Optimal Transport meets Probability, Statistics and Machine Learning

Guillaume Carlier (Université paris Dauphine),
Marco Cuturi (ENSAE Paristech),
Brendan Pass (University of Alberta),
Carola Schoenlieb (Cambridge).

CMO, Oaxaca, May 1-May 5 2017

1 Overview of the Field

Optimal transport (OT) theory was born at the turn of the XVIIIth century with the work of Gaspard Monge who formulated the problem of *remblais and déblais*. In modern language, the problem consists in finding a measure-preserving map minimizing a total transport cost. After important contributions of Kantorovich in the 1940's who relaxed the problem thanks to the notion of transport plan and introduced a powerful dual formulation, the subject was tremendously revitalized in the late 1980's with the seminal work of Yann Brenier [7] who solved the quadratic cost case and subsequent important contributions by McCann, Gangbo, Caffarelli, Otto, Villani and many others. The connections with fluid dynamics, geometry, functional inequalities and the textbooks of Fields medalist Cédric Villani [29, 30] as well as the more applied-math oriented book by Filippo Santambrogio [26] contributed a lot to popularize optimal transport.

Optimal transport has long standing connections to probability, which have been amplified in recent years. For example, variants of the classical optimal transport problem have arisen in connection to applications in financial mathematics (including transport problems with additional dependence constraints and martingale optimal transport, where versions with several, or even infinitely many, prescribed marginals are also of interest) and Schrödinger's problem of minimizing the relative entropy of stochastic processes with fixed initial and final laws (Georgiou, Léonard).

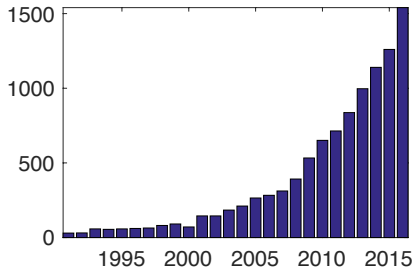

Figure 1: Evolution of papers with the words "optimal transport" and "Wasserstein" in the titles, according to Google Scholar.

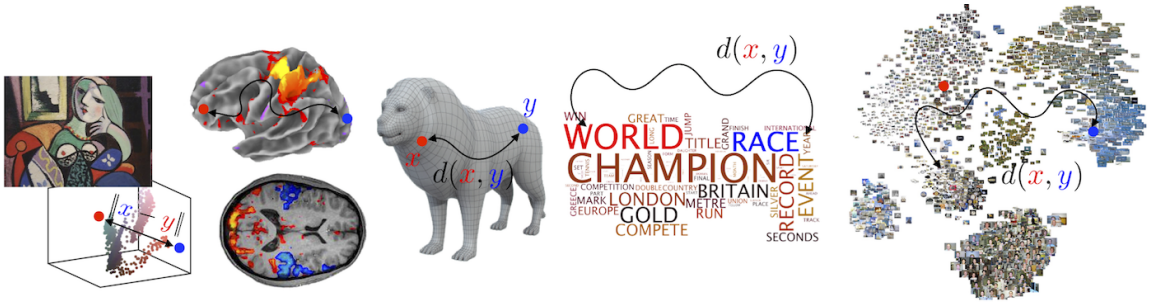

Figure 2: Overview of typical applications where OT already plays a crucial role, or is likely to have a strong impact. From let to right: color processing of image, activation detection in brain recordings, 3D shape editing, natural language processing, high dimensional machine learning for vision.

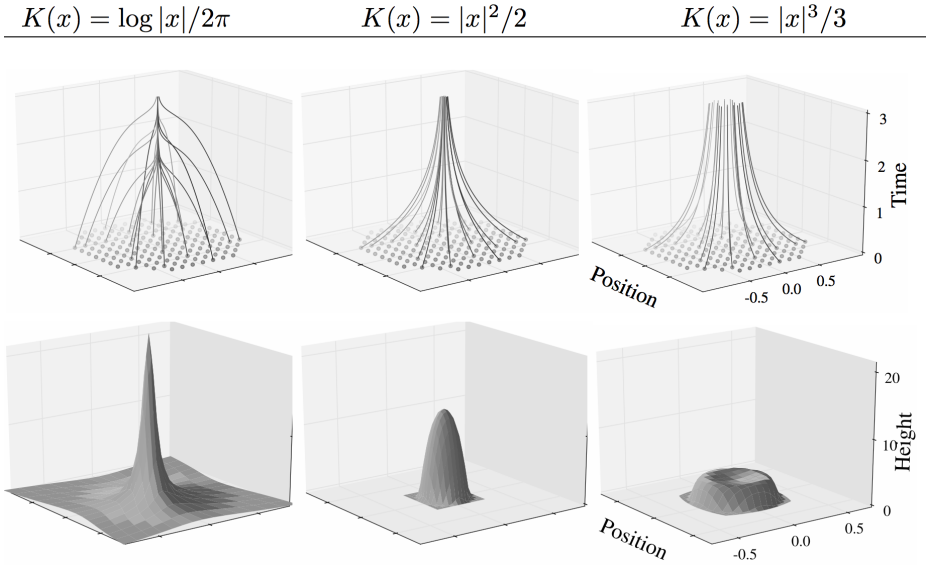

Figure 3: Evolution of the solution of Wasserstein gradient flow under an attraction potential, which generates singular aggregation patterns (credit Katy Craig).

The recent period (posterior to 2012) has seen an impressive development of efficient numerical methods and solvers for optimal transport in particular the entropic regularization approach (Cuturi [15]) and the semi-discrete approach (Mérigot, Oudet [23], Lévy [22]) have made it possible to solve accurately large scale optimal transport problems. At about the same time, optimal transport became more and more popular in image processing, statistics and even machine learning.

Each of these application fields brings forth significant new challenges, both theoretical and computational; in addition to addressing these, the organizers of this workshop were convinced that bringing together leaders in optimal transport with experts in probability and statistics might uncover even more connections and would stimulate research on both sides. Let us emphasize that the presence of Robert McCann, a world leader in the field of OT, during the whole workshop was particularly stimulating as well as his lecture which closed the meeting and was devoted to new and general results on the concavity of the so-called principal-agent problem, a central problem in economics which perfectly illustrated the power of OT ideas.

The workshop was indeed a great success, according to all participants and the organizers. It gathered 41 participants, and consisted of a combination of short and long talks as well as an open problems session and free time for discussions and collaborations. It is our belief that this workshop should be the first of a fruitful series to be held in the future.

2 Recent Developments and Open Problems

Analysis of Sinkhorn beyond the OT case As already pointed out, the entropic regularization approach has been extremely succesful in the recent years and Cuturi has emphasized its connection with Sinkhorn matrix scaling algorithm [27, 28] and [17]. It is well-known that the convergence of Sinkhorn algorithm for entropic regularization of discrete OT is linear, as can be proved by using Hilbert projective metric. It turns out that no such result is available for the multi-marginal extensions or to some non-commutative versions (transport of matrices for instance) as was pointed out during an open problems session of the workshop by Tryphon Georgiou. The multi-marginal version of Sinkhorn is known to converge but finding rates requires new tools. Another challenging problem, important in mathematical finance (see below) is martingale optimal transport, the multi-marginal extension is natural and almost nothing is known on the structure of solutions to such problems, for which it would be important to have good numerical methods, it turns out that entropic regularization performs rather badly on such problems...

OT for high dimensional machine learning The applications of OT methods are currently mostly focussed on low dimensional problems (typical examples being in social sciences or imaging sciences). Extending these successes to high dimensional problems in machine learning raises many theoretical, numerical and algorithmic challenges. In particular, OT distances are known to be difficult to estimate from discrete samples in high dimension, and current semi-discrete (to match discrete samples to continuous distributions) OT solvers lack scalability. An important area of both recent developpements (see in particular Section 3.4) and opportunity for major breakthrough is thus to be able to define useful and tractable OT-inspired metrics for high-dimensional statistical inference. Unleashing the geometry of OT in these high dimensional settings is likely be one of the key driving force for the future of machine learning, a typical example being deep neural nets training.

3 Presentation Highlights

3.1 OT and probability

3.1.1 Martingale Optimal transport

From an OT perspective, Martingale Optimal transport (MOT) is a version the classical OT problem with a further constraint that the transport plan is a martingale. From a probabilistic perspective, thinking in continuous time, it is the problem of selecting a solution to the Skorokhod embedding problem with an additional optimality property. Finally, from a financial perspective, it is the problem of computing range of no-arbitrage prices (primal) and robust hedging strategies which enforce these (dual), when given market quoted prices for co-maturing vanilla options (calls). MOT offers an exciting interaction of the three fields. It brought tremendous new geometrical insights into structure of (optimal) Skorokhod embeddings and it links naturally with martingale inequalities. In exchange, probabilistic methods offer explicit transport plans and the problem, when compared with OT, features an intricate structure of polar sets. Jan Obloj, gave a long talk during which he presented a panorama of the field with emphasis on some of the most recent contributions. Gaoyue Guo gave a short talk about approximation and stability of the MOT problem as well as numerical methods for MOT on the real line which clearly emphasized the differences with the standard OT case. Finally, Young-Heon Kim presented results on some fine structural properties of optimal transport plans, these results are the only structural ones for MOT in more than one dimension.

3.1.2 Schrödinger bridges: from classical to quantum

Christian Léonard introduced a large deviation principle leading to the Schrödinger bridge problem (SBP) and gave a new proof of the HWI inequality, based on this entropy minimization problem, This was an excellent occasion to present several properties of entropic transport. Tryphon Georgiou discussed generalizations of the Schrödinger Bridge Problem to the setting of matrix-valued and vector-valued distributions. Matrix-valued OMT in particular allows one to define a Wasserstein geometry on the space of density matrices of quantum mechanics and, as it turns out, the Lindblad equation of open quantum systems (quantum diffusion)

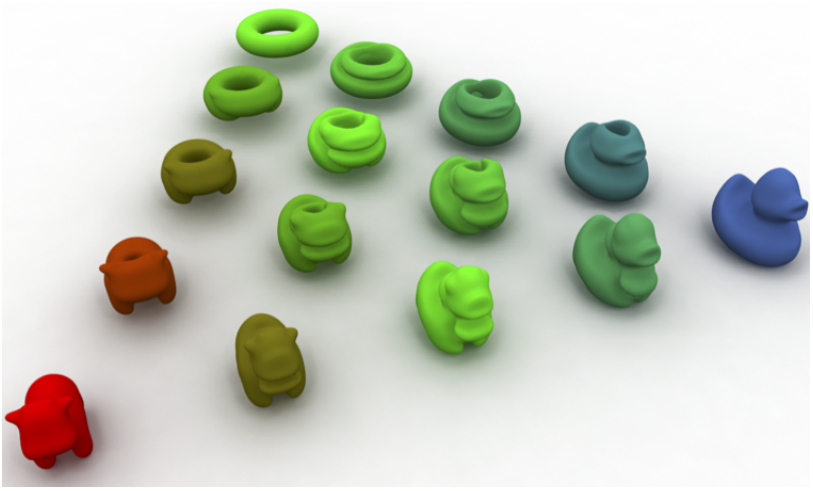

Figure 4: Computation of OT barycenter using entropic regularization (which is similar to Schrödinger's problem)

turns out to be exactly the gradient flow of the von Neumann quantum entropy in this sense. Tryphon Georgiou explained how OT and SBP ideas can be generaized to the setting of matrix-valued and vector-valued distributions.

3.2 OT: theory and numerics

3.2.1 Numerics

Bruno Lévy's talk gave a state of the art talk on numerical semi-discrete OT (a measure with a density is transported to a sum of Dirac masses). This setting is very well adapted to a computer implementation, because the transport map is determined by a vector of parameters (associated with each Dirac mass) that maximizes a concave function (Kantorovich dual). An efficient numerical solution mechanism requires to carefully orchestrate the interplay of geometry and numerics. On geometry, two algorithms to efficiently compute Laguerre cells were presented, one that uses arbitrary precision predicates, and one that uses standard double-precision arithmetics. On numerical aspects, when implementing a Newton solver (a 3D version of Kitagawa Merigot Thibert [21]), the main difficulty is to assemble the Hessian of the objective function, a sparse matrix with a non-zero pattern that changes during the iterations. By exploiting the relation between the Hessian and the 1-skeleton of the Laguerre diagram, it is possible to efficiently construct the Hessian. An important output of the method is that the algorithm can be used to simulate incompressible Euler fluids (Merigot, Gallouet [18]).

3.2.2 Gradient flows, modelling

One important application of OT concerns the Wasserstein gradient flow approach to evolution PDEs, a powerful theoretical tool but also becoming relevant from a numerical perspective as well. Indeed, the gradient flow structure gives rise to a variational scheme by means of the minimising movement scheme (also called JKO scheme, after the seminal work of Jordan, Kinderlehrer and Otto [20], see [2] for an overview of the theory) which constitutes a discrete-time minimization problem for the energy. While the scheme has been used originally for analytical aspects, a number of authors have explored the numerical potential of this scheme. Bertram Düring presented Lagrangian schemes where instead of the density, the evolution of a time-dependent homeomorphism that describes the spatial redistribution of the density is considered and illustrated these ideas on nonlinear diffusions of porous medium type as well as on the fourth-order Derrida-Lebowitz-Speer-Spohn equation. Clarice Poon presented new results on the JKO scheme for the total variation, motivated by denoising problems in image processing and a related highly nonlinear fourth-order PDE (for which OT arguments enable to prove some kind of maximum principle). Wasserstein gradient flows can also be

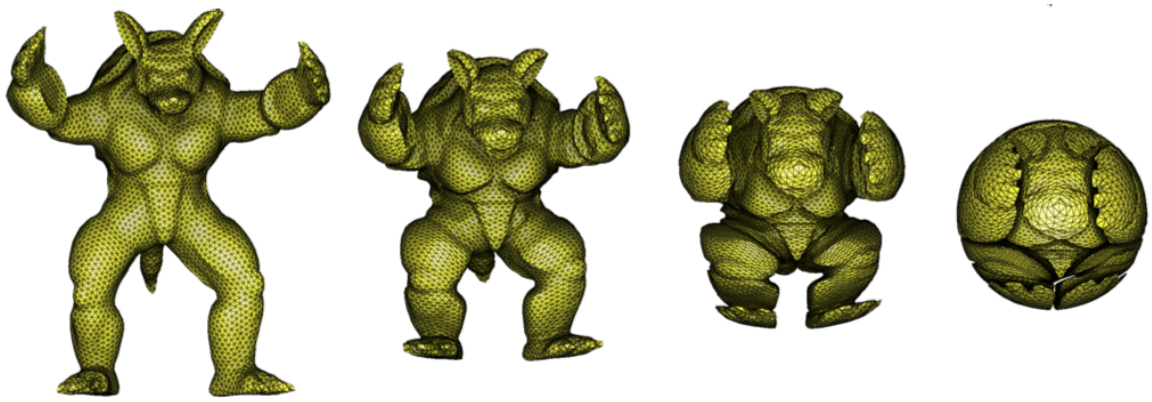

Figure 5: Application of semi-discrete OT to 3-D mesh warping (Bruno Levy)

used to study complex models where aggregation effects enter in competition with diffusion. Developing deterministic particle methods for problems involving diffusion poses unique challenges, particularly when nonlocal interaction terms are present. Katy Craig presented a new blob method for diffusion and degenerate diffusion, inspired by blob methods from classical fluid mechanics. The computational fluid dynamics approach to OT, pioneered by Benamou and Brenier [4] is also a cornerstone of OT and can be generalized in several directions for modeling and/or numerical purposes. It is worth mentioning here that it can be recovered as the zero-noise limit of the Schrödinger bridge problem (SBP), see paragraph 3.1.2. Christoph Brune presented a joint segmentation-optimal transport model; for studying vascular structures in 4D biomedical imaging, it is of great importance to automatically determine the velocity of flow in video sequences, for example blood flow in vessel networks. New optimal transport models focusing on direction and segmentation are investigated in this model to find an accurate displacement between two density distributions. By incorporating fluid dynamics constraints, one can obtain a realistic description of the displacement. With an a-priori given segmentation of the network structure, transport models can be improved. However, a segmentation is not always known beforehand and this is why developing joint segmentation-optimal transport models is important.

3.3 Multi-marginal OT, Wasserstein barycenters, applications

Multi-marginal OT (MMOT) problems consist in finding a plan between N probability measures which minimize the average of some cost function. Such problems have received a lot of attention due to their importance in applied settings such as density functional theory (DFT) in quantum chemistry (see the works of Cotar, Friesecke and coauthors [12], [13] and Buttazzo, De Pascale, Gori-Giorgi [10]), matching problems in economics ([11], [25]), symmetric OT problems (Ghoussoub and Maurey [19]) or fluid dynamics with Brenier's relaxed formulation of incompressible Euler equation ([6], [8],[9]) following the seminal work of Arnol'd [3], not that in the latter case there are infinitely many (one for each time) marginal constraints corresponding to incompressibility. Contrary to the well-established theory of two marginals OT where under a well-known twist condition, one knows that optimal plans have a graphical structure, the structure of MMOT is much less understood, with the notable exception of the results of Brendan Pass [24].Indeed, even simple costs may lead to optimal maps which are fractal (see [16])! A case which is well understood though is the quadratic problem corresponding to the notion of Wasserstein barycenters introduced by Agueh and Carlier [1] and which is becoming quite poular in statistics and image processing and for which efficient solvers exist (see [14], [5]). Elsa Cazelles presented a regularized version of the Wasserstein barycenter, in her talk she addressed the convergence of the regularized empirical barycenter of a set of n i.i.d. random probability measures towards their population counterpart and discussed the rate of convergence. This approach is appropriate for the statistical analysis of both discrete and absolutely continuous random measures. Esteban Tabak presented an optimal-transport based methodology is proposed for the explanation of variability in data in which the central idea is to estimate and simulate conditional distributions by mapping them optimally to their Wasserstein barycenter. Sanvesh Srivastava explained how to use Wasserstein barycenters in a tractable

scalable bayesian method and illustrated this approach on a large movie ratings database where this method has superior empirical performance. Codina Cotar gave a long talk on the interplay between MMOT and DFT for repulsive costs of Coulomb or Riesz type, she also presented new results and open problems on next order terms in DFT. Jun Kitagawa, presented a multi-marginal partial transport problem, for which contrarily to the two-marginal case, the active regions do not vary monotonically with the amount of mass to be transported, this is another indication of the significantly different behavior of solutions compared to the two marginal case.

3.4 OT, statistics and data sciences

3.4.1 Wasserstein based statistics

The use of OT criteria is becoming more and more popular in statistics either for inference or testing purposes. This poses new challenges both theoretical and computational so as to lead to tractable methods in which the geometry of OT plays a crucial role. As outlined by Espen Bernton, interest in the Wasserstein distance for statistical inference has been mainly theoretical, due to computational limitations. Bernton's talk aimed at showing the feasibility of the Wasserstein approach for inference in generated models. In purely generative models, one can simulate data given parameters but not necessarily evaluate the likelihood. In Berton's talk, the Wasserstein distances between empirical distributions of observed data and empirical distributions of synthetic data drawn from such models are used to estimate their parameters. Thanks to recent advances in numerical transport, the computation of these distances has become feasible for relatively large data sets, up to controllable approximation errors. Eustasio Del Barrio showed that a CLT holds for the empirical transportation cost under mild moment and smoothness requirements. The limiting distributions are Gaussian and admit a simple description in terms of the optimal transportation potentials.

3.4.2 Bayesian analysis

Integration against an intractable probability measure is among the fundamental challenges of statistical inference, particularly in the Bayesian setting. The starting point of Youssef Marzouk's approach seeks a deterministic coupling of the measure of interest with a tractable "reference" measure (e.g., a standard Gaussian) Yet characterizing such a map—e.g., representing, constructing, and evaluating it—grows challenging in high dimensions. We will present links between the conditional independence structure of the target measure and the existence of certain low-dimensional couplings, induced by transport maps that are sparse or decomposable. Marzouk also described conditions, common in Bayesian inverse problems, under which transport maps have a particular low-rank structure. This analysis not only facilitates the construction of couplings in high-dimensional settings, but also suggests new inference methodologies. For instance, in the context of nonlinear and non-Gaussian state space models, this led to new variational algorithms for filtering, smoothing, and online parameter estimation. These algorithms implicitly characterize—via a transport map—the full posterior distribution of the sequential inference problem using only local operations while avoiding importance sampling or resampling. Sebastian Reich reviewed the use of optimal coupling argument in the design of sequential Monte Carlo methods for data assimilation and sequential Bayesian inference. He also addressed their efficient implementation using the Sinkhorn approximation and second-order corrections. Jean-Micel Loubes talk focused on forecasting a Gaussian process indexed by probability distributions. For this, he provided a family of positive definite kernels built using transportation based distances, gave a probabilistic understanding of these kernels and characterized the corresponding stochastic processes, finally, he addressed the asymptotic properties of the forecast process.

3.5 Machine Learning

Dejan Slepcev's long talk opened the workshop, in this inspiring lecture, Slepcev adressed from a variational perspective the general regression problem of semi-supervised learning: given real-valued labels on a small subset of data recover the function on the whole data set while taking into account the information provided by a large number of unlabeled data points. Objective functionals modeling this regression problem involve terms rewarding the regularity of the function estimate while enforcing agreement with the labels provided. Deep neural networks have achieved significant success in a number of challenging engineering problems.

There is consensus in the community that some form of smoothing of the loss function is needed, an important tool is the stochastic gradient descent for the loss function, which is a non convex function of a very large number of variables. Adam Oberman presented a new PDE approach based on the Hamilton-Jacobi-Bellman equation to address this problem and presented a new algorithm which involves auxiliary convex minimization problems, he explained also how tools from OT can be used to analyze convergence performance of this method. Jianbo Ye, motivated by machine learning applications of OT presented two new numerical methods, the first one is based on a Bregman ADMM approach to solve optimal transport and Wasserstein barycenter and the second one is a simulated annealing approach to solve Wasserstein minimization problems. Rémi Flamary presented an OT based approach to unsupervised domain adaptation problems, where one wants to estimate a prediction function in a given target domain without any labeled sample by exploiting the knowledge available from a source domain where labels are known, this corresponds to the minimization of a generalization bound, and provides an efficient algorithmic solution, for which convergence is proved. The versatility of the approach, both in terms of class of hypothesis or loss functions was demonstrated with real world classification and regression problems, for which the method reaches or surpasses state-of-the-art results. Peyman Mohajerin Esfahani considered stochastic programs where the distribution of the uncertain parameters is only observable through a finite training dataset. Constructing a (Wasserstein) ball in the space of probability distributions centered at the uniform distribution on the training samples, one then seeks decisions that perform best in view of the worst-case distribution within this Wasserstein ball. The state-of-the-art methods for solving the resulting distributionally robust optimization (DRO) problems rely on global optimization techniques, which quickly become computationally excruciating. In Esfahani's talk, he demonstrated that, under mild assumptions, the DRO problems over Wasserstein balls can in fact be reformulated as finite convex programs—in many interesting cases even as tractable linear programs. He further discussed performance guarantees as well as connection to the celebrated regularization techniques in the Machine Learning literature.

4 Outcome of the Meeting

This meeting, the first of this kind, gathered experts from different research communities around OT questions. A first outcome was to spread the main ideas and techniques behind state of the art numerical methods. The second outcome was the identification of new challenges either theoretical or computational posed by OT methods in the fields of statistics and machine learning. All participants were very enthusiastic about the scientific content of this meeting, the cross-fertilization between the different subjects and their impact on their own research subjects. It is the belief of the organizers that this meeting is the first of a long series of forthcoming scientific events which may have a significant impact on the development of OT methods in the aforementioned fields.

References

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

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] Vladimir Arnold. Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In *Annales de l'institut Fourier*, volume 16, pages 319–361, 1966.

[4] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[5] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[6] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[7] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[8] Yann Brenier. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the euler equations. *Communications on pure and applied mathematics*, 52(4):411–452, 1999.

[9] Yann Brenier. Generalized solutions and hydrostatic approximation of the euler equations. *Physica D: Nonlinear Phenomena*, 237(14):1982–1988, 2008.

[10] Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Phys. Rev. A*, 85:062502, Jun 2012.

[11] G. Carlier and I. Ekeland. Matching for teams. *Econom. Theory*, 42(2):397–418, 2010.

[12] C. Cotar, G. Friesecke, and C. Kluppelberg. Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.

[13] Codina Cotar, Gero Friesecke, and Brendan Pass. Infinite-body optimal transport with coulomb cost. *Calculus of Variations and Partial Differential Equations*, pages 1–26, 2013.

[14] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML), JMLR W&CP*, volume 32, 2014.

[15] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[16] Simone Di Marino, Augusto Gerolin, and Luca Nenna. Optimal transportation theory with repulsive costs. *arXiv preprint arXiv:1506.04565*, 2015.

[17] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735, 1989.

[18] T. Gallouët and Q. Mérigot. A Lagrangian scheme for the incompressible Euler equation using optimal transport. *ArXiv e-prints*, May 2016.

[19] Nassif Ghoussoub and Bernard Maurey. Remarks on multi-marginal symmetric Monge-Kantorovich problems. *Discrete Contin. Dyn. Syst.*, 34(4):1465–1480, 2014.

[20] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[21] J. Kitagawa, Q. Mérigot, and B. Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *ArXiv e-prints*, March 2016.

[22] Bruno Lévy. A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM Math. Model. Numer. Anal.*, 49(6):1693–1715, 2015.

[23] Quentin Mérigot and Édouard Oudet. Discrete optimal transport: complexity, geometry and applications. *Discrete Comput. Geom.*, 55(2):263–283, 2016.

[24] B. Pass. Uniqueness and Monge solutions in the multimarginal optimal transportation problem. *SIAM Journal on Mathematical Analysis*, 43(6):2758–2775, 2011.

[25] Brendan Pass. Multi-marginal optimal transport and multi-agent matching problems: uniqueness and structure of solutions. *Discrete Contin. Dyn. Syst.*, 34(4):1623–1639, 2014.

[26] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[27] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

[28] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly*, 74:402–405, 1967.

[29] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[30] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.