

Measuring Overlap, or Measuring Displacement

— KL and Wasserstein Are Not Competitors: What Happens When the Supports Come Apart, and Why the Rate Changes Character at $d = 3$ — [Paper 139]

Yuuki YAMAGISHI (山岸 勇輝)

Yamagishi Complexity Torus Lab, Japan

August 2026

— *From thinking to syncing.* —

ABSTRACT

Paper 137 wrote that the KL divergence is not a distance. **This paper sets its counterpart, the Wasserstein distance, beside it, and writes that the two are not competitors but answer different questions. NO NEW MATHEMATICAL THEOREM AND NO NEW PHYSICAL LAW IS CLAIMED.** First, the Wasserstein distance is a metric. Over 3000 random triples of samples, **asymmetries found: 0; failures of the triangle inequality: 0.** The KL divergence of Paper 137 failed the same test **25.4% of the time** (Section 2). **Second, and this is the core. What separates them is disjoint supports.** For two point masses a distance d apart, **KL is infinite for every d** while W_2 equals d exactly (checked at $d = 0.5, 1, 5, 100$). While the supports overlap both see the separation: for Gaussians of equal width **KL grows as $d^2/2$ and W_2 as d (12.500000 against 5 at $d = 5$, agreeing exactly with the closed form).** **How different are two distributions is two questions: do they overlap, and how far apart are they** (Section 3). **Third, Wasserstein is not the better object; it asks for more.** It requires a metric on the underlying space. **On three unordered categories KL is well defined at 0.223805, while Wasserstein cannot be formed until someone says how far red is from green** (Section 4). **Fourth, the heat equation can be written twice.** For the heat kernel started from a point, $W_2 = \sqrt{2t}$ (checked at $t = 0.25, 1, 4$). **This is $[t] = L^2$, the parabolic scaling of Paper 119. And by Otto's theorem the heat equation is the gradient flow of entropy in this metric: one equation, two descriptions** (Section 5). **Fifth, the rate of convergence changes character at $d = 3$.** Measuring W_1 between two samples of the **same** distribution (true value 0), a fourfold sample improves the estimate by a factor **2.18 at $d = 1$, 1.59 at $d = 3$ and 1.22 at $d = 8$. The rate is $n^{-1/2}$ up to $d = 2$ and $n^{-1/d}$ from $d = 3$: the central-limit rate and the covering rate change places there** (Section 6).

Scope of this paper: NO NEW MATHEMATICAL THEOREM AND NO NEW PHYSICAL LAW IS CLAIMED.

Monge, Kantorovich, Otto's gradient flow and the Fournier–Guillin rates are **all standard**. What this paper adds is only **counting the metric properties beside KL's, writing disjoint support as the separator, writing the asymmetry of prerequisites, and checking the two regimes of the convergence rate numerically. It is not claimed that Wasserstein is better than KL. On the numerics: at $d = 1$ the transport cost is computed exactly by sorting. For $d \geq 2$ it is computed by log-domain Sinkhorn with entropic regularisation, and the implementation was validated against the exact one-dimensional answer (relative error 0.1 to 1.1%). The sliced Wasserstein distance was deliberately not used: it averages one-dimensional projections, its rate is dimension-free, and it would have hidden the very effect under test.**

1. Notation and position

For two probability measures μ, ν write the p -Wasserstein distance as

$$W_p(\mu, \nu) = \left(\inf_{\gamma} \int \|x - y\|^p d\gamma(x, y) \right)^{1/p}, \tag{1}$$

the infimum being over couplings γ with marginals μ and ν .

Position: existing results are laid side by side and their reach written down. No new theorem is claimed. **This paper continues Paper 137 and pairs with it.**

2. The Wasserstein distance is a metric

Verification 2.1: Over 3000 triples of samples drawn from Gaussians on the line, symmetry and the triangle inequality were tested for W_1 .

	Wasserstein (here)	KL (Paper 137)
asymmetries	0 / 3000	always asymmetric
triangle failures	0 / 3000	25.4%

(QED)

Observation 2.2

Wasserstein satisfies the metric axioms, and therefore gives a consistent ordering of which is further from which — exactly what KL could not give in Paper 137.

3. The separator is disjoint support

Verification 3.1 (while they overlap): Shifting a Gaussian of fixed width by d , the numerics here give

d	KL (numeric)	$d^2/2$	W_2
0.5	0.125000	0.125000	0.5
1.0	0.500000	0.500000	1.0
5.0	12.500000	12.500000	5.0

Exact agreement with the closed form. While the supports overlap, both see the separation (KL as d^2 , W_2 as d). (QED)

Verification 3.2 (when they do not): For two point masses a distance d apart,

d	KL	W_2
0.5	infinite	0.5
1.0	infinite	1.0
5.0	infinite	5.0
100.0	infinite	100.0

KL returns the same value for every d and cannot tell them apart. W_2 returns d itself. (QED)

Observation 3.3 (this is the core)

The question of how different two distributions are is in fact two questions.

question	the tool	when supports are disjoint
do they overlap	KL (Paper 137)	infinite, and nothing further
how far apart are they	Wasserstein	returns the distance

One word, “apartness”, has been lumping the two together (Paper 17).

4. But the prerequisites differ

Observation 4.1

Equation (1) contains $\|x - y\|$: Wasserstein requires a metric on the underlying space. The definition of KL, $\sum p \log(p/q)$, contains no such thing.

Verification 4.2: For $p = (0.5, 0.3, 0.2)$ and $q = (0.2, 0.5, 0.3)$ on three unordered categories, this paper obtains $D(p||q) = 0.223805$. Wasserstein, by contrast, cannot be formed until someone says how far red is from green. (QED)

Observation 4.3 (not a ranking)

Wasserstein is not simply the better object. It asks for more input, and where that input does not exist it cannot be formed. Two tools, two prerequisites (Paper 95).

5. The heat equation can be written twice

Verification 5.1: The heat kernel started from a point is $\rho(\cdot, t) = N(0, 2t)$, so $W_2(\delta_0, \rho_t) = \sqrt{2t}$. At $t = 0.25, 1, 4$ this paper obtains 0.708606, 1.414860 and 2.829112 (against 0.707107, 1.414214, 2.828427). (QED)

Observation 5.2 (the scaling)

$W_2 \propto \sqrt{t}$ is $[t] = L^2$ — the parabolic scaling Paper 119 wrote for its fourth clock.

Observation 5.3 (Otto's theorem)

The heat equation is the gradient flow of entropy in the Wasserstein metric (Otto 2001; Jordan–Kinderlehrer–Otto 1998). **One equation, two descriptions** — as a partial differential equation, and as steepest descent. The second reading has meaning only once this metric is in place.

And the corpus contains the same shape: Paper 110 wrote that Ricci flow is a renormalisation-group flow. **“One equation, two descriptions”** happened there too, at a different address (there the string worldsheet, here the space of probability measures).

6. The rate changes character at $d = 3$

Verification 6.1: W_1 between two samples of the same standard normal — the true value is 0, so what is measured is pure estimation error.

d	$n = 64$	$n = 256$	ratio	predicted
1	0.22411	0.10263	2.18	2.00
2	0.55029	0.27474	2.00	2.00
3	0.82414	0.51940	1.59	1.59
5	1.48053	1.06582	1.39	1.32
8	2.26564	1.86390	1.22	1.19

The prediction is $4^{\min(1/2, 1/d)}$. (QED)

Observation 6.2 (two regimes)

The rate is $n^{-1/2}$ up to $d = 2$ and $n^{-1/d}$ from $d = 3$. So the curse of dimensionality begins at $d = 3$, where $1/d$ drops below $1/2$ — the central-limit rate and the covering rate change places there.

Observation 6.3 (this is a dimension that selects)

Paper 114 separated facts that select a dimension from facts that do not. The $d = 3$ here is on the selecting side — though what is selected is not a geometry but which of two rates wins. And it shares its number with the $d \geq 3$ of James–Stein in Paper 133 while having a different root (there admissibility, here covering numbers).

Observation 6.4 (exactness and measurability)

Satisfying the metric axioms does not make a quantity easy to estimate. The shape seen in Paper 136 returns: being right and being measurable are different.

7. Conclusion

1. Wasserstein is a metric (Ver. 2.1): 0 asymmetries and 0 triangle failures in 3000 triples, against 25.4% for KL in Paper 137.
2. The separator is disjoint support (Ver. 3.2). KL is infinite at $d = 0.5$ and at $d = 100$ alike; W_2 returns d .
3. “How different” is two questions (Obs. 3.3): do they overlap, and how far apart are they.
4. But the prerequisites differ (Ver. 4.2). Wasserstein demands a metric on the underlying space and KL does not. This is a difference of prerequisites, not a ranking.
5. The heat equation can be written twice (Ver. 5.1, Obs. 5.3). $W_2 = \sqrt{2t}$, that is $[t] = L^2$ (Paper 119), and by Otto's theorem it is a gradient flow.
6. The rate changes character at $d = 3$ (Ver. 6.1): $n^{-1/2}$ up to $d = 2$, $n^{-1/d}$ from $d = 3$. Being a metric does not make it measurable (Paper 136).

Closing: KL and Wasserstein are not in competition. One measures overlap and the other measures displacement. Where the supports come apart the difference shows most sharply: one can say only “infinitely different”, and the other answers with a distance. Which to use is decided by which of the two questions one is holding.

References

1. Kantorovich, L. V. (1942). *On the translocation of masses*. Doklady Akademii Nauk SSSR, 37.
2. Jordan, R., Kinderlehrer, D. & Otto, F. (1998). *The variational formulation of the Fokker-Planck equation*. SIAM Journal on Mathematical Analysis, 29.

3. Otto, F. (2001). *The geometry of dissipative evolution equations: the porous medium equation*. Communications in Partial Differential Equations, 26.
4. Villani, C. (2009). *Optimal Transport: Old and New*. Springer.
5. Fournier, N. & Guillin, A. (2015). *On the rate of convergence in Wasserstein distance of the empirical measure*. Probability Theory and Related Fields, 162.
6. Cuturi, M. (2013). *Sinkhorn distances: lightspeed computation of optimal transport*. Advances in Neural Information Processing Systems, 26.
7. Yamagishi, Y. (2026). *The Equals Sign Has Distinct Roots Too* [Paper 16], 3rd ed. Zenodo. 10.5281/zenodo.21600471
8. Yamagishi, Y. (2026). *The 8π Is an Artefact of the Units* [Paper 17], 3rd ed. Zenodo. 10.5281/zenodo.21600813
9. Yamagishi, Y. (2026). *Two Entropies* [Paper 88]. Zenodo. 10.5281/zenodo.21747809
10. Yamagishi, Y. (2026). *“Dimension” Is Not One Thing* [Paper 94]. Zenodo. 10.5281/zenodo.21758581
11. Yamagishi, Y. (2026). *Convention and Fact* [Paper 95]. Zenodo. 10.5281/zenodo.21758621
12. Yamagishi, Y. (2026). *Reading the Interior from the Boundary* [Paper 99]. Zenodo. 10.5281/zenodo.21758790
13. Yamagishi, Y. (2026). *The Crossing of Royal Roads* [Paper 100]. Zenodo. 10.5281/zenodo.21758864
14. Yamagishi, Y. (2026). *Extracting Signal from Noise* [Paper 106]. Zenodo. 10.5281/zenodo.21836984
15. Yamagishi, Y. (2026). *Ricci Flow IS a Renormalization Group Flow* [Paper 110]. Zenodo. 10.5281/zenodo.21837476
16. Yamagishi, Y. (2026). *Modules Are Not Walls but Limits* [Paper 113]. Zenodo. 10.5281/zenodo.21962996
17. Yamagishi, Y. (2026). *Scale Invariance Fixes an Exponent — in Four Different Ways* [Paper 117]. Zenodo. 10.5281/zenodo.22005515
18. Yamagishi, Y. (2026). *The Fourth Clock* [Paper 119]. Zenodo. 10.5281/zenodo.22036353
19. Yamagishi, Y. (2026). *The Fisher Metric Is Unique Only on the Classical Side* [Paper 133]. Zenodo. 10.5281/zenodo.22080550
20. Yamagishi, Y. (2026). *The Equality Is Exact, and the Average Is Held by the Tail* [Paper 136]. Zenodo. 10.5281/zenodo.22081230
21. Yamagishi, Y. (2026). *The KL Divergence Is Not a Distance, but Its Second Derivative Is a Metric* [Paper 137]. Zenodo. 10.5281/zenodo.22081369
22. Yamagishi, Y. (2026). *Which Part Is the Hamiltonian Is a Convention* [Paper 138]. Zenodo. 10.5281/zenodo.22081512

— *From thinking to syncing.* —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)