapalli. 2016. "Shadow prices in territory division". *Networks and Spatial Economics* 16(3):893–931.

Chzhen, E., C. Denis, M. Hebiri, L. Oneto and M. Pontil. 2020. "Fair regression with Wasserstein barycenters". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Volume 33, 7321–7331.

Del Barrio, E., J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. 1999. "Tests of goodness of fit based on the L2-Wasserstein distance". *Annals of Statistics* 27(4):1230–1239.

Ford, L. R. and D. R. Fulkerson. 1956. "Solving the transportation problem". *Management Science* 3(1):24–32.

Fu, M. C. 2006. "Gradient estimation". In *Handbooks in Operations Research and Management Science, vol. 13: Simulation*, edited by S. G. Henderson and B. L. Nelson, 575–616. North-Holland Publishing, Amsterdam.

Gangbo, W. and R. J. McCann. 1996. "The geometry of optimal transportation". *Acta Mathematica* 177:113–161.

Genevay, A., M. Cuturi, G. Peyré, and F. Bach. 2016. "Stochastic optimization for large-scale optimal transport". In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Volume 29, 3440–3448: Curran Associates, Inc.

Hallin, M., G. Mordant, and J. Segers. 2021. "Multivariate goodness-of-fit tests based on Wasserstein distance". *Electronic Journal of Statistics* 15(1):1328–1271.

Hartmann, V. and D. Schuhmacher. 2020. "Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case". *Mathematical Methods of Operations Research* 92(1):133–163.

Kim, S. 2006. "Gradient-based simulation optimization". In *Proceedings of the 2006 Winter Simulation Conference*, 159–167.

Klatt, M., C. Tameling, and A. Munk. 2020. "Empirical regularized optimal transport: Statistical theory and applications". *SIAM Journal on Mathematics of Data Science* 2(2):419–443.

Kushner, H. J. and G. Yin. 2003. *Stochastic approximation and recursive algorithms and applications (2nd ed.)*. Springer.

Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. 2015. "From Word Embeddings To Document Distances". In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, Volume 37 of *Proceedings of Machine Learning Research*, 957–966: PMLR.

Levin, V. 1999. "Abstract cyclical monotonicity and Monge solutions for the general Monge-Kantorovich problem". *Set-Valued Analysis* 7(1):7–32.

Luenberger, D. G. and Y. Ye. 1984. *Linear and nonlinear programming*, Volume 2. Springer.

Maggi, F. 2023. *Optimal mass transport on Euclidean spaces*. Cambridge University Press.

Mattila, P. 1995. *Geometry of sets and measures in Euclidean spaces : fractals and rectifiability*. Cambridge studies in advanced mathematics ; 44. Cambridge [England] ;: Cambridge University Press.

Pasupathy, R. and S. Kim. 2011. "The stochastic root-finding problem: Overview, solutions, and open questions". *ACM Transactions on Modeling and Computer Simulation* 21(3):19:1–19:23.

Peyré, G., M. Cuturi, *et al*. 2019. "Computational optimal transport: With applications to data science". *Foundations and Trends® in Machine Learning* 11(5-6):355–607.

Pratelli, A. 2007. "On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation". *Annales de l'Institut Henri Poincare* B43(1):1–13.

Rabin, J., S. Ferradans, and N. Papadakis. 2014. "Adaptive color transfer with relaxed optimal transport". In *Proceedings of the 2014 IEEE International Conference on Image Processing*, 4852–4856. IEEE.

Rigollet, P. and J. Weed. 2019. "Uncoupled isotonic regression via minimum Wasserstein deconvolution". *Information and Inference: A Journal of the IMA* 8(4):691–717.

Torous, W., F. Gunsilius, and P. Rigollet. 2021. "An optimal transport approach to causal inference". *arXiv preprint arXiv:2108.05858*.

Villani, C. 2021. *Topics in optimal transportation*, Volume 58. American Mathematical Society.

Yurochkin, M., S. Claici, E. Chien, F. Mirzazadeh and J. M. Solomon. 2019. "Hierarchical optimal transport for document representation". In *Advances in neural information processing systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Volume 32, 1601–1611.

Zhu, Y. and I. O. Ryzhov. 2024. "Optimal data-driven hiring with equity for underrepresented groups". *Production and Operations Management (to appear)*.

AUTHOR BIOGRAPHIES

ILYA O. RYZHOV is a Professor of Operations Management at the Robert H. Smith School of Business, University of Maryland. His research interests include stochastic optimization, statistics, and applications in public sector operations research. He serves as Associate Editor at *Operations Research* and *INFORMS Journal on Computing*. He won I-SIM's Outstanding Paper Award in 2017, and was recognized by WSC's Best Theoretical Paper award competition on three occasions (winner in 2012, finalist in 2009 and 2016). His email address is iryzhov@rhsmith.umd.edu and his website is https://sites.google.com/umd.edu/iryzhov.

RAGHU PASUPATHY is Professor of Statistics at Purdue University. His current research interests lie broadly in stochastic optimization, uncertainty quantification, and simulation methodology. He has been actively involved with the Winter Simulation Conference for the past 20 years. Raghu Pasupathy's email address is pasupath@purdue.edu, and his web page https://web.ics.purdue.edu/~pasupath contains links to papers, software codes, and other material.

HARSHA HONNAPPA is an Associate Professor of Industrial Engineering at Purdue University. His research interests as an applied probabilist encompass stochastic modeling, optimization and control, with applications to machine learning, simulation and statistical inference. His research is supported by the National Science Foundation, including an NSF CAREER award, the Department of Defense, and the Purdue Research Foundation. He is an editorial board member at Operations Research, Operations Research Letters and Queueing Systems journals. His email address is honnappa@purdue.edu and his website is https://engineering.purdue.edu/SSL.