EDUSCHOOL
ACADEMIC RESEARCH & PUBLISHERS

INTERNATIONAL JOURNAL OF PURE SCIENCE RESEARCH STUDIES (IJPSRS)
(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online:

Optimal Transport: Theory, Numerical Methods and Applications in Applied Mathematics

Assanu Augustine

Assistant Professor, Department of Mathematics, Marian College Kuttikkanam (Autonomous), Kottayam, India

Article Information

Received: 4th May 2026
Received in revised form: 6th June 2026
Accepted: 8th July 2026
Available online: 14th August 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.5281/zenodo.21916551

Abstract

Optimal transport asks for the cheapest way of rearranging one distribution of mass into another, and the answer supplies a metric on probability measures that respects the geometry of the underlying space. This review traces the subject from the formulation posed by Monge in 1781, through the linear relaxation introduced by Kantorovich, to the characterisation of optimal maps as gradients of convex functions. The resulting Wasserstein distances behave differently from divergences based on pointwise comparison of densities: they remain finite and informative for measures with disjoint support, and the shortest paths they induce displace mass rather than fade one profile into another. The dynamic formulation connects these paths to a fluid mechanical problem and identifies several diffusion equations as gradient flows of familiar energies. Numerical work has changed the field. Entropic regularisation reduces the linear program to matrix scaling iterations that are simple, parallel and differentiable, at the cost of blurring the transport plan by an amount controlled by the regularisation parameter. Semi-discrete and sliced methods offer alternatives when accuracy matters more than speed. Applications now include image retrieval, generative modelling, domain adaptation and the reconstruction of developmental trajectories from single-cell measurements.

Keywords: Optimal Transport, Wasserstein Distance, Kantorovich Duality, Sinkhorn Algorithm, Gradient Flow.

1. INTRODUCTION

The problem has a concrete origin. Monge asked how to move a pile of soil into an embankment of prescribed shape while minimising the total work, where the work of moving a grain equals its mass times the distance travelled .[1] Posed as a search over maps that push one measure onto another, the question resisted analysis for a century and a half, largely because such a map need not exist. If the source is a single atom and the target is spread over two locations, no map can split it.

Kantorovich removed the obstacle by replacing maps with couplings, that is, joint measures whose marginals are the prescribed source and target .[2] Mass may then be split, the feasible set is convex and compact in a suitable topology, and the problem becomes an infinite-dimensional linear program with a well-behaved dual. The relaxation is not merely a technical convenience: it produced the duality theory on which most later analysis rests, and it was developed alongside the work on resource allocation for which its author later shared a Nobel prize in economics.

Interest broadened considerably once it became clear that the minimal cost defines a metric on probability measures with useful geometric properties, and that this metric linearises certain nonlinear partial differential equations. Two comprehensive treatments set out the resulting theory.[3,4] This paper reviews the formulations, the structure of optimal solutions, the algorithms that made large problems tractable, and the applications that followed.

2. FORMULATIONS AND THE STRUCTURE OF SOLUTIONS

Let two probability measures be given on a metric space together with a cost function specifying the price of moving a unit of mass from one point to another. The Monge problem seeks a measurable map, pushing the first measure onto the second, that minimises the integrated cost. The Kantorovich problem seeks instead a coupling of the two measures minimising the integral of the cost against that coupling. Every map induces a coupling, so the second problem has a value no larger than the first, and under mild conditions the two values coincide.

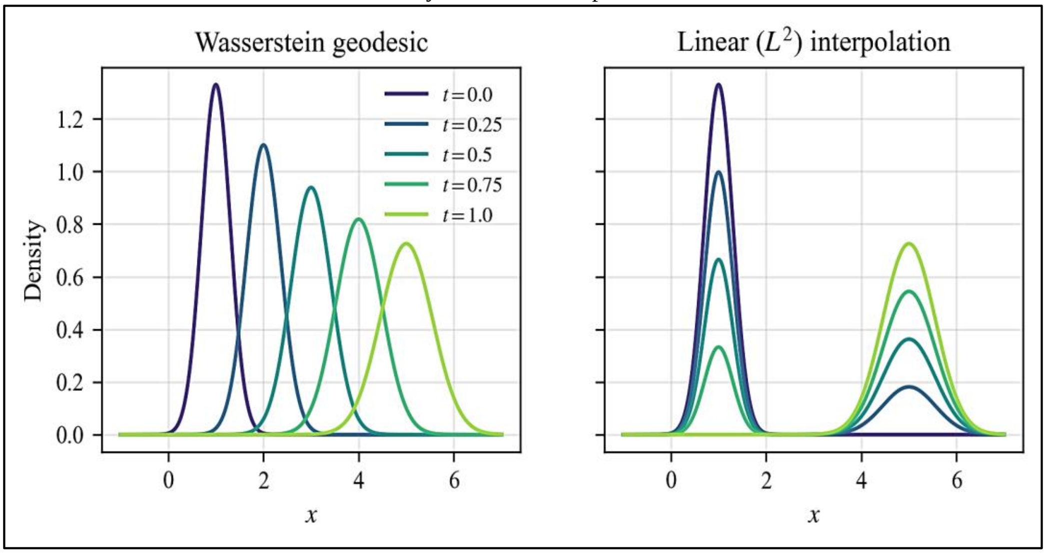
Duality is what makes the relaxed problem workable. The dual seeks a pair of potentials whose sum never exceeds the cost and which maximise the difference of their integrals against the two measures. At the optimum the constraint is active exactly on the support of the optimal coupling, a complementary slackness condition with a clear interpretation as a price system: the potentials are the prices at which mass would be bought at the source and sold at the target.

For quadratic cost on Euclidean space the structure of solutions is completely understood. Brenier proved that if the source measure is absolutely continuous, then the optimal coupling is induced by a map, that map is the gradient of a convex function, and it is unique almost everywhere .[5] The result identifies optimal transport with the polar factorisation of vector fields and connects it to the Monge-Ampère equation, since the Jacobian equation satisfied by the gradient of the potential is of that type.

The minimal cost raised to a suitable power defines the Wasserstein distances. Their most useful feature is that they compare measures through the geometry of the space rather than pointwise. Two narrow bumps separated by a small distance are close in this metric and far apart in any divergence based on overlap of densities. The associated shortest paths reflect the same principle: McCann showed that interpolating between two measures along a Wasserstein geodesic displaces mass continuously.[6] Fig. 1 contrasts this displacement interpolation with ordinary linear interpolation of densities, which creates a spurious second mode instead of moving the first.

One case is worth stating explicitly because it underlies several algorithms. On the real line with a convex cost the optimal map is the monotone rearrangement, obtained by composing the inverse distribution function of the target with the distribution function of the source. The transport cost then reduces to an integral of the distance between quantile functions, so the problem is solved by sorting rather than by optimisation. The observation explains why one-dimensional projections are useful in practice, and it supplies the closed form used in Fig. 1, where the geodesic between two Gaussian measures interpolates the mean and the standard deviation linearly.

Fig 1: Interpolation between two Gaussian densities along a Wasserstein geodesic (left) and by linear combination (right). The geodesic translates and rescales the profile, while the linear path passes through a bimodal density that neither endpoint resembles.



3. THE DYNAMIC FORMULATION AND GRADIENT FLOWS

Benamou and Brenier recast the quadratic problem as a control problem in which a density evolves under a velocity field subject to the continuity equation, and the cost becomes the integrated kinetic energy .[7] The reformulation converts a static assignment into a fluid mechanical problem, which is convex in the pair consisting of density and momentum and can therefore be attacked by standard splitting methods. It also makes the geodesic structure explicit and provides the natural setting for a Riemannian interpretation of the space of measures.

The consequence with the widest reach concerns evolution equations. Jordan, Kinderlehrer and Otto showed that the Fokker-Planck equation is the gradient flow of the free energy with respect to the quadratic Wasserstein metric, in the precise sense that an implicit scheme alternating between minimising energy and penalising transport distance converges to its solution .[8] The same construction applies to the porous medium equation and to a range of aggregation and diffusion models. The general theory of such flows in metric spaces was developed subsequently, and it now provides both existence proofs and structure-preserving numerical schemes for equations that are difficult to treat by other means.[4]

4. COMPUTATION

Discretising both measures on n points turns the Kantorovich problem into a linear assignment problem. Network simplex and auction algorithms solve it exactly, with worst-case cost that grows roughly as the cube of n, which becomes impractical well before the sizes encountered in imaging or machine learning.

Entropic regularisation changed this. Adding a multiple of the entropy of the coupling makes the objective strictly convex, and the optimality conditions then state that the solution is a diagonal rescaling of a fixed matrix obtained by exponentiating the negative cost. Alternating the two scalings is the Sinkhorn iteration, which involves only matrix-vector products and therefore runs efficiently on parallel hardware.[9] The regularised objective is also differentiable with respect to its inputs, which is why the method was adopted quickly in learning applications.[10] Convergence to a fixed accuracy is achieved in a number of operations that is near linear in the number of entries of the cost matrix.[11]

The price is blurring. Fig. 2 shows plans computed by Sinkhorn iteration for the same pair of one-dimensional densities at three values of the regularisation parameter. At the smallest value the plan concentrates near the graph of the monotone map that solves the exact problem; as the parameter grows the plan spreads and eventually approaches the product of the marginals. Small values recover accuracy but require more iterations and invite numerical underflow, which is handled in practice by working with logarithms.

Other approaches avoid regularisation altogether. Semi-discrete methods exploit the fact that when the target is a finite set of atoms the dual problem is a smooth concave maximisation over their weights, whose gradient is computed from a power diagram; the resulting schemes give exact transport at a cost dominated by geometric computation.[12] Sliced methods project both measures onto random lines, where the one-dimensional problem is solved by sorting, and average the results.[13] Convolutional methods replace the cost matrix by heat kernel applications on a grid or mesh and never form it explicitly.[14] Table 1 compares these options.

Fig 2: Entropic transport plans between two mixtures of Gaussians, computed by Sinkhorn iteration at three values of the regularisation parameter. Support concentrates on the monotone map as the parameter decreases.

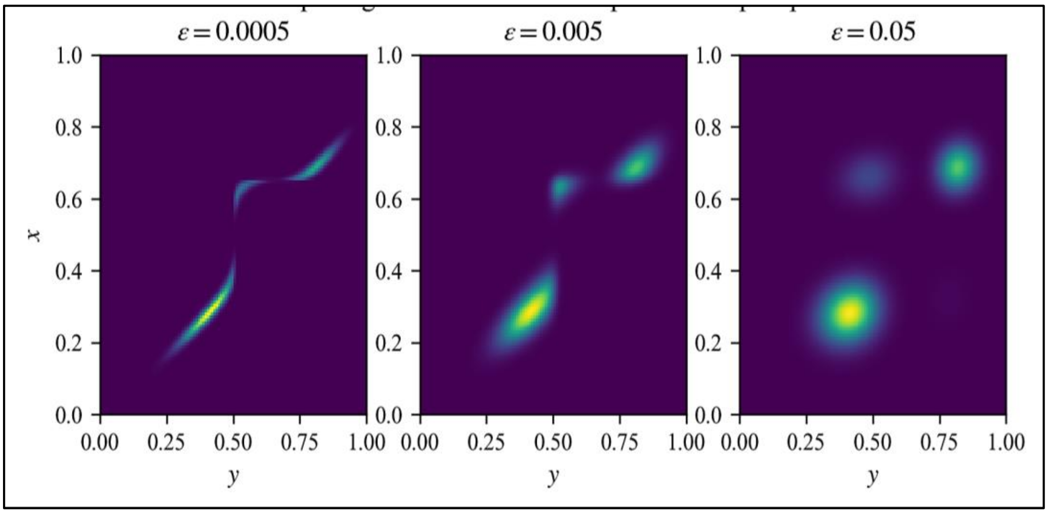


Table 1. Numerical approaches to discrete optimal transport Costs are the scalings usually quoted for each method rather than worst-case bounds.

Method	Typical cost for n points	Output	Suited to
Network simplex	Order $n^3 \log n$	Exact plan	Small problems, exact solutions
Sinkhorn iteration	Order n^2 per iteration	Blurred plan	Large dense problems, learning [9,11]
Semi-discrete	Order $n \log n$ per iteration in low dimension	Exact map per	Continuous source, discrete target [12]
Sliced projection	Order $Ln \log n$ for L direction	Distance only	Fast comparison of point clouds [13]
Convolutional	Linear in grid size	Blurred plan	Measures on grids and meshes [14]

5. STATISTICAL BEHAVIOUR IN HIGH DIMENSION

A practical question is how well the transport cost between two measures can be estimated from samples. The answer is discouraging in high dimension: the expected error of the empirical Wasserstein distance decreases as the sample size raised to the power minus one over the dimension, so the sample size required for a fixed accuracy grows exponentially.[15] This rate reflects the difficulty of covering a high-dimensional space rather than any weakness of the estimator, and it improves when the measures concentrate near a low-dimensional set.

Entropic regularisation improves the statistical picture as well as the computational one. The regularised divergence converges at a rate governed by the regularisation parameter rather than by the dimension, which is one reason regularised quantities are preferred in learning applications even when exact computation would be affordable.[16] The bias introduced by regularisation is partly removed by the debiased form, which subtracts the self-transport terms.

6. APPLICATIONS

Image retrieval provided an early application. Rubner and colleagues compared colour and texture histograms using what they called the earth mover's distance, which is the transport cost with a ground metric on feature space, and reported retrieval performance better than bin-by-bin comparison.[17] The advantage comes from the same property that distinguishes the metric theoretically: neighbouring bins are treated as similar.

In machine learning the metric was adopted for generative modelling. Training a generator by minimising a Wasserstein distance to the data distribution, with the dual potential represented by a constrained network, avoids the vanishing gradients that arise when the model and the data have disjoint support, a situation ordinary divergences cannot distinguish from complete mismatch.[18] Transport also underpins domain adaptation, where a classifier trained on one dataset is applied to another after aligning the two distributions.[19]

A striking recent use is in developmental biology. Single-cell sequencing destroys the cell it measures, so a time course yields unmatched samples from a sequence of distributions rather than trajectories. Treating consecutive time points as source and target of a transport problem recovers probable ancestor and descendant relationships, and applying this to reprogramming experiments identified the developmental routes taken by cells that succeed and by those that do not.[20] The inference relies on the assumption that cells move a small distance in expression space between sampling times, which is testable but not automatic. Barycentres in Wasserstein space provide the corresponding notion of averaging several distributions, and the applied literature collected in reference gives further worked examples.[21,22]

7. EXTENSIONS BEYOND THE BALANCED PROBLEM

Two restrictions in the classical setting limit its use, and both have been relaxed.

The first is that the two measures must have equal total mass. Many applications violate this: cells divide and die between sampling times, image features appear and disappear, and measurement noise creates spurious mass. Unbalanced formulations replace the hard marginal constraints by penalties, usually a divergence between the marginals of the coupling and the prescribed measures, so that mass may be created or destroyed at a price. The scaling algorithms used for the balanced entropic problem extend to this setting with little modification, which is one reason the formulation was adopted quickly .[23] The penalty parameter controls the trade-off between moving mass and discarding it, and results depend on it in ways that are not always reported.

The second restriction is that both measures must live in the same space, since otherwise the cost of moving a point to another is undefined. Comparing a shape in three dimensions with a graph, or two point clouds recorded in incomparable coordinates, requires a different construction. The Gromov-Wasserstein distance compares the internal distance structures instead of the points themselves, minimising over couplings the discrepancy between pairwise distances in the two spaces.[24] The resulting problem is quadratic and non-convex, so algorithms find local minima, but the formulation gives a principled way to match objects that no rigid alignment relates. It is now used for matching single-cell datasets from different measurement technologies and for correspondence problems in geometry processing.

Both extensions inherit the computational advantages of entropic regularisation and both inherit its bias. Reported distances are therefore comparable only when the regularisation parameter, the penalty weights and the stopping rule are stated, and the literature is not consistent on this point.

8. CONCLUSION

Optimal transport has moved from a specialised topic in analysis to a working tool in computation and statistics within about two decades. The theory is settled for the quadratic cost on Euclidean space, where existence, uniqueness and the structure of the optimal map are known, and the metric it defines has a geometric interpretation that other comparisons between distributions lack. Entropic regularisation removed the computational barrier at the price of a controllable bias, and the alternatives available when that bias is unacceptable are now reasonably mature. The main open difficulty is statistical: estimating transport costs from samples remains expensive in high dimension, and the low-dimensional structure that makes applications work is usually assumed rather than verified. Methods that exploit such structure explicitly are the natural direction for further work.

REFERENCES

1. Monge G. Mémoire sur la théorie des déblais et des remblais. In: Histoire de l'Académie Royale des Sciences de Paris. Paris: Imprimerie Royale; 1781. p. 666–704.
2. Kantorovich LV. On the translocation of masses. Dokl Akad Nauk SSSR. 1942;37(7-8):199–201.
3. Villani C. Optimal transport: old and new. Berlin: Springer; 2009.
4. Ambrosio L, Gigli N, Savaré G. Gradient flows in metric spaces and in the space of probability measures. 2nd ed. Basel: Birkhäuser; 2008.
5. Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Commun Pure Appl Math. 1991;44(4):375–417.
6. McCann RJ. A convexity principle for interacting gases. Adv Math. 1997;128(1):153–79.
7. Benamou JD, Brenier Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer Math. 2000;84(3):375–93.
8. Jordan R, Kinderlehrer D, Otto F. The variational formulation of the Fokker-Planck equation. SIAM J Math Anal. 1998;29(1):1–17.
9. Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. Adv Neural Inf Process Syst. 2013;26:2292–300.
10. Peyré G, Cuturi M. Computational optimal transport. Found Trends Mach Learn. 2019;11(5-6):355–607.
11. Altschuler J, Weed J, Rigollet P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. Adv Neural Inf Process Syst. 2017;30:1964–74.
12. Mérigot Q. A multiscale approach to optimal transport. Comput Graph Forum. 2011;30(5):1583–92.
13. Bonneel N, Rabin J, Peyré G, Pfister H. Sliced and Radon Wasserstein barycenters of measures. J Math Imaging Vis. 2015;51(1):22–45.
14. Solomon J, de Goes F, Peyré G, et al. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. ACM Trans Graph. 2015;34(4):66.
15. Weed J, Bach F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli. 2019;25(4A):2620–48.
16. Genevay A, Chizat L, Bach F, Cuturi M, Peyré G. Sample complexity of Sinkhorn divergences. Proc Mach Learn Res. 2019;89:1574–83.
17. Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vis. 2000;40(2):99–121.
18. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. Proc Mach Learn Res. 2017;70:214–23.
19. Courty N, Flamary R, Tuia D, Rakotomamonjy A. Optimal transport for domain adaptation. IEEE Trans Pattern Anal Mach Intell. 2017;39(9):1853–65.
20. Schiebinger G, Shu J, Tabaka M, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell. 2019;176(4):928–43.
21. Agueh M, Carlier G. Barycenters in the Wasserstein space. SIAM J Math Anal. 2011;43(2):904–24.
22. Santambrogio F. Optimal transport for applied mathematicians. Basel: Birkhäuser; 2015.
23. Chizat L, Peyré G, Schmitzer B, Vialard FX. Scaling algorithms for unbalanced optimal transport problems. Math Comput. 2018;87(314):2563–609.
24. Mémoli F. Gromov-Wasserstein distances and the metric approach to object matching. Found Comput Math. 2011;11(4):417–87.