Wasserstein metric

From Wikipedia, the free encyclopedia

In mathematics, the **Wasserstein distance** or **Kantorovich–Rubinstein metric** is a distance function defined between probability distributions on a given metric space M. It is named after Leonid Vaseršteĭn.

Intuitively, if each distribution is viewed as a unit amount of earth (soil) piled on M, the metric is the minimum "cost" of turning one pile into the other, which is assumed to be the amount of earth that needs to be moved times the mean distance it has to be moved. This problem was first formalised by Gaspard Monge in 1781. Because of this analogy, the metric is known in computer science as the earth mover's distance.

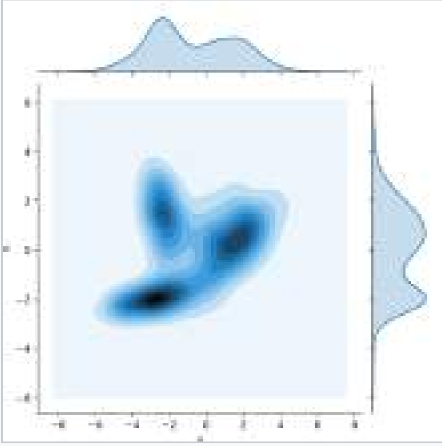
The name "Wasserstein distance" was coined by R. L. Dobrushin in 1970, after learning of it in the work of Leonid Vaseršteĭn on Markov processes describing large systems of automata[1] (Russian, 1969). However the metric was first defined by Leonid Kantorovich in *The Mathematical Method of Production Planning and Organization*[2] (Russian original 1939) in the context of optimal transport planning of goods and materials. Some scholars thus encourage use of the terms "Kantorovich metric" and "Kantorovich distance". Most English-language publications use the German spelling "Wasserstein" (attributed to the name "Vaseršteĭn" (Russian: Васерштейн) being of Yiddish origin).

Definition [edit]

Let (M,d) be a metric space that is a Polish space. For $p \in [1,+\infty]$, the Wasserstein p-distance between two probability measures μ and ν on M with finite p-moments is

$W_p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\mathbf{E}_{(x,y)\sim\gamma} d(x,y)^p\right)^{1/p}$, where $\Gamma(\mu,\nu)$ is the set of all couplings of μ and ν; $W_\infty(\mu,\nu)$ is defined to be $\lim_{p\to+\infty} W_p(\mu,\nu)$ and corresponds to a supremum norm. Here, a coupling γ is a joint probability measure on $M \times M$ whose marginals are μ and ν on the first and second factors, respectively. This means that for all measurable $A \subset M$, it fulfills $\gamma(A \times M) = \mu(A)$ and $\gamma(M \times A) = \nu(A)$.

Intuition and connection to optimal transport [edit]



Two one-dimensional distributions μ and ν, plotted on the x and y axes, and one possible joint distribution that defines a transport plan between them. The joint distribution/transport plan is not unique

One way to understand the above definition is to consider the optimal transport problem. That is, for a distribution of mass $\mu(x)$ on a space X, we wish to transport the mass in such a way that it is transformed into the distribution $\nu(x)$ on the same space; transforming the 'pile of earth' μ to the pile ν. This problem only makes sense if the pile to be created has the same mass as the pile to be moved; therefore without loss of generality assume that μ and ν are probability distributions containing a total mass of 1. Assume also that there is given some cost function

$$c(x, y) \geq 0$$

that gives the cost of transporting a unit mass from the point x to the point y. A transport plan to move μ into ν can be described by a function $\gamma(x, y)$ which gives the amount of mass to move from x to y. You can imagine the task as the need to move a pile of earth of shape μ to the hole in the ground of shape ν such that at the end, both the pile of earth and the hole in the ground completely vanish. In order for this plan to be meaningful, it must satisfy the following properties:

1. the amount of earth moved out of point x must equal the amount that was there to begin with; that is,

$$\int \gamma(x, y)\, \mathrm{d}y = \mu(x),$$

   and

2. the amount of earth moved into point y must equal the depth of the hole that was there at the beginning; that is,

$$\int \gamma(x, y)\, \mathrm{d}x = \nu(y).$$

That is, that the total mass moved *out of* an infinitesimal region around x must be equal to $\mu(x)\mathrm{d}x$ and the total mass moved *into* a region around y must be $\nu(y)\mathrm{d}y$. This is equivalent to the requirement that γ be a joint probability distribution with marginals μ and ν. Thus, the infinitesimal mass transported from x to y is $\gamma(x,y)\,\mathrm{d}x\,\mathrm{d}y$, and the cost of moving is $c(x,y)\gamma(x,y)\,\mathrm{d}x\,\mathrm{d}y$, following the definition of the cost function. Therefore, the total cost of a transport plan γ is

$$\iint c(x,y)\gamma(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int c(x,y)\,\mathrm{d}\gamma(x,y).$$

The plan γ is not unique; the optimal transport plan is the plan with the minimal cost out of all possible transport plans. As mentioned, the requirement for a plan to be valid is that it is a joint distribution with marginals μ and ν; letting Γ denote the set of all such measures as in the first section, the cost of the optimal plan is

$$C = \inf_{\gamma\in\Gamma(\mu,\nu)} \int c(x,y)\,\mathrm{d}\gamma(x,y).$$

If the cost of a move is simply the distance between the two points, then the optimal cost is identical to the definition of the W_1 distance.

Examples [edit]

Point masses [edit]

Deterministic distributions [edit]

Let $\mu_1 = \delta_{a_1}$ and $\mu_2 = \delta_{a_2}$ be two degenerate distributions (i.e. Dirac delta distributions) located at points a_1 and a_2 in $\mathbb{R}$. There is only one possible coupling of these two measures, namely the point mass $\delta_{(a_1,a_2)}$ located at $(a_1, a_2) \in \mathbb{R}^2$. Thus, using the usual absolute value function as the distance function on $\mathbb{R}$, for any $p \geq 1$, the p-Wasserstein distance between μ_1 and μ_2 is

$$W_p(\mu_1,\mu_2) = |a_1 - a_2|.$$

By similar reasoning, if $\mu_1 = \delta_{a_1}$ and $\mu_2 = \delta_{a_2}$ are point masses located at points a_1 and a_2 in $\mathbb{R}^n$, and we use the usual Euclidean norm on $\mathbb{R}^n$ as the distance function, then

$$W_p(\mu_1,\mu_2) = \|a_1 - a_2\|_2.$$

Empirical distributions [edit]

One dimension [edit]

If P is an empirical measure with samples $X_1,\ldots,X_n$ and Q is an empirical measure with samples $Y_1,\ldots,Y_n$, the distance is a simple function of the order statistics:

$$W_p(P,Q) = \left(\frac{1}{n}\sum_{i=1}^{n}\|X_{(i)} - Y_{(i)}\|^p\right)^{1/p}.$$

Higher dimensions [edit]

If P and Q are empirical distributions, each based on n observations, then

$$W_p(P,Q) = \inf_\pi \left(\frac{1}{n} \sum_{i=1}^{n} \|X_i - Y_{\pi(i)}\|^p \right)^{1/p},$$

where the infimum is over all permutations π of n elements. This is a linear assignment problem, and can be solved by the Hungarian algorithm in cubic time.

Normal distributions [edit]

Let $\mu_1 = \mathcal{N}(m_1, C_1)$ and $\mu_2 = \mathcal{N}(m_2, C_2)$ be two non-degenerate Gaussian measures (i.e. normal distributions) on $\mathbb{R}^n$, with respective expected values m_1 and $m_2 \in \mathbb{R}^n$ and symmetric positive semi-definite covariance matrices C_1 and $C_2 \in \mathbb{R}^{n\times n}$. Then,[3] with respect to the usual Euclidean norm on $\mathbb{R}^n$, the 2-Wasserstein distance between μ_1 and μ_2 is

$$W_2(\mu_1, \mu_2)^2 = \|m_1 - m_2\|_2^2 + \mathrm{trace}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right).$$

where $C^{1/2}$ denotes the principal square root of C. Note that the second term (involving the trace) is precisely the (unnormalised) Bures metric between C_1 and C_2. This result generalises the earlier example of the Wasserstein distance between two point masses (at least in the case $p = 2$), since a point mass can be regarded as a normal distribution with covariance matrix equal to zero, in which case the trace term disappears and only the term involving the Euclidean distance between the means remains.

One-dimensional distributions [edit]

Let $\mu_1, \mu_2 \in P_p(\mathbb{R})$ be probability measures on $\mathbb{R}$, and denote their cumulative distribution functions by $F_1(x)$ and $F_2(x)$. Then the transport problem has an analytic solution: Optimal transport preserves the order of probability mass elements, so the mass at quantile q of μ_1 moves to quantile q of μ_2. Thus, the p-Wasserstein distance between μ_1 and μ_2 is

$$W_p(\mu_1, \mu_2) = \left(\int_0^1 \left|F_1^{-1}(q) - F_2^{-1}(q)\right|^p \mathrm{d}q \right)^{1/p},$$

where F_1^{-1} and F_2^{-1} are the quantile functions (inverse CDFs). In the case of $p = 1$, a change of variables leads to the formula[4]

$$W_1(\mu_1, \mu_2) = \int_{\mathbb{R}} |F_1(x) - F_2(x)| \ \mathrm{d}x.$$

Applications [edit]

The Wasserstein metric is a natural way to compare the probability distributions of two variables *X* and *Y*, where one variable is derived from the other by small, non-uniform perturbations (random or deterministic).

In computer science, for example, the metric W_1 is widely used to compare discrete distributions, *e.g.* the color histograms of two digital images; see earth mover's distance for more details.

In their paper 'Wasserstein GAN', Arjovsky et al.[5] use the Wasserstein-1 metric as a way to improve the original framework of generative adversarial networks (GAN), to alleviate the vanishing gradient and the mode collapse issues. The special case of normal distributions is used in a Frechet inception distance.

The Wasserstein metric has a formal link with Procrustes analysis, with application to chirality measures,[6] and to shape analysis.[7]

In computational biology, Wasserstein metric can be used to compare between persistence diagrams of cytometry datasets.[8]

The Wasserstein metric also has been used in inverse problems in geophysics.[9]

The Wasserstein metric is used in integrated information theory to compute the difference between concepts and conceptual structures.[10]

The Wasserstein metric and related formulations have also been used to provide a unified theory for shape observable analysis in high energy and collider physics datasets.[11][12]

Properties [edit]

Metric structure [edit]

It can be shown that W_p satisfies all the axioms of a metric on the Wasserstein space $\boldsymbol{P}_p(M)$ consisting of all Borel probability measures on M having finite pth moment. Furthermore, convergence with respect to W_p is equivalent to the usual weak convergence of measures plus convergence of the first pth moments.[13]

Dual representation of W_1 [edit]

The following dual representation of W_1 is a special case of the duality theorem of Kantorovich and Rubinstein (1958): when μ and ν have bounded support,

$$W_1(\mu,\nu) = \sup\left\{ \int_M f(x)\,\mathrm{d}(\mu-\nu)(x) \,\middle|\, \text{continuous } f : M \to \mathbb{R}, \operatorname{Lip}(f) \leq 1 \right\},$$

where Lip(f) denotes the minimal Lipschitz constant for f. This form shows that W_1 is an integral probability metric.

Compare this with the definition of the Radon metric:

$$\rho(\mu,\nu) := \sup\left\{ \int_M f(x)\,\mathrm{d}(\mu-\nu)(x) \,\middle|\, \text{continuous } f : M \to [-1,1] \right\}.$$

If the metric d of the metric space (M,d) is bounded by some constant C, then

$$2W_1(\mu,\nu) \le C\rho(\mu,\nu),$$

and so convergence in the Radon metric (identical to **total variation convergence** when M is a Polish space) implies convergence in the Wasserstein metric, but not vice versa.

Proof [edit]

The following is an intuitive proof which skips over technical points. A fully rigorous proof is found in.[14]

Discrete case: When M is discrete, solving for the 1-Wasserstein distance is a problem in linear programming:

$$\begin{cases} \min_\gamma \sum_{x,y} c(x,y)\gamma(x,y) \\ \sum_y \gamma(x,y) = \mu(x) \\ \sum_x \gamma(x,y) = \nu(y) \\ \gamma \ge 0 \end{cases}$$

where $c : M \times M \to [0,\infty)$ is a general "cost function".

By carefully writing the above equations as matrix equations, we obtain its dual problem:[15]

$$\begin{cases} \max_{f,g} \sum_x \mu(x)f(x) + \sum_y \nu(y)g(y) \\ f(x) + g(y) \le c(x,y) \end{cases}$$

and by the duality theorem of linear programming, since the primal problem is feasible and bounded, so is the dual problem, and the minimum in the first problem equals the maximum in the second problem. That is, the problem pair exhibits *strong duality*.

For the general case, the dual problem is found by converting sums to integrals:

$$\begin{cases} \sup_{f,g} \mathbb{E}_{x\sim\mu}[f(x)] + \mathbb{E}_{y\sim\nu}[g(y)] \\ f(x) + g(y) \le c(x,y) \end{cases}$$

and the *strong duality* still holds. This is the **Kantorovich duality theorem**. Cédric Villani recounts the following interpretation from Luis Caffarelli:[16]

> Suppose you want to ship some coal from mines, distributed as μ, to factories, distributed as ν. The cost function of transport is c. Now a shipper comes and offers to do the transport for you. You would pay him $f(x)$ per coal for loading the coal at x, and pay him $g(y)$ per coal for unloading the coal at y.
>
> For you to accept the deal, the price schedule must satisfy $f(x) + g(y) \le c(x,y)$. The Kantorovich duality states that the shipper can make a price schedule that makes you pay almost as much as you would ship yourself.

This result can be pressed further to yield:

Theorem (Kantorovich-Rubenstein duality) — When the probability space Ω is a metric space, then for any fixed $K > 0$,

$$W_1(\mu,\nu) = \frac{1}{K} \sup_{\|f\|_L \le K} \mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{y\sim\nu}[f(y)]$$

where $\|\cdot\|_L$ is the Lipschitz norm.

Proof

It suffices to prove the case of $K = 1$. Start with

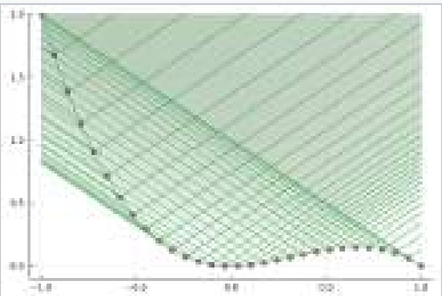
$$W_1(\mu,\nu) = \sup_{f(x)+g(y)\le d(x,y)} \mathbb{E}_{x\sim\mu}[f(x)] + \mathbb{E}_{y\sim\nu}[g(y)].$$

Then, for any choice of g, one can push the term higher by setting $f(x) = \inf_y d(x,y) - g(y)$, making it an infimal convolution of $-g$ with a cone. This implies $f(x) - f(y) \le d(x,y)$ for any x, y, that is, $\|f\|_L \le 1$.

Thus,

$$\begin{aligned} W_1(\mu,\nu) &= \sup_g \sup_{f(x)+g(y)\le d(x,y)} \mathbb{E}_{x\sim\mu}[f(x)] + \mathbb{E}_{y\sim\nu}[g(y)] \\ &= \sup_g \sup_{\|f\|_L\le 1, f(x)+g(y)\le d(x,y)} \mathbb{E}_{x\sim\mu}[f(x)] + \mathbb{E}_{y\sim\nu}[g(y)] \\ &= \sup_{\|f\|_L\le 1} \sup_{g, f(x)+g(y)\le d(x,y)} \mathbb{E}_{x\sim\mu}[f(x)] + \mathbb{E}_{y\sim\nu}[g(y)]. \end{aligned}$$

Next, for any choice of $\|f\|_L \le 1$, g can be optimized by setting $g(y) = \inf_x d(x,y) - f(x)$. Since $\|f\|_L \le 1$, this implies $g(y) = -f(y)$.



Infimal convolution of a cone with a curve. Note how the lower envelope has slope ≤ 1, and how the lower envelope is *equal* to the curve on the parts where the curve itself has slope ≤ 1.

The two infimal convolution steps are visually clear when the probability space is $\mathbb{R}$.

For notational convenience, let $\square$ denote the infimal convolution operation.

For the first step, where we used $f = \operatorname{cone} \square (-g)$, plot out the curve of $-g$, then at each point, draw a cone of slope 1, and take the lower envelope of the cones as f, as shown in the diagram, then f cannot increase with slope larger than 1. Thus all its secants have slope $\left|\frac{f(x)-f(y)}{x-y}\right| \leq 1$.

For the second step, picture the infimal convolution $\operatorname{cone} \square (-f)$, then if all secants of f have slope at most 1, then the lower envelope of $\operatorname{cone} \square (-f)$ are just the cone-apices themselves, thus $\operatorname{cone} \square (-f) = -f$.

1D Example. When both μ, ν are distributions on $\mathbb{R}$, then integration by parts give

$$\mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{y\sim\nu}[f(y)] = \int f'(x)(F_\nu(x) - F_\mu(x))\,\mathrm{d}x,$$

thus

$$f(x) = K \cdot \operatorname{sign}(F_\nu(x) - F_\mu(x)).$$

Fluid mechanics interpretation of W_2 [edit]

Benamou & Brenier found a dual representation of W_2 by fluid mechanics, which allows efficient solution by convex optimization.[17][18]

Given two probability densities p, q on $\mathbb{R}^n$,

$$W_2^2(p,q) = \min_{\mathbf{v}} \int_0^1 \int_{\mathbb{R}^n} \|\mathbf{v}(\mathbf{x},t)\|^2 \rho(\mathbf{x},t)\, d\mathbf{x}\, dt$$

where $\mathbf{v}$ ranges over velocity fields driving the continuity equation with boundary conditions on the fluid density field:

$$\dot\rho + \nabla\cdot(\rho\mathbf{v}) = 0 \quad \rho(\cdot,0) = p,\ \rho(\cdot,1) = q$$

That is, the mass should be conserved, and the velocity field should transport the probability distribution p to q during the time interval $[0,1]$.

Equivalence of W_2 and a negative-order Sobolev norm [edit]

Under suitable assumptions, the Wasserstein distance W_2 of order two is Lipschitz equivalent to a negative-order homogeneous Sobolev norm. More precisely, if we take M to be a connected Riemannian manifold equipped with a positive measure π, then we may define for $f\colon M \to \mathbb{R}$ the seminorm

$$\|f\|_{\dot H^1(\pi)}^2 = \int_M \|\nabla f(x)\|^2\, \pi(\mathrm{d}x)$$

and for a signed measure μ on M the dual norm

$$\|\mu\|_{\dot{H}^{-1}(\pi)} = \sup \left\{ |\langle f, \mu \rangle| \,\middle|\, \|f\|_{\dot{H}^1(\pi)} \leq 1 \right\}.$$

Then any two probability measures μ and ν on M satisfy the upper bound [19]

$$W_2(\mu, \nu) \leq 2 \, \|\mu - \nu\|_{\dot{H}^{-1}(\pi)}.$$

In the other direction, if μ and ν each have densities with respect to the standard volume measure on M that are both bounded above by some $0 < C < \infty$, and M has non-negative Ricci curvature, then [20] [21]

$$\|\mu - \nu\|_{\dot{H}^{-1}(\pi)} \leq \sqrt{C} \, W_2(\mu, \nu).$$

Separability and completeness [edit]

For any $p \geq 1$, the metric space ($\boldsymbol{P}_p(M)$, W_p) is separable, and is complete if (M, d) is separable and complete.[22]

Wasserstein distance for $p = \infty$ [edit]

It is also possible to consider the Wasserstein metric for $p = \infty$. In this case, the defining formula becomes:

$$W_\infty(\mu, \nu) = \lim_{p \to +\infty} W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \gamma\text{-}\operatorname{essup} d(x, y),$$

where $\gamma\text{-}\operatorname{essup} d(x, y)$ denotes the essential supremum of $d(x, y)$ with respect to measure γ. The metric space ($\boldsymbol{P}_\infty(M)$, W_∞) is complete if (M, d) is separable and complete. Here, $\boldsymbol{P}_\infty$ is the space of all probability measures with bounded support.[23]

See also [edit]

- Hutchinson metric
- Lévy metric
- Lévy–Prokhorov metric
- Fréchet distance
- Total variation distance of probability measures
- Transportation theory
- Earth mover's distance
- Wasserstein GAN
- Kolmogorov–Smirnov test

References [edit]

1. ^ Vaserstein LN (1969). "Markov processes over denumerable products of spaces, describing large systems of automata" (PDF). *Problemy Peredači Informacii*. **5** (3): 64–72.
2. ^ Kantorovich LV (1939). "Mathematical Methods of Organizing and Planning Production". *Management Science*. **6** (4): 366–422. doi:10.1287/mnsc.6.4.366. JSTOR 2627082.

3. ^ Olkin I, Pukelsheim F (October 1982). "The distance between two random vectors with given dispersion matrices". *Linear Algebra and Its Applications*. **48**: 257–263. doi:10.1016/0024-3795(82)90112-4. ISSN 0024-3795.
4. ^ S. S. Vallander, Theory of Probability and its Applications, 1974, Volume 18, Issue 4, Pages 784–786 DOI: https://doi.org/10.1137/1118101
5. ^ Arjovsky M, Chintala S, Bottou L (July 2017). "Wasserstein Generative Adversarial Networks". *International Conference on Machine Learning 214-223*: 214–223.
6. ^ Petitjean M (2002). "Chiral mixtures" (PDF). *Journal of Mathematical Physics*. **43** (8): 4147–4157. Bibcode:2002JMP....43.4147P. doi:10.1063/1.1484559. S2CID 85454709.
7. ^ Petitjean M (2004). "From shape similarity to shape complementarity: toward a docking theory". *Journal of Mathematical Chemistry*. **35** (3): 147–158. doi:10.1023/B:JOMC.0000033252.59423.6b. S2CID 121320315.
8. ^ Mukherjee S, Wethington D, Dey TK, Das J (March 2022). "Determining clinically relevant features in cytometry data using persistent homology". *PLOS Computational Biology*. **18** (3): e1009931. arXiv:2203.06263. Bibcode:2022PLSCB..18E9931M. doi:10.1371/journal.pcbi.1009931. PMC 9009779. PMID 35312683.
9. ^ Frederick, Christina; Yang, Yunan (2022-05-06). "Seeing through rock with help from optimal transport". *Snapshots of Modern Mathematics from Oberwolfach*. doi:10.14760/SNAP-2022-004-EN.
10. ^ Oizumi, Masafumi; Albantakis, Larissa; Tononi, Giulio (2014-05-08). "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0". *PLOS Computational Biology*. **10** (5): e1003588. Bibcode:2014PLSCB..10E3588O. doi:10.1371/journal.pcbi.1003588. PMC 4014402. PMID 24811198.
11. ^ Ba, Demba; Dogra, Akshunna S.; Gambhir, Rikab; Tasissa, Abiy; Thaler, Jesse (2023-06-29). "SHAPER: can you hear the shape of a jet?". *Journal of High Energy Physics*. **2023** (6): 195. arXiv:2302.12266. Bibcode:2023JHEP...06..195B. doi:10.1007/JHEP06(2023)195. ISSN 1029-8479. S2CID 257205971.
12. ^ "Awards, fellowships and the shape of physics: News from the College | Imperial News | Imperial College London". *Imperial News*. 2023-03-29. Retrieved 2023-10-31.
13. ^ Clement P, Desch W (2008). "An elementary proof of the triangle inequality for the Wasserstein metric". *Proceedings of the American Mathematical Society*. **136** (1): 333–339. doi:10.1090/S0002-9939-07-09020-X.
14. ^ Villani, Cédric (2003). "Chapter 1: The Kantorovich Duality". *Topics in optimal transportation*. Providence, RI: American Mathematical Society. ISBN 0-8218-3312-X. OCLC 51477002.
15. ^ Matoušek, Jiří; Gärtner, Bernd (2007), "Duality of Linear Programming", *Understanding and Using Linear Programming*, Universitext, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 81–104, doi:10.1007/978-3-540-30717-4_6, ISBN 978-3-540-30697-9, retrieved 2022-07-15
16. ^ Villani, Cédric (2003). "1.1.3. The shipper's problem.". *Topics in optimal transportation*. Providence, RI: American Mathematical Society. ISBN 0-8218-3312-X. OCLC 51477002.
17. ^ Benamou, Jean-David; Brenier, Yann (2000-01-01). "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". *Numerische Mathematik*. **84** (3): 375–393. doi:10.1007/s002110050002. ISSN 0945-3245. S2CID 1100384.
18. ^ Finlay, Chris; Jacobsen, Joern-Henrik; Nurbekyan, Levon; Oberman, Adam (2020-11-21). "How to Train Your Neural ODE: the World of Jacobian and Kinetic Regularization". *International Conference on Machine Learning*. PMLR: 3154–3164. arXiv:2002.02798.

19. ^ Peyre R (October 2018). "Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance". *ESAIM: Control, Optimisation and Calculus of Variations*. **24** (4): 1489–1501. doi:10.1051/cocv/2017050. ISSN 1292-8119. (See Theorem 2.1.)
20. ^ Loeper G (July 2006). "Uniqueness of the solution to the Vlasov–Poisson system with bounded density". *Journal de Mathématiques Pures et Appliquées*. **86** (1): 68–79. arXiv:math/0504140. doi:10.1016/j.matpur.2006.01.005. ISSN 1292-8119. (See Theorem 2.9.)
21. ^ Peyre R (October 2018). "Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance". *ESAIM: Control, Optimisation and Calculus of Variations*. **24** (4): 1489–1501. doi:10.1051/cocv/2017050. (See Theorem 2.5.)
22. ^ Bogachev VI, Kolesnikov AV (October 2012). "The Monge–Kantorovich problem: achievements, connections, and perspectives". *Russian Mathematical Surveys*. **67** (5): 785–890. Bibcode:2012RuMaS..67..785B. doi:10.1070/RM2012v067n05ABEH004808. S2CID 121411457.
23. ^ Givens, Clark R; Shortt, Rae Michael (1984). "A class of Wasserstein metrics for probability distributions". *Michigan Mathematical Journal*. **31** (2): 231–240. doi:10.1307/mmj/1029003026.

Further reading [edit]

- Ambrosio L, Gigli N, Savaré G (2005). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Basel: ETH Zürich, Birkhäuser Verlag. ISBN 978-3-7643-2428-5.
- Jordan R, Kinderlehrer D, Otto F (January 1998). "The variational formulation of the Fokker–Planck equation". *SIAM Journal on Mathematical Analysis*. **29** (1): 1–17 (electronic). CiteSeerX 10.1.1.6.8815. doi:10.1137/S0036141096303359. ISSN 0036-1410. MR 1617171. S2CID 13890235.
- Rüschendorf L (2001) [1994], "Wasserstein metric", *Encyclopedia of Mathematics*, EMS Press
- Villani C (2008). *Optimal Transport, Old and New*. Springer. ISBN 978-3-540-71050-9.

External links [edit]

- "What is the advantages of Wasserstein metric compared to Kullback–Leibler divergence?". *Stack Exchange*. August 1, 2017.

Categories: Measure theory | Metric geometry | Theory of probability distributions | Statistical distance