

1

Introduction

The basic problem in optimal transport (hereafter, OT) can be best exemplified by the problem of assigning workers to jobs: given the distribution of a population of workers with heterogeneous skills, and given the distribution of jobs with heterogeneous characteristics, how should one assign workers to firms in order to maximize the total economic output? The economic output, of course, will depend on the complementarities between workers' skills and job characteristics; some assignments generate higher total output than others. This problem and its variants are known under several names: mass transportation, optimal assignment, matching with transferable utility, optimal coupling, Monge–Kantorovich, and Hitchcock being the more common. Of course, the multiplicity of names reflects small variations in the formulation of the problem, but also the stunning diversity of applications this theory has found.

1.1 A NUMBER OF ECONOMIC APPLICATIONS

To describe OT as a general framework for labor market assignment problems, as we just did, is somewhat overrestrictive. While labor economics is certainly one use of the theory, which we will discuss below, there is much more to it. Indeed, an impressive number of seemingly unrelated problems in economics have the structure of an OT problem. Here are some examples, without any attempt at exhaustivity.

– *Matching models* are models in which two populations, such as men and women on the marriage market, workers and machines etc., must be assigned into pairs. Each pair generates a surplus which depends on the characteristics of both partners. One question deals with the characterization of the *optimal assignment*: what is the assignment a central planner would choose in order to maximize the total utility surplus? Another question deals with the *equilibrium assignment*: letting partners match in a decentralized manner, what are the equilibrium matching patterns and transfers? The Monge–Kantorovich theorem, introduced in chapter 2, implies that the answers to these two questions coincide: any optimal solution chosen by the central

planner can be obtained at equilibrium and conversely, any equilibrium assignment is also optimal. Thus OT theory will provide a powerful welfare theorem in matching models with transferable utility.

– *Models of differentiated demand* are models where consumers who choose a differentiated good (say, a house) have unobserved variations in preferences. These types of models are often called *hedonic models* when the measure of the quality of the good is continuous, and *discrete choice models* when it is discrete. A central problem in these models is the identification of preferences. By imposing assumptions on the distribution of the variation in preferences, one is able to identify the preferences on the basis of distribution of the demanded qualities. It turns out that these preferences happen to be the solution to the dual Monge–Kantorovich problem. This approach is explained in sections 9.2 and 9.4. In this context, OT therefore provides a constructive identification strategy.

– Some *incomplete econometric models* can be addressed using OT theory. In some problems, data are incomplete or missing, which creates a partial identification issue. For instance, income is sometimes reported only in tax brackets, therefore a model using the distribution of income as a source of identification may be incomplete in the sense that several values of the parameters may be compatible with the observed distribution. The problem of determining the identified set, namely, the set of parameters compatible with the observed distribution, can be reformulated as an OT problem. OT problems enjoy nice computational properties that make them efficiently computable. Hence the OT approach to partial identification is practical, as it allows fast computation of the identified set. Section 9.1 elaborates on this.

– *Quantile methods* are useful econometric and statistical techniques for analyzing distributions and dependence between random variables. They include among others quantile regression, quantile treatment effect, and least absolute deviation estimation. In dimension one, a quantile map is simply the inverse of the cumulative distribution function. As we shall see in chapter 4, quantile maps are very closely connected to an OT problem. In particular, OT provides a way to define a generalization of the notion of a quantile; see sections 9.4 and 9.5.

– In *contract theory*, multidimensional principal–agents problems may be reformulated as OT problems, as seen in section 9.6. This reformulation has useful econometric implications, as it allows us to infer each agent’s unobserved type based on the observed choices, assuming that the distribution of types is known.

– Some *derivative pricing* questions can be answered using OT, in particular the problem of bounds on derivative prices. A derivative is a financial asset whose value depends on the value of one or several other traded assets, called underlyings. Derivatives with several underlyings are often hard to price in

practice, as their value depends not only on the distribution of the value of each underlying, but also on the joint distribution. Often, the distribution of each underlying is known, as it can be recovered from market prices. The Monge–Kantorovich theorem is then useful to analyze bounds of prices of derivatives with multiple underlyings. An example of this method is provided in section 9.7. Similarly, OT is useful in *risk management* problems, where the measure of a given risk often depends on the joint distribution of several risks whose marginal distribution is known but whose dependence structure is unknown. Providing bounds on these measures of risk can then be rephrased as an OT problem.

1.2 A MIX OF TECHNIQUES

Surprisingly, in several other scientific disciplines, OT has allowed old problems to be revisited, and has brought new insights into them. This is the case in astrophysics (where OT has been used to model the early universe), in meteorology (where it has been used to model atmospheric fronts), in image analysis (where it provides convenient interpolation tools), and even in pure mathematics (it is insightful for analyzing Ricci curvature in Riemannian geometry). What is the reason for this apparent universality? Why is OT so prevalent? One answer may come from the very strong link between OT and convex analysis. Convex analysis is a most useful tool in many sciences, and OT is a way to revisit convex analysis in depth. As we shall see, one can learn about convex analysis almost entirely from the sole point of view of OT. In fact, the latter allows—at no extra cost—a significant generalization of convex analysis, which will be described in section 7.1. Hence, it should not be surprising that many problems in economics and other disciplines have a natural reformulation as an OT problem.

Moreover, developing an in-depth knowledge of OT will help the reader to discover, or rediscover, a number of tools. Indeed, one interesting feature of OT, especially from the point of view of a student eager to learn useful techniques “on the go,” is that it is connected to a number of important methods from various fields. OT is a mix of different techniques, and this text will contain a number of “crash courses” on a variety of topics. Let us briefly discuss a few of these topics and how they will occur in the book.

– *Linear programming* will underlie much of these notes. While optimal transportation in the general (continuous) case is an infinite-dimensional linear programming problem, and needs to be studied with more specific tools, we will see in chapter 3 that it boils down in the discrete case to a prototypical linear programming problem. In chapter 3 and appendix B we will spell out the basics of linear programming, with no prerequisite knowledge on the topic.

– More generally, this book will make heavy use of *large-scale optimization* methods. Indeed, the optimal transportation problem is a linear programming problem of a particular sort in the sense that it has a very sparse structure: the matrix of constraints contains many zeros. When computing these problems using linear programming algorithms, this fact calls for the use of large-scale optimization techniques, which take the sparsity of the matrix of constraints into consideration. We will demonstrate the interest of recognizing the sparse structure of the problem by giving computational examples written in R interfaced with Gurobi, a state-of-the-art linear programming solver.

– *Convex analysis* will be met in several places within this book. In the first place, OT problems are convex optimization problems, as are all linear programming problems. Also, we will see in section 6 that a special case of the OT problem, when the surplus is quadratic and the distributions are continuous, yields solutions that are convex functions. Chapter 6 will then be the occasion to revisit convex analysis from the point of view of optimal transport.

– The general setting of *network flow problems* will be studied in chapter 8. These extend discrete OT problems, and are among the most useful and best-studied problems in operations research. In a minimum cost flow problem, one seeks to send mass from a number of source nodes to a number of demand nodes through the network along paths of intermediate nodes in a way which minimizes the total transportation cost. Minimum cost flow problems combine a shortest path problem (find the cheapest path from one supply to one demand node) and an OT problem (find the optimal assignment between supply and demand nodes associated with the optimal cost between any pair of nodes). There is a continuous extension of this theory—not discussed in these notes—whose cornerstone result is McCann’s theorem on optimal transportation on manifolds.

– Finally, these notes will also incidentally feature some tools for *spatial data analysis* and *computational geometry* (introducing Voronoi cells, power diagrams, and the Hotelling location game); for *supermodularity*; and for *matrix theory*. The list goes on, but it should by no means discourage the reader. Again, these notes were written to be as self-contained as possible, in the hope that the reader will develop a working knowledge of the mix of techniques that is required for an in-depth understanding of OT.

1.3 BRIEF HISTORY

The history of OT starts with a French mathematician and statesman, Gaspard Monge (1746–1818, see figure 2.1), who is also the inventor of descriptive geometry, and the founder of École Polytechnique. Monge formulated the problem for the first time in 1781 out of civil engineering

preoccupations. As we will see in chapter 2, Monge was concerned with a particularly difficult variant of the problem, and the solution he gave was incomplete. Despite significant efforts, nineteenth-century mathematicians failed to overcome the difficulty. The problem remained unsolved until 1941, when the great Soviet mathematician Leonid Kantorovich (1912–1986, see figure 2.2), and independently a few years after him, Koopmans and his collaborators, introduced the relaxation technique described in chapter 2, allowing the problem to be relaxed into a linear programming problem. Duality provided a powerful tool to analyze the problem and its properties, and to provide an economic interpretation. The second half of the twentieth century mostly focused on the discrete assignment problem, detailed in chapter 3. It was only at the end of the 1980s and in the 1990s, with the work of Brenier, Knott, Rachev, Rüschendorf, Smith, McCann, Gangbo, and others, that Kantorovich duality was put to efficient use to fully solve Monge’s problem with quadratic costs and make a complete connection with convex duality, as described in chapter 6. This discovery, whose most striking formulation is Brenier’s theorem, presented in section 6.2, sparked a renewed interest in the topic, and OT is currently a very active area of research in mathematics and many applied sciences.

The interest in the numerical computation of OT problems (in its finite-dimensional version) is almost as old as the problem itself, although it was studied independently. It seems that the first efficient assignment algorithm (known today as the “Hungarian algorithm”) was discovered by Carl Jacobi around 1850, and later rediscovered in the 1950s with the work of Kuhn, Munkres, König, and Egerváry.

1.4 LITERATURE

There are good sources on OT in the mathematical literature. Primary references include two excellent, and fairly recent monographs by Cédric Villani. *Topics in Optimal Transportation* [148] is a set of great lecture notes, written in an intuitive way. The intellectual debt the present text owes to the former cannot be overstated. We will very often refer to it, and suggest that the reader should study both texts in parallel. *Optimal Transport: Old and New* [149] is the definitive reference on the topic, written in a more encyclopedic way, which is why we do not recommend it as an introduction. Both of these books, even the first one, require a significant investment; further, the author’s favorite application is fluid mechanics rather than economics. For these reasons, economists do not always find it easy to appropriate this material. Another good source is Rachev and Rüschendorf’s two volume treatise *Mass Transportation Problems* [118]. Although somewhat outdated given the progress in the literature over the last twenty years, it

contains insightful discussions and examples that are not found in Villani's texts. Finally, Santambrogio's recent book, *Optimal Transport for Applied Mathematicians* [132], has an original perspective on the topic and offers some interesting computational considerations. None of these texts has a focus on economics; in contrast, at least two high quality sets of introductory lecture notes are explicitly aimed at economic applications, despite being written in the mathematical tradition. One is Carlier's unpublished 2010 lecture notes [32], which can easily be found online. The other one is Ekeland's lecture notes [50], which appeared in 2010 in a special issue of *Economic Theory* dedicated to economic applications of OT.

On the other hand, the economic literature is somewhat terse on up-to-date reference texts for OT and its economic applications. There is little or no mention of the topic in the main graduate microeconomics textbooks. The classic treatise by Roth and Sotomayor [125] on two-sided matching deals mostly with models with nontransferable utility, that do not belong in the category of OT problems. It has one excellent chapter on the optimal assignment in the finite-dimensional (i.e., discrete) case, with which chapter 3 of the present text partially overlaps, but it covers none of the more advanced topics. Vohra [150] has excellent coverage of much of the basic mathematical machinery needed for the present book, and has a concise, yet informative section on assignment problems, but is also restricted to the finite-dimensional case. Another enlightening text by the same author, Vohra [151], offers a unifying reformulation of mechanism design theory using network flows; however, it also predominantly focuses on the discrete case, and it deals only with mechanism design applications.

Given the central importance of OT in the field, it is somewhat strange that the economic literature is missing an introductory text on the topic. We believe that the time has come for such a book.

1.5 ABOUT THESE NOTES

The purpose of the present notes is precisely to guide economists through this topic, by highlighting the potential for economic applications, and cutting short the part of the theory which is not of primordial importance for the latter. These notes are therefore intended as a complement to, rather than as a substitute for Villani's text mentioned above, [148], which we strongly suggest that the reader should read in parallel.

Because the purpose of this book is not to replace [148], but rather to complement it, this book is written in such a way that the formal statements, theorems and propositions, will be mathematically correct, but the proofs will sometimes be only sketched, or sometimes be shown under a set of stronger assumptions. Our style of exposition therefore draws inspiration

from texts on mathematical physics, and we will at times content ourselves with a “sketch of proof” explaining why a result is true without providing a proof that mathematicians would receive as acceptable. As a result, this text will be self-contained for readers who are content with results and their economic intuition; but readers who want to see full proofs will often be referred to [148], or Villani’s other monograph [149].

Another contrast between Villani’s text and the present one is the focus on computation in the latter. Economists (or, more precisely, econometricians) need to take their models to data. Economists are happy to know about the existence of a solution, but they worry if they cannot compute it in a reasonable amount of time. Complementing mathematical results with algorithms is quite natural as OT problems are closely linked to linear programming and optimal assignments, which are computationally tractable optimization problems for which there are well-developed efficient solutions. Hence, computation will be inherently part of this book, and examples labeled “Programming Example” will provide details on implementations. Our approach, however, has been strongly biased. Rather than looking for the most efficient method adapted to a given particular problem, we have sought to demonstrate that general purpose linear programming techniques, combined with the use of libraries to handle sparse matrices and matrix algebra, yield satisfactory results for most applications we will discuss. The demonstration codes are therefore written in R, which is an open-access mathematical programming language and allows easy and quick prototyping of programs. Although it is not advisable to use R directly for optimization, it can easily be interfaced with most optimization solvers. We will make frequent use of Gurobi, a state-of-the-art linear programming solver, which is commercial software, but is provided for free to the academic community (www.gurobi.com). The full set of programs is provided via this book’s web page at <http://press.princeton.edu/titles/10870.html>. The reader is strongly encouraged to try their own programs and compare them with those provided.

As this text is intended as a graduate course, and as learning requires practice, a number of exercises are provided throughout the book. Some of them (labeled “M”) are intended to develop mathematical agility; some (labeled “C”) help the reader to get used to computational techniques; others (labeled “E”) are intended to build economic intuition.

1.6 ORGANIZATION OF THIS BOOK

Let us briefly summarize the content of each chapter.

Chapter 2 states the Monge–Kantorovich problem and provides the duality result in a fairly general setting. The primal problem is interpreted as

the central planner's problem of determining the optimal assignment of workers to firms, while the dual problem is interpreted as the invisible hand's problem of obtaining a system of decentralized equilibrium prices. In general, the primal problem always has a solution (which means that an optimal assignment of workers to jobs exists), but the dual does not: the optimal assignment cannot always be decentralized by a system of prices. However, as we shall see later, the cases where the dual problem does not have a solution are rather pathological, and in all of the cases considered in the rest of the book, both the primal and the dual problems have solutions.

Chapter 3 considers the finite-dimensional case, which is the case when the marginal probability distributions are discrete with finite support. In this case the Monge–Kantorovich problem becomes a finite-dimensional linear programming problem; the primal and the dual solutions are related by complementary slackness, which is interpreted in terms of stability. The solutions can be conveniently computed by linear programming solvers, and we will show how to do this using some matrix algebra and Gurobi.

Chapter 4 considers the univariate case, when both the worker and the job are characterized by a scalar attribute. The important assumption of positive assortative matching, or supermodularity of the matching surplus, will be introduced and discussed. As a consequence, the primal problem has an explicit solution (an optimal assignment) which is related to the probabilistic notion of a quantile transform, and the dual problem also has an explicit solution (a set of equilibrium prices), which are obtained from the solution to the primal problem. As a consequence, the Monge–Kantorovich problem is explicitly solved in dimension one under the assumption of positive assortative matching.

Chapter 5 considers the case when the attributes are d -dimensional vectors, the matching surplus is the scalar product, the distribution of workers' attributes is continuous, and the distribution of the firms is discrete. The geometry of the optimal assignment of workers to firms is discussed and related to the important notion of power diagrams in computational geometry. The optimal assignment map is shown to be the gradient of a piecewise affine convex function, and the equilibrium prices of the firms are shown to be the solution to a finite-dimensional convex minimization problem. We will discuss how to perform this computation in practice.

Chapter 6 still considers the case when the attributes are d -dimensional vectors and the surplus is the scalar product; it still assumes that the distribution of the workers' attributes is continuous, but it relaxes the assumption that the distribution of the firms' attributes is discrete. This setting allows us to entirely rediscover convex analysis, which is introduced from the point of view of optimal transport. As a consequence, Brenier's polar factorization theorem is given, which provides a vector extension for the scalar notions of quantile and rearrangement.

Chapter 7 considers a case with a more general surplus function. This is the place to show that when the scalar-product surplus is replaced by a more general function, much of the machinery put in place in chapter 6 goes through. In particular, it is possible to generalize convex analysis in a natural way, and to obtain generalized notions of convex conjugates, of convexity, and of a subdifferential that are perfectly suited to the problem. A general result on the existence of dual minimizers will be given, as well as sufficient conditions for the existence of a solution to the Monge problem.

Chapter 8 considers the optimal network flow problem, which is a generalization of the optimal assignment problem considered in chapter 3. In optimal flow problems, one considers a network of cities, or edges, to move a distribution of mass on supply nodes to a distribution of mass on demand nodes. The difference from a standard optimal assignment problem is that the matching surplus associated with moving from a supply location to a demand location is not necessarily directly defined; instead, there are several paths from the supply location to the demand location, among these some yield maximal surplus. Therefore, both the optimal assignment problem and the shortest path problem are instances of the optimal flow problem; these instances are representative in the sense that any optimal flow problem may be decomposed into an assignment problem and a number of shortest path problems. We will show how to easily compute these problems using linear programming.

Chapter 9 offers a selection of applications to economics: partial identification in econometrics in section 9.1, inversion of demand systems in section 9.2, computation of hedonic equilibria in section 9.3, identification via vector quantile methods in section 9.4, quantile regression in section 9.5, multidimensional screening in section 9.6, and pricing of financial derivatives in section 9.7.

Chapter 10 concludes with perspectives on computation, duality, and equilibrium.

1.7 NOTATION AND CONVENTIONS

Throughout this text we have tried to strike a good balance between mathematical precision and ease of exposition. A *probability measure* will always mean a Borel probability measure; a *set* will always mean a measurable set; a *continuous probability* or *continuous distribution* means a probability measure which is absolutely continuous with respect to the Lebesgue measure; a *convex* function means a convex function which is not identically $+\infty$.

Usual abbreviations will be used: *c.d.f.* means the cumulative distribution function; *p.d.f.* means probability density function; *a.s.* means almost surely; depending on the context, *s.t.* means such that or subject to.

Given a smooth function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the gradient of f at x , denoted $\nabla f(x)$, is the vector of partial derivatives $(\partial f(x)/\partial x_1, \dots, \partial f(x)/\partial x_d)$. Given a function $f : \mathbb{R}^k \rightarrow \mathbb{R}^l$, $Df(x)$ is the Jacobian matrix of f at x , that is, the matrix of partial derivatives $(\partial f^i(x)/\partial x^j)$, $1 \leq i \leq l$, $1 \leq j \leq k$. The Dirac mass at x_0 , denoted δ_{x_0} , is the probability distribution whose c.d.f. is $F(x) = 1 \{x_0 \leq x\}$. The notation $A \succeq_{\text{spd}} B$ means $A - B$ is symmetric, positive semidefinite. Given a compact set C of \mathbb{R}^d , the notation $\mathcal{U}(C)$ denotes the uniform distribution on C , which is simply denoted \mathcal{U} when $C = [0, 1]$ is the unit interval. The set $L^1(P)$ is the set of functions which are integrable with respect to P . $X \sim P$ means X has distribution P . Perhaps less standard notation is $\mathcal{M}(P, Q)$, which denotes the set of couplings of P and Q , which is the set of probability measures π such that if $(X, Y) \sim \pi$, then $X \sim P$ and $Y \sim Q$.

Throughout this book, we shall try to have consistent notation, which we now summarize. Some exceptions are made in chapter 9, where we sometimes choose to use notation that is more traditional in the given application.

Agents: The types or attributes of workers are denoted by $x \in \mathcal{X}$, and the types or attributes of firms (or jobs) are denoted by $y \in \mathcal{Y}$. Whenever needed, i indexes individual workers, and j indexes individual firms.

Probabilities: A probability measure on \mathcal{X} is denoted by P and on \mathcal{Y} by Q . The probability measure of observing pair (x, y) is denoted by $\pi \in \mathcal{M}(P, Q)$; according to the context (continuous or discrete), this probability measure may be identified with its p.d.f. or with its probability mass function in the discrete case.

Utilities: The pretransfer surplus of a type x when matched with a type y is denoted by $\alpha(x, y)$, and the pretransfer surplus of a type y when matched with a type x is denoted by $\gamma(x, y)$. The matching surplus of a couple (x, y) is denoted by $\Phi(x, y) = \alpha(x, y) + \gamma(x, y)$. The equilibrium monetary transfer (wage) from y to x within pair (x, y) is denoted by $w(x, y)$. The posttransfer indirect surplus of a type x is denoted by $u(x)$, and the posttransfer indirect surplus of a type y is denoted by $v(y)$.

Whenever types x and y are discrete, we will prefer to use subscript notation, that is, Φ_{xy} instead of $\Phi(x, y)$.