Computational Optimal Transport

with Applications to Data Sciences
Other titles in Foundations and Trends® in Machine Learning

*Non-convex Optimization for Machine Learning*
Prateek Jain and Purushottam Ka

*Kernel Mean Embedding of Distributions: A Review and Beyond*
Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Scholkopf
ISBN: 978-1-68083-288-4

*Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions*
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-222-8

*Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 2 Applications and Future Perspectives*
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-276-1

*Patterns of Scalable Bayesian Inference*
Elaine Angelino, Matthew James Johnson and Ryan P. Adams
ISBN: 978-1-68083-218-1

*Generalized Low Rank Models*
Madeleine Udell, Corinne Horn, Reza Zadeh and Stephen Boyd
Computational Optimal Transport

with Applications to Data Sciences

Gabriel Peyré
CNRS and DMA, ENS

Marco Cuturi
Google and CREST/ENSAE
Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2018, Volume 11, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
# Contents

1 **Introduction** 3

2 **Theoretical Foundations** 8
   2.1 Histograms and Measures 9
   2.2 Assignment and Monge Problem 11
   2.3 Kantorovich Relaxation 16
   2.4 Metric Properties of Optimal Transport 23
   2.5 Dual Problem 28
   2.6 Special Cases 36

3 **Algorithmic Foundations** 45
   3.1 The Kantorovich Linear Programs 46
   3.2 C-Transforms 48
   3.3 Complementary Slackness 50
   3.4 Vertices of the Transportation Polytope 51
   3.5 A Heuristic Description of the Network Simplex 55
   3.6 Dual Ascent Methods 60
   3.7 Auction Algorithm 63

4 **Entropic Regularization of Optimal Transport** 70
   4.1 Entropic Regularization 71
   4.2 Sinkhorn’s Algorithm and Its Convergence 77
4.3 Speeding Up Sinkhorn’s Iterations 89
4.4 Stability and Log-Domain Computations 94
4.5 Regularized Approximations of the Optimal Transport Cost 98
4.6 Generalized Sinkhorn 101

5 Semidiscrete Optimal Transport 104
5.1 \(c\)-Transform and \(\bar{c}\)-Transform 105
5.2 Semidiscrete Formulation 106
5.3 Entropic Semidiscrete Formulation 109
5.4 Stochastic Optimization Methods 113

6 \(W_1\) Optimal Transport 117
6.1 \(W_1\) on Metric Spaces 118
6.2 \(W_1\) on Euclidean Spaces 120
6.3 \(W_1\) on a Graph 121

7 Dynamic Formulations 124
7.1 Continuous Formulation 124
7.2 Discretization on Uniform Staggered Grids 128
7.3 Proximal Solvers 129
7.4 Dynamical Unbalanced OT 132
7.5 More General Mobility Functionals 134
7.6 Dynamic Formulation over the Paths Space 134

8 Statistical Divergences 138
8.1 \(\varphi\)-Divergences 139
8.2 Integral Probability Metrics 145
8.3 Wasserstein Spaces Are Not Hilbertian 151
8.4 Empirical Estimators for OT, MMD and \(\varphi\)-divergences 155
8.5 Entropic Regularization: Between OT and MMD 159

9 Variational Wasserstein Problems 161
9.1 Differentiating the Wasserstein Loss 162
9.2 Wasserstein Barycenters, Clustering and Dictionary Learning 168
9.3 Gradient Flows 180
9.4 Minimum Kantorovich Estimators 188
Computational Optimal Transport
Gabriel Peyré¹ and Marco Cuturi²

¹ CNRS and DMA, ENS; gabriel.peyre@ens.fr
² Google and CREST/ENSAE; cuturi@google.com

ABSTRACT
Optimal transport (OT) theory can be informally described using the words of the French mathematician Gaspard Monge (1746–1818): A worker with a shovel in hand has to move a large pile of sand lying on a construction site. The goal of the worker is to erect with all that sand a target pile with a prescribed shape (for example, that of a giant sand castle). Naturally, the worker wishes to minimize her total effort, quantified for instance as the total distance or time spent carrying shovelfuls of sand. Mathematicians interested in OT cast that problem as that of comparing two probability distributions—two different piles of sand of the same volume. They consider all of the many possible ways to morph, transport or reshape the first pile into the second, and associate a “global” cost to every such transport, using the “local” consideration of how much it costs to move a grain of sand from one place to another. Mathematicians are interested in the properties of that least costly transport, as well as in its efficient computation. That smallest cost not only defines a distance between distributions, but it also entails a rich geometric structure on the space of probability distributions. That structure is canonical in the sense that it borrows key geometric properties of the underlying “ground” space on which these distributions are defined. For instance, when the underlying space is Euclidean, key concepts such
as interpolation, barycenters, convexity or gradients of functions extend naturally to the space of distributions endowed with an OT geometry.

OT has been (re)discovered in many settings and under different forms, giving it a rich history. While Monge’s seminal work was motivated by an engineering problem, Tolstoi in the 1920s and Hitchcock, Kantorovich and Koopmans in the 1940s established its significance to logistics and economics. Dantzig solved it numerically in 1949 within the framework of linear programming, giving OT a firm footing in optimization. OT was later revisited by analysts in the 1990s, notably Brenier, while also gaining fame in computer vision under the name of earth mover’s distances. Recent years have witnessed yet another revolution in the spread of OT, thanks to the emergence of approximate solvers that can scale to large problem dimensions. As a consequence, OT is being increasingly used to unlock various problems in imaging sciences (such as color or texture processing), graphics (for shape manipulation) or machine learning (for regression, classification and generative modeling).

This paper reviews OT with a bias toward numerical methods, and covers the theoretical properties of OT that can guide the design of new algorithms. We focus in particular on the recent wave of efficient algorithms that have helped OT find relevance in data sciences. We give a prominent place to the many generalizations of OT that have been proposed in but a few years, and connect them with related approaches originating from statistical inference, kernel methods and information theory. All of the figures can be reproduced using code made available in a companion website\(^1\). This website hosts the book project Computational Optimal Transport. You will also find slides and computational resources.

\(^1\)https://optimaltransport.github.io/
The shortest path principle guides most decisions in life and sciences: When a commodity, a person or a single bit of information is available at a given point and needs to be sent at a target point, one should favor using the least possible effort. This is typically reached by moving an item along a straight line when in the plane or along geodesic curves in more involved metric spaces. The theory of optimal transport generalizes that intuition in the case where, instead of moving only one item at a time, one is concerned with the problem of moving simultaneously several items (or a continuous distribution thereof) from one configuration onto another. As schoolteachers might attest, planning the transportation of a group of individuals, with the constraint that they reach a given target configuration upon arrival, is substantially more involved than carrying it out for a single individual. Indeed, thinking in terms of groups or distributions requires a more advanced mathematical formalism which was first hinted at in the seminal work of Monge (1781). Yet, no matter how complicated that formalism might look at first sight, that problem has deep and concrete connections with our daily life. Transportation, be it of people, commodities or information, very rarely involves moving only one item. All major economic problems, in logistics, production
planning or network routing, involve moving distributions, and that thread appears in all of the seminal references on optimal transport. Indeed Tolstoi (1930), Hitchcock (1941) and Kantorovich (1942) were all guided by practical concerns. It was only a few years later, mostly after the 1980s, that mathematicians discovered, thanks to the works of Brenier (1991) and others, that this theory provided a fertile ground for research, with deep connections to convexity, partial differential equations and statistics. At the turn of the millenium, researchers in computer, imaging and more generally data sciences understood that optimal transport theory provided very powerful tools to study distributions in a different and more abstract context, that of comparing distributions readily available to them under the form of bags-of-features or descriptors.

Several reference books have been written on optimal transport, including the two recent monographs by Villani (2003; 2009), those by Rachev and Rüschendorf (1998; 1998) and more recently that by Santambrogio (2015). As exemplified by these books, the more formal and abstract concepts in that theory deserve in and by themselves several hundred pages. Now that optimal transport has gradually established itself as an applied tool (for instance, in economics, as put forward recently by Galichon (2016)), we have tried to balance that rich literature with a computational viewpoint, centered on applications to data science, notably imaging sciences and machine learning. We follow in that sense the motivation of the recent review by Kolouri et al. (2017) but try to cover more ground. Ultimately, our goal is to present an overview of the main theoretical insights that support the practical effectiveness of OT and spend more time explaining how to turn these insights into fast computational schemes. The main body of Chapters 2, 3, 4, 9, and 10 is devoted solely to the study of the geometry induced by optimal transport in the space of probability vectors or discrete histograms. Targeting more advanced readers, we also give in the same chapters, in light gray boxes, a more general mathematical exposition of optimal transport tailored for discrete measures. Discrete measures are defined by their probability weights, but also by the location at which these weights are defined. These locations are usually taken in a continuous metric space, giving a second important degree of freedom.
to model random phenomena. Lastly, the third and most technical layer of exposition is indicated in dark gray boxes and deals with arbitrary measures that need not be discrete, and which can have in particular a density w.r.t. a base measure. This is traditionally the default setting for most classic textbooks on OT theory, but one that plays a less important role in general for practical applications. Chapters 5 to 8 deal with the interplay between continuous and discrete measures and are thus targeting a more mathematically inclined audience.

The field of computational optimal transport is at the time of this writing still an extremely active one. There are therefore a wide variety of topics that we have not touched upon in this survey. Let us cite in no particular order the subjects of distributionally robust optimization (Shafieezadeh Abadeh et al., 2015; Esfahani and Kuhn, 2018; Lee and Raginsky, 2018; GAO et al., 2018), in which parameter estimation is carried out by minimizing the worst possible empirical risk of any data measure taken within a certain Wasserstein distance of the input data; convergence of the Langevin Monte Carlo sampling algorithm in the Wasserstein geometry (Dalalyan and Karagulyan, 2017; Dalalyan, 2017; Bernton, 2018); other numerical methods to solve OT with a squared Euclidian cost in low-dimensional settings using the Monge-Ampère equation (Froese and Oberman, 2011; Benamou et al., 2014; Sulman et al., 2011) which are only briefly mentioned in Remark 2.25.

**Notation**

- $[n]$: set of integers $\{1, \ldots, n\}$.
- $\mathbb{1}_{n,m}$: matrix of $\mathbb{R}^{n \times m}$ with all entries identically set to 1. $\mathbb{1}_n$: vector of ones.
- $\mathbb{I}_n$: identity matrix of size $n \times n$.
- For $u \in \mathbb{R}^n$, $\text{diag}(u)$ is the $n \times n$ matrix with diagonal $u$ and zero otherwise.
- $\Sigma_n$: probability simplex with $n$ bins, namely the set of probability vectors in $\mathbb{R}_+^n$. 

Full text available at: http://dx.doi.org/10.1561/2200000073

Full text available at: http://dx.doi.org/10.1561/2200000073
Introduction

- \((a, b)\): histograms in the simplices \(\Sigma_n \times \Sigma_m\).
- \((\alpha, \beta)\): measures, defined on spaces \((X, Y)\).
- \(\frac{d\alpha}{d\beta}\): relative density of a measure \(\alpha\) with respect to \(\beta\).
- \(\rho_\alpha = \frac{d\alpha}{dx}\): density of a measure \(\alpha\) with respect to Lebesgue measure.
- \((\alpha = \sum_i a_i \delta_{x_i}, \beta = \sum_j b_j \delta_{y_j})\): discrete measures supported on \(x_1, \ldots, x_n \in X\) and \(y_1, \ldots, y_m \in Y\).
- \(c(x, y):\) ground cost, with associated pairwise cost matrix \(C_{i,j} = (c(x_i, y_j))_{i,j}\) evaluated on the support of \(\alpha, \beta\).
- \(\pi\): coupling measure between \(\alpha\) and \(\beta\), namely such that for any \(A \subset X, \pi(A \times Y) = \alpha(A)\), and for any subset \(B \subset Y, \pi(X \times B) = \beta(B)\). For discrete measures \(\pi = \sum_{i,j} P_{i,j} \delta_{x_i,y_j}\).
- \(\mathcal{U}(\alpha, \beta)\): set of coupling measures, for discrete measures \(\mathbf{U}(a, b)\).
- \(\mathcal{R}(c)\): set of admissible dual potentials; for discrete measures \(\mathbf{R}(C)\).
- \(T : X \rightarrow Y\): Monge map, typically such that \(T_\sharp \alpha = \beta\).
- \((\alpha_t)_{t=0}^1\): dynamic measures, with \(\alpha_{t=0} = \alpha_0\) and \(\alpha_{t=1} = \alpha_1\).
- \(v\): speed for Benamou–Brenier formulations; \(J = \alpha v\): momentum.
- \((f, g)\): dual potentials, for discrete measures \((f, g)\) are dual variables.
- \((u, v) \stackrel{\text{def}}{=} (e^{f/\varepsilon}, e^{g/\varepsilon})\): Sinkhorn scalings.
- \(K \stackrel{\text{def}}{=} e^{-C/\varepsilon}\): Gibbs kernel for Sinkhorn.
- \(s\): flow for \(W_1\)-like problem (optimization under divergence constraints).
- \(L_C(a, b)\) and \(L_c(\alpha, \beta)\): value of the optimization problem associated to the OT with cost \(C\) (histograms) and \(c\) (arbitrary measures).
• $W_p(a, b)$ and $W_p(\alpha, \beta)$: $p$-Wasserstein distance associated to ground distance matrix $D$ (histograms) and distance $d$ (arbitrary measures).

• $\lambda \in \Sigma_S$: weight vector used to compute the barycenters of $S$ measures.

• $\langle \cdot, \cdot \rangle$: for the usual Euclidean dot-product between vectors; for two matrices of the same size $A$ and $B$, $\langle A, B \rangle \overset{\text{def}}{=} \text{tr}(A^\top B)$ is the Frobenius dot-product.

• $f \oplus g(x, y) \overset{\text{def}}{=} f(x) + g(y)$, for two functions $f : \mathcal{X} \to \mathbb{R}, g : \mathcal{Y} \to \mathbb{R}$, defines $f \oplus g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

• $f \oplus g \overset{\text{def}}{=} f^\top_m + 1_n g^\top \in \mathbb{R}^{n \times m}$ for two vectors $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$.

• $\alpha \otimes \beta$ is the product measure on $\mathcal{X} \times \mathcal{Y}$, i.e. $\int_{\mathcal{X} \times \mathcal{Y}} g(x, y) \text{d}(\alpha \otimes \beta)(x, y) = \int_{\mathcal{X}} g(x, \cdot) \text{d}\alpha(x) \text{d}\beta(y)$.

• $a \otimes b \overset{\text{def}}{=} ab^\top \in \mathbb{R}^{n \times m}$.

• $u \odot v = (u_i v_i) \in \mathbb{R}^n$ for $(u, v) \in (\mathbb{R}^n)^2$. 

Full text available at: http://dx.doi.org/10.1561/2200000073


References


References


Oliker, V. and L. D. Prussner. 1989. “On the numerical solution of the equation \[ \frac{\partial^2 z}{\partial x^2} \frac{\partial^2 z}{\partial y^2} - \left( \frac{\partial^2 z}{\partial x \partial y} \right)^2 = f \] and its discretizations, I”. *Numerische Mathematik*. 54(3): 271–293.


Full text available at: http://dx.doi.org/10.1561/2200000073


References


Tolstoı, A. 1930. “Metody nakhozhdeniya naimen’shego summovogo kilome-trazha pri planirovanii perevozok v prostranstve (Russian; Methods of finding the minimal total kilometrage in cargo transportation planning in space)”. TransPress of the National Commissariat of Transportation: 23–55.


